Concepts in terms of other Concepts: a method for the analysis of collective word use

Gustavo Cilleruelo Calderón, Nataliya Stepanova, Emily Allaway & Alexandra Birch
{g.cilleruelo-calderon, n.p.stepanova}@sms.ed.ac.uk, {emily.allaway, a.birch}@ed.ac.uk
School of Informatics, University of Edinburgh

Abstract

Sometimes how concepts are used in day-to-day talk differs from the expectations set in explicit, socially accepted, definitions. For example, many occupations which are inherently gender-neutral exhibit a gender bias in their usage across different corpora. We propose a method to quantify interactions between concepts in collections of data using autoregressive language models. Instead of prompting or extracting internal representations (i.e. embeddings), we measure differences in token probability distributions on minimal pairs of subject-verb-complement sentences, where the main noun in the subject varies across different 'concept-words'. We demonstrate the expressivity of this approach with a case study where we recover well-known social biases in language from contemporary internet data: ranking concepts in relation to *men-women* reveals gender biases in what would be expected to be gender neutral vocabulary. We also show how our method can be used to examine concepts across different domains – time periods, authors or ideologies – by selecting sentences from different corpora (e.g. news articles from the 19th century). We extend this to arbitrary conceptual 'dimensions', enabling the study of concepts in terms of other concepts with state-of-the-art language models.

Code and data are available at gustavocilleruelo.com/concepts.

Keywords: concepts; generics; language models; social bias; gender

Introduction

Concepts have an explicit and public character, an 'intuitive' meaning, that sometimes differs from their actual use in linguistic and social practice (Haslanger, 1995, 2005). Some of these concepts are about groups of people, e.g. *patients*, *women* or *judges*. It is an important task in social psychology to understand how these terms are collectively formed and used, beyond (or despite!) their explicit definitions.

The extensive literature on linguistic gender biases lays out the issue clearly: apparently gender neutral concepts (such as *nurse* or *soldier*) become gendered when corpus statistics about their actual use are analyzed (Bailey et al., 2022). How these words are used *collectively* is what determines what they imply and suggest. Especially in cases where social groups are involved, we should critique and understand when observed "social meaning" differs from the normative "dictionary" definition.

This work presents a *novel computational method to study relationships between concepts that uses autoregressive language models on collections of sentences*. We use these models not with prompting nor for their internal representations, but as probabilistic models of language: we compare probability distributions on certain tokens before and after interventions on the target concept word. Because we analyze *collections of sentences in context*, results are linked to the source documents rather than the model. We demonstrate this flexibility by analyzing concepts in news articles from the 19th century *without* having to retrain models from scratch. Although in this work we focus on social groups, like *judges* or *victims*, the core ideas we present can easily be extended to the study of other parts of language.

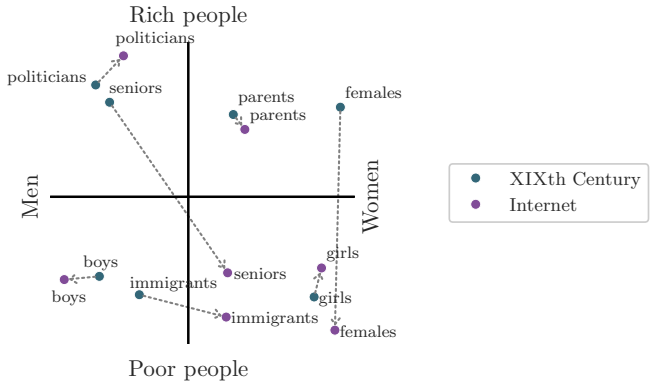

Figure 1: Our method can embed concepts in spaces defined in terms of other arbitrary concepts. Here we visualize meaning shifts for some social kinds between 19th century and internet data. Note how some words have remained stable, while others undergone massive changes.

Current distributional methods that analyze concepts and linguistic social biases rely on word embeddings (vector representations) and study the geometrical properties of those vectors as a proxy for word meaning. While this is feasible with static embeddings, e.g. GloVe (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013), and bidirectional encoders (BERT; Devlin et al., 2018), these algorithms have quickly become small and outdated in comparison to autoregressive language models (also commonly referred to as large language models). Autoregressive models have a very different training objective: predicting the next token (Radford et al., 2019). This means, for example, that internal vector representations of a token in an autoregressive model do not have information about the token's right context, which makes it

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

hard to extend embedding methods to these families of models of language, which are state-of-the-art across almost all benchmarks.

The method we present not only enables the use of large, autoregressive language models to measure semantic differences between words, but also embeds the concepts referred to by these words in a space defined in terms of *other concepts*. As an example, Figure 1 situates social kinds as they appear on the internet and 19th century sentences on dimensions defined by *men-women* and *rich people-poor people*. In terms of gender, the term *seniors* was closer to *men* during the 19th century than in today's internet. But this is not the only 'conceptual dimension' in which *seniors* can be analyzed, and we see that, although the word was closer to *rich people* in the 19th century, nowadays it has a different, less optimistic, economic connotation.

Further in the paper we introduce our method in more detail, explaining how it uses language models to measure the impact of interventions on concept terms. We demonstrate our method's expressivity by visualizing the gender bias on the internet and in 19th century data. We also discuss how this analysis can be extended to arbitrary 'concept-dimensions'. Finally, we mention possible uses of our method in social psychology, philosophy and cognitive science.

Background

Concept analysis with language models

Many prior works have investigated social bias by using language models to obtain quantified representations of social concepts (Bailey et al., 2022; Bailey et al., 2024; Giulianelli et al., 2020; Haket & Daniels, 2025; Ju et al., 2024; Shani et al., 2023). Since language models are trained on large sets of linguistic data, their learned representations reflect the underlying linguistic distribution of the corpora. This allows researchers to form and test hypotheses about how different social groups are discussed in different domains.

Recent computational and distributional analyses of concepts use have often relied on BERT-based language models (bi-drectional encoders) because they provide an easy and simple method of obtaining contextualized word embeddings, which provide a computational basis for testing various hypotheses. On a high level, a common approach consists of: *(i)* extracting embeddings for the relevant concepts from a model, and *(ii)* using some measure of distance or probabilistic association to rank concepts along the social dimension of interest.

Applying similar methods to study gender bias, Bailey et al. (2022) and Bailey et al. (2024) use word embeddings and empirically observe the androcentric bias (a man is the "standard" person and a woman is the gendered version, also observed by Chang and McKeown (2019)). Ju et al. (2024) use word embeddings paired with information-theoretic metrics to conclude that "typically female" occupations are often accompanied by explicit gender mentions. Deligianni and Horne (2023) use topic analysis to study gender bias on Reddit.

Some works study how concepts change and evolve over time. Loureiro et al. (2022) study concepts on Twitter by training different models on quarterly data. From a more historical perspective, Giulianelli et al. (2020) map the distribution change of several concepts from 1810 to 2009, finding that their proposed metrics were effective in capturing broadening, narrowing, and metaphorization in word meaning.

In general, BERT-based approaches rely on word embeddings by models that are small by current standards, and may therefore not capture many subtleties in the semantics of complex collective concepts. Recent work has moved to using larger state-of-the-art autoregressive language models and can be broadly characterized into two streams (Rowe et al., 2025): *(i)* extrinsic evaluation approaches detect bias through prompting and assessing the output of specific tasks (e.g., Wan et al., 2023); and *(ii)* probing of internal probability distributions, comparing the likelihoods of generating different sentences (e.g., Mitchell et al., 2025).

We focus on the second stream because our proposed method falls into that category. Probing into a model's internal biases usually consists of modeling the change in likelihood of a minimal pair of sentences (e.g. gendered vs. non-gendered). Pikuliak et al. (2024) inspect changes on the probability distribution over the token of interest (e.g. in *he/she*, *woman/man*). Mitchell et al. (2025) compare probabilities between a stereotype and a contrastive sentence on autoregressive language models, inspired by Nangia et al. (2020), by comparing averaged log likelihoods of the common tokens to detect preference. These comparisons rely on annotated sentences, where the tokens that correspond to the stereotype are known, so they may not be scalable to large collections of data. Regardless of the overall approach, these works focus on evaluating the bias *of the language model*, rather than using a language model to study how the concepts are used in language.

Additionally, the use of collective concepts crucially depends on the surrounding sentential context – simply measuring the co-occurrence rates of the group and attribute may not be a sufficient indicator of stereotyping rates (Davani et al., 2024). Our target sentences are evaluated *in context* (rather than as isolated sentences), which is more reflective of how claims about social groups are actually encountered.

Language Models

This work uses language models as probabilistic models of language: we are not concerned with prompting the models in a question-answer setting, nor with their internal vector representations, but with the probability distributions for the next token given a context. Given a textual context c, an autoregressive model of language θ and a vocabulary $\mathcal{V}$, the model of language is a function that returns a probability distribution over all tokens $x \in \mathcal{V}$:

$$\theta(c) = P(\mathcal{V})$$

For each token, the probability that it follows from c is given by $\theta(x|c) = p(x|c)$. This definition is for *autoregressive* language models, which are trained to predict the next token,

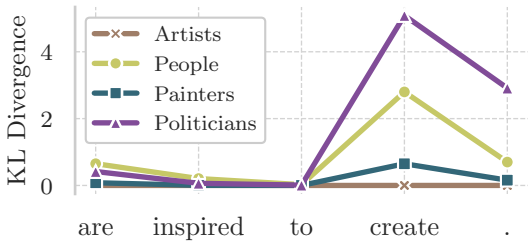

Figure 2: KL-divergences with respect to the original sentence (which has *Artists* as subject) for *Painters*, *Politicians* and *People*. As expected, *Painters* keeps the token probability distribution very close to *Artists* on the 'create' token. Note the KL-divergence with respect to the same distribution (*Artists* with itself) is 0 (OLMO3-32B).

without information about words in the right context. While other types of language models exist, such as bi-directional encoders (like BERT (Devlin et al., 2018)), which have different training objectives, we focus on autoregressive language models as they are the current state-of-the-art architectures. Their success on a variety of tasks make them good candidates for the distributional study of meaning and concepts in language (Baroni, 2021; Grindrod, 2024a). Our main results are presented with OLMO3-32B (Olmo et al., 2025), a competitive fully open-source pre-trained model.

There is an ongoing debate as to the extent to which findings about language usage obtained from language models can be generalized to inferences about the nature of language overall (McCoy et al., 2024). In this paper, we assume that language models are a good way to track linguistic activity across a large group of people and therefore serve as good models of the "social object" aspect of language use (Grindrod, 2024a, 2024b).

It may be important to note that we are not making any connections between the algorithm for language modeling and the cognitive processes that take part in human language production: we provide *a tool for the distributional analysis of textual sources*. While language distribution, as well as the information measures derived from language models (Giulianelli et al., 2024), may be closely related to some cognitive processes, making claims about such connections would require a much different empirical and theoretical elaboration.

Throughout the rest of the paper, we refer to *autoregressive language models* simply as *language models*.

Methodology

Meaning from Token Probabilities

In this work we deal with sentences consisting of a simple subject-verb-complement (SVC) structure where the 'concept-word' is a plural noun in the main noun phrase of the subject. Throughout this work we refer to postverbal complements coming strictly after the verb, simply as *complements*. Let us call these plural nouns GoP, as we are interested in groups of people. We target sentences with the following structure:

$$\text{(modifiers) GoP (modifiers) } V\ C. \tag{1}$$

We study sentences with this syntax as they are often used to predicate general properties of the GoP, and what is said about a group of people (e.g. *kids* or *workers*) is closely related on how these are understood collectively. Modifiers can be adverbs, determiners or clauses that complement the main plural noun or verb.

Intuitively: the things we say after *people are*, *firemen are* or *women are* are different, and this difference is related to *how different the concepts themselves are*. By changing the GoP (for example, substituting *firemen* for one of *people*, *women* or *men*), we obtain minimal pairs, for which the only difference is the plural noun in the subject. This modification will, given a language model θ, affect the probability distributions over tokens of the verb complements C.

In order to quantify how modifying the GoP word changes the type of properties that would be predicated, we use KL-divergence, a standard method to compare probability distributions:

$$D(p||q) = \sum_{x \in \mathcal{V}} p(x) \log \frac{p(x)}{q(x)} \tag{2}$$

In our case, p is the probability distribution on a token in the object at position i, and q the probability distribution at that same position on the modified sentence of the minimal pair. We go over a detailed example of this comparison in the next section.

Comparing token probability distributions this way, we can define how close two GoP-words are in a particular sentence by measuring *their maximum KL-divergence on a complement token for a given minimal pair*:

$$\max_{i \in C} D(p_i||q_i) \tag{3}$$

When GoP are conceptually similar (in context), the possible things to be said in the object will be close, and this is reflected in their respective token probability distributions. A similar method, also comparing token probability distributions on postverbal tokens after interventions, is introduced in Cilleruelo et al. (2026) for the study of generics and quantification.

Example: *Artists are inspired to create.*

In this example, we are going to situate *artists* in relation to *painters*, *politicians* and *people* in the following target sentence in context:

> Music is one of the biggest; dancers are inspired to dance because of music. Singers are inspired to sing. *Artists* are inspired to create.

This will give us 3 minimal pairs, one for each plural noun with respect to the original. For example for *politicians* we would have "Music is one of the biggest; dancers are inspired to dance because of music. Singers are inspired to sing.

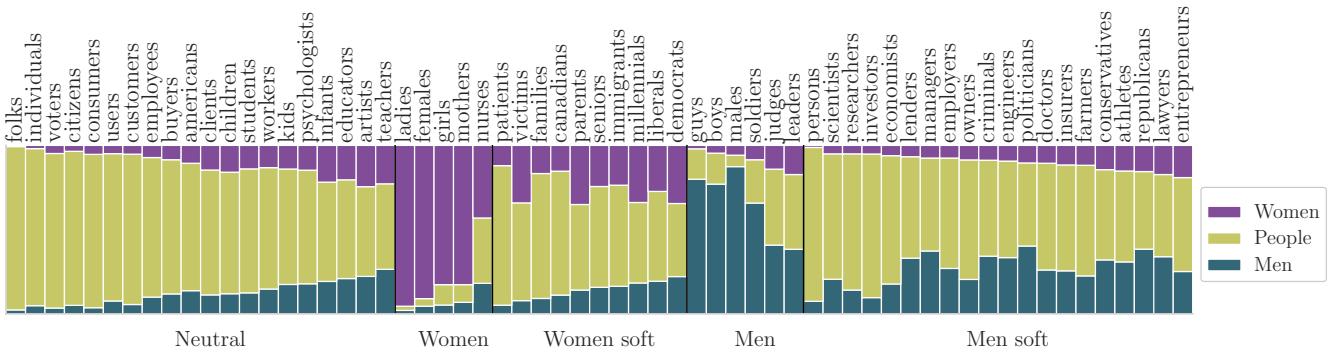

Figure 3: Percentage of sentences about different groups of people for which *women*, *people* or *men* are closer in meaning on internet data, with 1000 sentences in context for each social group.

Politicians are inspired to create." and calculate the probability distributions on each token, which will differ from the original (*artists*) sentence. Ideally, the fact that *artists* and *politicians* are usually inspired to do different things should affect the token probability distributions from the language model on certain tokens in the post-verbal complements.

Figure 2 shows the KL-divergence of the modified sentences with respect to the original at each token. Note how for *artists* it is always 0 since the KL-divergence of a probability distribution with respect to itself is 0. On the key token of the predicate, 'create', we observe how *painters* keeps the token probability distribution the most similar, while *politician* is very different, more so than for a neutral term like *people*.

The analysis of a single sentence can hardly give any insight into the heterogeneous meaning that words can have across contexts. In the experiments that follow, we choose a *dimension*, i.e. a set of two concepts (in this case *politicians-painters*), and calculate their closeness with different target words (for example *artists*), comparing also to *people* as a control. There may be cases where the context, GoP noun or predicate are unrelated to the words in the dimension, and in those cases we expect the generalist , *people*, to have the lowest KL-divergence with respect to the original.

Data: from the internet and the 19th century

Internet. To easily get sentences from the internet with social groups of people in the subject position, we filter sentences from MGen: a dataset of over 3 million generics and 1 million quantified sentences in context (Cilleruelo et al., 2025). Although this dataset is originally designed to study generic and quantified sentences, we use it as a general source of natural sentences in context that predicate things about groups of people. We filter the dataset using a simple prompt on a frontier language model (GPT4o; OpenAI et al., 2024) to label whether the first word of each sentence (excluding quantifiers) is a plural noun about a group of people. Words with more than 1000 occurrences as the first word in the subject are verified by the authors to indeed be about groups of people. We thus end up with a total of 108 GoP words, with 1000 sentences sampled for each (note that some of these are omitted from the visualizations for clarity).

19th century. The AmericanStories dataset contains around 20 million digitalized news articles from the 19th and early 20th centuries (Dell et al., 2023). We filter sentences from documents between 1845 and 1899 with a simple heuristic, selecting sentences with one of the social kinds from before (from MGen) in subject position. We again select 1000 samples per GoP word for 92 of those words (some common words on the internet are very rare in the 19th century data).

All samples from both collections have *3 sentences of left context*. This reduces the uncertainty of the language model under the particular language distribution of interest (in this case, generalist internet text and news articles from the 19th century).

The gender dimension

To motivate the expressive power of these comparisons through KL-divergence on complement tokens, we first study the groups of people along the 'gender dimension'. We measure the KL-divergence after changing the noun in the original sentence for *women*, *men* and *people* (what we refer to as an 'intervention'). Conceptually, for a given social group of people, e.g. *teachers*, we estimate which of the gendered words (or the neutral case in *people*) keeps the possible tokens in the complement more similar to the original sentence.

For the 1000 samples of each GoP word in the internet data derived from MGen, we get the percentage of those for which the various interventions – *women*, *men* and *people* – are the closest, and calculate 95% confidence intervals (Newcombe, 1998; Wilson, 1927) to establish when these differences are significant. We argue that the differences in percentages are a decomposition of the original concept across the dimension implicitly defined by *women-people-men*.

As can be seen in Figure 3, this gives a natural clustering of concepts into five categories:

(i) *Neutral.* The difference in percentage for the *men* and *women* interventions is not statistically significant.

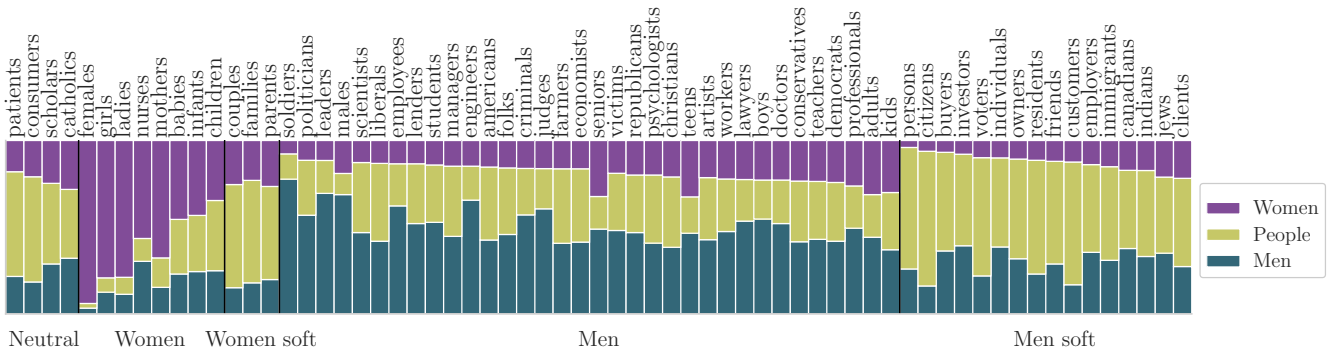

Figure 4: Language reflects a male-dominated America in the 19th century. *Neutral* and *Women* concepts are an minority, with the majority of social classes being associated with *Men*.

(ii) *Women.* The percentage for the *women* intervention is significantly larger than for both *men* and *people*.

(iii) *Women soft.* The percentage for the *women* intervention is significantly larger than for *men*, but not the neutral concept *people*.

(iv) *Men.* The percentage for the *men* intervention is significantly larger than for both *women* and *people*.

(v) *Men soft.* The percentage for the *men* intervention is significantly larger than for *women*, but not for *people*.

Figure 3 explicitly illustrates male bias in internet language, which can be a reflection of aspects of current social practice and structure. While many terms are gender neutral, social kinds that wield political power are biased towards *men*, e.g. *leaders*, *judges* or *politicians*, even though they should be, in principle, gender neutral. It is also notable how words related to *consumers* (e.g., *clients*, *buyers*, *costumers*) as well as *workers* and *employees* are gender neutral, while those concepts that imply control of economic resources, such as *owners*, *lenders*, *investors* or *employers*, all fall in the *Men soft* cluster.

The gender dimension in the 19th century

Social biases in society change throughout history. In this section, we demonstrate that our method can give insights into how concepts have changed from the 19th century to the internet today.

We reproduce the experimental setup from the previous section, this time situating the 'gender dimension' concepts in sentences from the AmericanStories dataset. We define the categories as before and discover a stark shift in the distribution of GoP targets across the five (especially *Men* and *Neutral*). Specifically, the *Men* category consists of a greater percentage of the GoP occupational target words in the 19th century data compared to the modern day internet, reflecting an almost exclusive male representation across most occupations in the American society of the time (Figure 4).

At the beginning of the 19th century, American society and most of the professions we analyze were strongly male-dominated. This changed throughout the century: around 1910 we see 'feminism', centered around suffrage reforms, appear as the term we know today (Blavatsky, 1888; Cruea, 2005). As seen in Figure 4, only social terms related to family have some *Woman* connotations, while all professions are strongly *Men*-gendered. Whereas the *Neutral* class covered many concepts on the internet data, here we observe a much more gendered language overall, with *people* being the majority class for few social kinds.

Concepts across different dimensions

In the previous sections we have focused only on one 'conceptual dimension' – gender, as it is well researched with numerous observed social biases. Our method is much more flexible, allowing the comparison of a concept in a sentence with any other that can fit the same syntactic role (i.e. is also a plural noun). To demonstrate this, we construct a scatterplot of various concepts across the gender *and* political dimensions. We model the political dimension with *liberals*, *conservatives* and *people* as neutral, similar to before.

First, in constructing the scatterplot we ignore samples that fall into the *Neutral* category (because this means these samples did not exhibit a significant difference across the measured conceptual axis). Next, we obtain the difference between the percentages of samples that are closest to both ends of the axis. This difference determines the position along that dimension.

It will be helpful to work through a concrete example. Suppose our target category is *soldiers*. Along the gender dimension, 80% of the 1000 samples were closest to the *men* intervention and 5% to *women* (with the remaining 15% closest to *people*, which we ignore). This means that the positioning of *soldiers* will be $0.05 - 0.75 = -0.70$ away from the *women* end of the axis (equivalently, 0.70 towards *men*). Supposing that along the political dimension, the same target concept *soldiers* was closest to the *conservatives* intervention 55% of the time and closest to the *liberals* intervention 35% of the time, we use the same logic as above to determine that its positioning along the political axis should be 0.20 towards *conservatives*. The final positioning of this example is there-

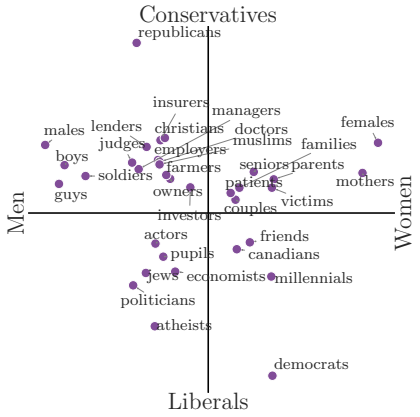

Figure 5: Some concepts situated in a space defined on internet data by gender (*men-women*) and political spectrum (*conservatives-liberals*).

fore $(-0.70, 0.20)$, so the target concept *soldiers* falls in the *men-conservatives* quadrant. In summary, the positioning of target concepts on the scatterplot is interpretable – the coordinates represent where the target concept lies in relation to the two chosen 'concept-axes' in a collection of sentences.

Figure 5 is suggestive of the *conservative-men* gaze often found in parts of the internet, but we leave a more fine-grained interpretation for future work. This approach can be extended to arbitrary dimensions defined by two concepts, even if these are not thematically opposed, for example *men-conservatives* and *young people-poor people*.

Discussion

We have demonstrated how intervening on subject plural nouns that refer to a group of people and measuring the effect of this intervention on the predicate yields *semantically expressive* relationships between concepts. This method can leverage state-of-the-art autoregressive language models without neither prompting nor operations on vector representations.

The core idea this paper presents is that to measure semantic variations between terms with an autoregressive language model, we should not examine some representation of the tokens of the term, but rather measure how interventions on that term (with grammatically compatible words) affect the probability distributions of tokens on the right (continuations). In this case study with social kinds, we have intervened on the subject by substituting the original term for some other and observed the impact of the substitution on the object tokens.

We have looked at gender bias in language to argue for the expressivity of this overall method and its usefulness for explicating the use of concepts across times, environments and ideologies. With the same pre-trained language model (OLMO3-32B), we have recovered very different gender biases for social concepts in internet (Figure 3) and 19th century data (Figure 4) respectively, matching our historical understanding of gender structures in the corresponding societies. This suggests that our method is, to a great extent, data-centric, i.e. what we recover are biases in a language distribution rather than in a language model, without the need of re-training. Because we get these expressive results on very limited data (just 1000 sentences in context for each social kind), this opens the door to a lightweight comparison of conceptual constructions in selected small collections (e.g. articles from a newspaper in a year). We believe this makes our method a useful tool for researchers interested in operative concepts across historical, ideological, or social communities. The fine-grained control over the data in which concepts are analyzed may be particularly relevant for *genealogical* approaches to concept analysis (Haslanger, 2005; Nietzsche, 1887): that is, historical descriptions of how concepts are embedded in evolving social practices.

We have proposed two visualizations, both based around comparing a concept with two others (e.g. *men* and *women*, what we have called 'dimension') and a neutral one (*people*). We believe the flexibility of being able to situate concepts in terms of other concepts, rather than in abstract vector spaces without interpretable axes, is a strength of the method we propose. Moreover, by including relatively long preceding contexts (3 sentences), we hope to capture semantic and pragmatic effects that arise from the higher-order role the analyzed sentence is playing in discourse. These two features of our method assume a central explanatory role of the *relation of concepts to other concepts*, as well as the *activities or social situations* these concepts are woven into.

Limitations & Future Work

We leave some design decisions and hyperparameters unexamined and offer only a limited commentary on the results. This is because our objective is to illustrate the method and show that it is expressive in reproducing known biases, leaving its application to open questions for future work. We focused only on sentences of the subject-verb-complement syntactic structure and intervened on the subject to compare probability distribution changes in the complements. We hypothesize that other types of syntactic structures and interventions can be explored. Experiments presented can also be easily extended to different language models, allowing for a possible comparison of they modeling of social concepts.

The natural next step in this investigation is to use this method as the empirical catalyst of a wider research project that studies the interactions between some particular collective concepts of interest.

Acknowledgments

No AI tool was used on the manuscript.

References

Bailey, A. H., Williams, A., & Cimpian, A. (2022). Based on billions of words on the internet, people = men. *Science Advances*, *8*(13). https://doi.org/10.1126/sciadv.abm2463

Bailey, A. H., Williams, A., Poddar, A., & Cimpian, A. (2024). Intersectional male-centric and white-centric biases in collective concepts. https://doi.org/10.31219/osf.io/72pg5

Baroni, M. (2021). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. https://doi.org/10.48550/ARXIV.2106.08694

Blavatsky, H. P. (1888). *The secret doctrine: The synthesis of science, religion, and philosophy* (Vol. 1-2). Theosophical Publishing Company.

Chang, S., & McKeown, K. (2019). Automatically inferring gender associations from language. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5745–5751. https://doi.org/10.18653/v1/d19-1579

Cilleruelo, G., Allaway, E., Haddow, B., & Birch, A. (2025). Mgen: Millions of naturally occurring generics in context. *Proceedings of the Society for Computation in Linguistics*, *8*(1), 11. https://doi.org/10.7275/scil.3147

Cilleruelo, G., Almotahari, M., Allaway, E., & Birch, A. (2026). Generics are not quantificational: A new path from language models to semantic theory. *Findings of the Association for Computational Linguistics: ACL 2026*.

Cruea, S. M. (2005). Changing ideals of womanhood during the nineteenth-century woman movement. *University Writing Program Faculty Publications*, *1*. https://scholarworks.bgsu.edu/gsw_pub/1

Davani, A. M., Gubbi, S., Dev, S., Dave, S., & Prabhakaran, V. (2024). Genil: A multilingual dataset on generalizing language. *arXiv preprint arXiv:2404.05866*.

Deligianni, A., & Horne, Z. (2023). Analysing hate speech towards women on reddit. https://doi.org/10.31234/osf.io/fsrq7

Dell, M., Carlson, J., Bryan, T., Silcock, E., Arora, A., Shen, Z., D'Amico-Wong, L., Le, Q., Querubin, P., & Heldring, L. (2023). American stories: A large-scale structured text dataset of historical u.s. newspapers.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Giulianelli, M., Del Tredici, M., & Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations, 3960–3973. https://doi.org/10.18653/v1/2020.acl-main.365

Giulianelli, M., Opedal, A., & Cotterell, R. (2024). Generalized measures of anticipation and responsivity in online language processing. https://doi.org/10.48550/ARXIV.2409.10728

Grindrod, J. (2024a). Modelling language. https://doi.org/10.48550/ARXIV.2404.09579

Grindrod, J. (2024b). Large language models and linguistic intentionality. *Synthese*, *204*(2). https://doi.org/10.1007/s11229-024-04723-8

Haket, N., & Daniels, R. (2025). Bert's conceptual cartography: Mapping the landscapes of meaning. *8*. https://doi.org/10.7275/SCIL.3145

Haslanger, S. (1995). Ontology and social construction. *Philosophical Topics*, *23*(2), 95–125. https://doi.org/10.5840/philtopics19952324

Haslanger, S. (2005). What are we talking about? the semantics and politics of social kinds. *Hypatia*, *20*(4), 10–26. https://doi.org/10.2979/hyp.2005.20.4.10

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ju, D., Ullrich, K., & Williams, A. (2024). Are female carpenters like blue bananas? a corpus investigation of occupation gender typicality. *Findings of the Association for Computational Linguistics: ACL 2024*, 4254–4274.

Loureiro, D., Barbieri, F., Neves, L., Espinosa Anke, L., & Camacho-collados, J. (2022). Timelms: Diachronic language models from twitter. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 251–260. https://doi.org/10.18653/v1/2022.acl-demo.25

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), e2322420121.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. https://arxiv.org/abs/1310.4546

Mitchell, M., Attanasio, G., Baldini, I., Clinciu, M., Clive, J., Delobelle, P., Dey, M., Hamilton, S., Dill, T., Doughman, J., et al. (2025). Shades: Towards a multilingual assessment of stereotypes in large language models. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 11995–12041.

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. (2020). Crows-pairs: A challenge dataset for measuring social biases in masked language models. *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 1953–1967.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods [Erratum in *Statistics in Medicine*. 1999 May 30;18(10):1293]. *Statistics in Medicine*, *17*(8), 873–890. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

Nietzsche, F. (1887). *On the genealogy of morals: A polemic*.

Olmo, T., : Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., Heineman, D., Groeneveld, D., Brahman, F., Timbers, F.,

Ivison, H., Morrison, J., Poznanski, J., Lo, K., Soldaini, L., Jordan, M., Chen, M., Noukhovitch, M., Lambert, N., ... Hajishirzi, H. (2025). Olmo 3. https://arxiv.org/abs/2512.13961

OpenAI, : Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., ... Malkov, Y. (2024). Gpt-4o system card. https://arxiv.org/abs/2410.21276

pandas development team, T. (2020, February). *Pandas-dev/pandas: Pandas* (Version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Pikuliak, M., Oresko, S., Hrckova, A., & Šimko, M. (2024). Women are beautiful, men are leaders: Gender stereotypes in machine translation and language modeling. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 3060–3083.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*.

Rowe, J., Klimaszewski, M., Guillou, L., Vallor, S., & Birch, A. (2025). Eurogest: Investigating gender stereotypes in multilingual language models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 32062–32084. https://doi.org/10.18653/v1/2025.emnlp-main.1632

Shani, C., Vreeken, J., & Shahaf, D. (2023). Towards concept-aware large language models. *arXiv preprint arXiv:2311.01866*.

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. Retrieved January 27, 2026, from http://www.jstor.org/stable/2276774