The Causal Influence of Grammatical Gender on Distributional Semantics

Karolina Stańczak[1,2,3,5] **Kevin Du**[3] **Adina Williams**[4]
Isabelle Augenstein[5] **Ryan Cotterell**[3]
[1]Mila – Quebec AI Institute, Canada [2]McGill University, Canada [3]ETH Zurich, Switzerland
[4]FAIR at Meta AI, USA [5]University of Copenhagen, Denmark
karolina.stanczak@mila.quebec kevin.du@inf.ethz.ch
adinawilliams@meta.com augenstein@di.ku.dk rcotterell@inf.ethz.ch

Abstract

How much meaning influences gender assignment across languages is an active area of research in linguistics and cognitive science. We can view current approaches as aiming to determine where gender assignment falls on a spectrum, from being fully arbitrarily determined to being largely semantically determined. For the latter case, there is a formulation of the neo-Whorfian hypothesis, which claims that even inanimate noun gender influences how people conceive of and talk about objects (using the choice of adjective used to modify inanimate nouns as a proxy for meaning). We offer a novel, causal graphical model that jointly represents the interactions between a noun's grammatical gender, its meaning, and adjective choice. In accordance with past results, we find a significant relationship between the gender of nouns and the adjectives that modify them. However, when we control for the meaning of the noun, the relationship between grammatical gender and adjective choice is near zero and insignificant.

1 Introduction

Approximately half of the world's languages have grammatical gender (Corbett, 2013a), a grammatical phenomenon that groups nouns together into classes that share morphosyntactic properties (Hockett, 1958; Corbett, 1991; Kramer, 2015). Among languages that have gender, there is variation in the number of gender classes; for example, some languages have only two classes, e.g., all Danish nouns are classed as either common or neuter, whereas others have significantly more, e.g., Nigerian Fula has around 20, depending on the variety (Arnott, 1967; Koval', 1979; Breedveld, 1995). Languages also vary with respect to how much gender assignment, i.e., how nouns are sorted into particular genders, is related to the form and the meaning of the noun (Corbett, 1991; Plaster and Polinsky, 2007; Corbett, 2013b, 2015; Kramer, 2020; Sahai and Sharma, 2021). Some languages group nouns into gender classes that are highly predictable from phonological (Parker and Hayward, 1985; Corbett, 1991, 2013b) or morphological (Corbett, 1991, 2013b; Corbett and Fraser, 2000) information, while others, such as the Dagestanian languages Godoberi and Bagwalal, seem to be predictable from meaning (Corbett, 1991; Corbett and Fraser, 2000; Corbett, 2015)—although, even for most of the strictly semantic systems, there are exceptions.

Despite this variation, gender assignment is rarely, if ever, wholly predictable from meaning alone. In many languages, there is a semantic core of nouns that are conceptually coherent (Aksenov, 1984; Corbett, 1991; Williams et al., 2019; Kramer, 2020) and a surround that is somewhat less semantically coherent. Axes along which genders are conceptually coherent often include semantic properties of animate nouns, with inanimate nouns appearing in the surround. For example, in Spanish, despite the fact that the nouns *table* (*mesa* in Spanish) and *woman* (*mujer* in Spanish) appear in the same gender (i.e., feminine), it is hard to imagine what meaning they share. Indeed, some linguists posit that gender assignment for inanimate nouns is effectively arbitrary (Bloomfield, 1935; Aikhenvald, 2000; Foundalis, 2002). And, to the extent that gender assignment is *not* fully arbitrary for inanimate nouns (Williams et al., 2021), many researchers argue there is no compelling evidence showing grammatical gender affects how we conceptualize objects (Samuel et al., 2019) or the distributional properties of language (Mickan et al., 2014).

However, not all researchers agree that non-arbitrariness in gender assignment, to the extent

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1672–1685, 2024. https://doi.org/10.1162/tacl_a_00723
Action Editor: Mark Johnson. Submission batch: 1/2024; Revision batch: 3/2024; Published 12/2024.

it exists, should be assumed to have no bearing on language production. Boroditsky (2003) famously argued for a *causal* relationship between the gender assigned to inanimate nouns and their usage, in a view colloquially known as the neo-Whorfian hypothesis after Benjamin Whorf (Whorf, 1956). Proponents of this view have studied human associations, under the assumption that people's perceptions of the genders of objects are strongly influenced by the grammatical genders these objects are assigned in their native language (Boroditsky and Schmidt, 2000; Semenuks et al., 2017). One manifestation of this perception is the choice of adjectives used to describe nouns (Semenuks et al., 2017). While this is an intriguing possibility, there are additional lexical properties of nouns that may act as confounders and, thus, finding statistical evidence for the causal effect of grammatical gender on adjective choice requires great care.

To facilitate a cleaner way to reason about the causal influence grammatical gender may have on adjective usage, we introduce a causal graphical model to represent the interactions between an inanimate noun's grammatical gender, its meaning, and the choice of its descriptors. This causal framework enables intervening on the values of specific factors to isolate the effects between various properties of languages. Our model explains the distribution of adjectives that modify a noun, conditioned on both a representation of the noun's meaning and the gender of the noun itself. Upon estimation of the parameters of the causal graphical model, we test the neo-Whorfian hypothesis beyond the anecdotal level. First, we validate our model by comparing it to the method presented in prior work without any causal intervention. Second, we employ our model with a causal intervention on the noun meaning to test the neo-Whorfian hypothesis. That is, we ask a counterfactual question: Had nouns been lexicalized with different grammatical genders but retained their same meanings, would the distribution of adjectives that speakers use to modify them have been different?

We employ our model on five languages that exhibit grammatical gender: four Indo-European languages (German, Polish, Portuguese, and Spanish) and one language from the Afro-Asiatic language family (Hebrew). We find that, at least in Wikipedia data, a noun's grammatical gender is indeed correlated with the choice of its descriptors. However, when controlling for a confounder, noun meaning, we present empirical evidence that noun gender has no significant effect on adjective usage.

2 A Primer on Grammatical Gender

In many languages with grammatical gender, adjectives, demonstratives, determiners, and other categories **agree** with the noun in gender, i.e., they will systematically change in form to indicate the grammatical gender of the noun they modify. Observe the following sentence, *A small dog sleeps under the tree.*, translated into two languages that exhibit grammatical gender (German and Polish):

a. ***Ein** klein**er** Hund schläft unter **dem** Baum.* (DE)
   a.M small.M dog.M sleeps under the.M tree.M

b. *Mały pies śpi pod drzew**em**.* (PL)
   a.M small.M dog.M sleeps under the.N tree.N

Because the German (DE) and Polish (PL) words for a dog, *Hund* and *pies*, are both assigned masculine gender, the adjectives in the respective languages, *klein* and *mały*, are morphologically gender-marked as masculine. Additionally, in German, the article, *dem*, is also gender-marked as masculine. The fact that gender is reflected by agreement patterns on other elements is generally taken to be a definitional property (Hockett, 1958; Corbett, 1991; Kramer, 2020) separating gender from other kinds of noun classification systems, such as numeral classifiers or declension classes.

It is an undeniable fact in many languages that morphological agreement reflects the gender of a noun in the *form* of other elements. However, one could imagine a similar process, such as analogical reasoning (Lucy, 2016), by which gender could influence both a noun's *meaning* and its form. If a noun's meaning were to influence its gender, then the noun meaning could also indirectly influence adjective usage, by way of the relationship between grammatical gender and adjective usage. There is ample statistical evidence that grammatical gender assignment is not fully arbitrary (Williams et al., 2019, 2021; Nelson, 2005; Sahai and Sharma, 2021). Such evidence is *prima facie* consistent with the idea that such influence is conceivably possible.

However, it is important to note that claims that noun gender influences meaning are by their very nature causal claims. The most famous example of such a causal claim is the neo-Whorfian view of gender (Boroditsky and Schmidt, 2000; Boroditsky, 2003; Semenuks et al., 2017), which states that a noun's grammatical gender *causally* affects meaning (e.g., adjective choice). Boroditsky and Schmidt (2000) discuss a laboratory experiment they have conducted showing that speakers of German chose stereotypically feminine adjectives to describe, for example, bridges, while speakers of Spanish chose stereotypically masculine adjectives. Boroditsky and Schmidt concluded that participant adjective choice reflected the fact that in German, the word for a bridge, *Brücke*, is grammatically feminine, while in Spanish, the word for a bridge, *puente*, is grammatically masculine. Their findings can be summed up in the following quote from Boroditsky and Schmidt (2000), ''people's ideas about the genders of objects are *strongly influenced* by the grammatical genders assigned to these objects in their native language'' (emphasis ours). Despite this clear causal formulation of the hypothesis, there has yet to be a modeling approach developed to test it. Moreover, subsequent studies have failed to replicate this result, raising into question the strength of this relationship between gender and adjective usage (Mickan et al., 2014) and inviting a study with more appropriate methodology.

Our paper builds on Williams et al.'s (2021) *correlational* study of noun meaning and its distributional properties and advances it to a *causal* one. While Williams et al. (2021) report non-trivial, statistically significant mutual information between the grammatical gender of a noun and its modifiers, e.g., adjectives that modify the noun, they do not control for other factors which might influence adjective usage, most notably the lexical semantics of the noun. Mutual information is symmetric and thus cannot speak to causation on its own. We are thus motivated by a potential common cause whereby the lexical semantics jointly influence a noun's grammatical gender *and* its distribution over modifiers and propose a causal model.

3 A Causal Graphical Model

The technical contribution of this work is a novel causal graphical model that jointly represents the relationship between the grammatical gender of a noun, its meaning, and descriptors. This model is depicted in Figure 1. If properly estimated, the model should enable us to measure the *causal* effect of grammatical gender on adjective choice in language. We first develop the necessary notation.

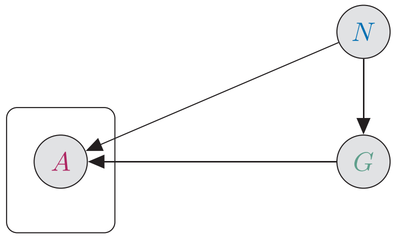

Figure 1: Causal graphical model relating noun semantics, gender, and adjective choice. The neo-Whorfian hypothesis posits that a noun's gender *causally* influences adjective choice. Correctly evaluating this hypothesis must also account for the relationship between the noun's meaning and adjective choice.

Notation We follow several font and coloring conventions to make our notation easier to digest. All base sets will be uppercase and in calligraphic font, e.g., $\mathcal{X}$. Elements of $\mathcal{X}$ will be lowercase and italicized, e.g., $x \in \mathcal{X}$. Subsets (including submultisets) will be uppercase and unitalicized, e.g., $\mathrm{X} \subset \mathcal{X}$. Random variables that draw their values from $\mathcal{X}$ will be uppercase and italicized, e.g., $p(X = x)$. We will use three colors. Those objects that relate to nouns will be in blue, those objects that relate to adjectives will be in purple, and those objects that relate to gender will be in green.

3.1 The Model

We assume there exists a set of noun meanings $\mathcal{N}$. In this paper, we assume that such meanings are representable by column vectors in $\mathbb{R}^D$. We denote the elements of $\mathcal{N}$ as $\boldsymbol{n} \in \mathbb{R}^D$. Additionally, we assume there exists an alphabet of adjectives $\mathcal{A}$. We denote an element of $\mathcal{A}$ as a. Finally, we assume there exists a language-dependent set of genders $\mathcal{G}$. In Spanish, for instance, we would have $\mathcal{G} = \{\text{FEM}, \text{MSC}\}$ whereas in German $\mathcal{G} = \{\text{FEM}, \text{MSC}, \text{NEU}\}$. We denote elements of $\mathcal{G}$ as g.

We now develop a generative model of the subset of lexical semantics relating to adjective choice. We wish to generate a set of $|\mathcal{N}|$ nouns, each of which is modified by a multiset of adjectives. We can view this model as a partial generative model of a corpus where we focus on generating noun types and adjective tokens. Generation from the model proceeds as follows:

$$\begin{aligned}
\boldsymbol{n} &\sim p_N(\cdot) \\
&\quad \text{(sample a noun meaning } \boldsymbol{n}\text{)} \\
g_{\boldsymbol{n}} &\sim p_G(\cdot \mid \boldsymbol{n}) \\
&\quad \text{(sample the gender } g_{\boldsymbol{n}} \text{ assigned to } \boldsymbol{n}\text{)} \\
a_{\boldsymbol{n}} &\sim p_A(\cdot \mid \boldsymbol{n}, g_{\boldsymbol{n}}) \\
&\quad \text{(sample adjectives } a_{\boldsymbol{n}} \text{ that modify } \boldsymbol{n}\text{)}
\end{aligned}$$

In this formulation, N is a $\mathcal{N}$-valued random variable, G is a $\mathcal{G}$-valued random variable, and A is a $\mathcal{A}$-valued random variable.

Written as a probability distribution, we have

$$p(\{\mathrm{A}_{\boldsymbol{n}}\}, \{g_{\boldsymbol{n}}\}, \mathrm{N}) = \prod_{\boldsymbol{n} \in \mathrm{N}} \prod_{a \in \mathrm{A}_{\boldsymbol{n}}} p_A(a \mid \boldsymbol{n}, g_{\boldsymbol{n}})\, p_G(g_{\boldsymbol{n}} \mid \boldsymbol{n})\, p_N(\boldsymbol{n}) \tag{1}$$

where $\mathrm{N} \subset \mathcal{N}$ is a subset of the set of noun meanings and each $g_{\boldsymbol{n}} \in \mathcal{G}$ is the gender of $\boldsymbol{n}$, and each $\mathrm{A}_{\boldsymbol{n}} \subset \mathcal{A}$ is a multisubset of $\mathcal{A}$ that contains the observed adjectives that modify $\boldsymbol{n}$. This model is represented graphically in Figure 1, where the arrow from N to G represents the dependence of G on N as shown in the conditional probability distribution $p_G(g_{\boldsymbol{n}} \mid \boldsymbol{n})$, and the arrows from N and G to A represent the dependence of A on N and G, as shown in the conditional probability distribution $p_A(a \mid \boldsymbol{n}, g_{\boldsymbol{n}})$.

Importantly, our model *generates* the lexical semantics of noun types. This means that a sample from it generates a new noun, whose semantics we may never have seen before. If we are able to estimate such a model well, we can use the basics of causal inference to estimate the causal effect gender has on adjective usage. Specifically, as is clear from Figure 1, the only confounder between gender and adjective selection in our proposed model is the semantics of the noun.[1]

3.2 Intervention

Thus, to the extent that the modeler believes our model p is a reasonable generative model of lexical semantics, we apply Pearl's backdoor criterion to get a causal effect (Pearl, 1993). One does so by applying the do-calculus, which results in the following gender-specific distribution over adjectives

$$p(a \mid \mathrm{do}(G = g)) = \sum_{\boldsymbol{n} \in \mathcal{N}} p_A(a \mid G = g, \boldsymbol{n})\, p_N(\boldsymbol{n}) \tag{2}$$

where for simplicity, $\mathcal{N}$ is assumed to be at most countable, despite being a subset of $\mathbb{R}^D$. We are now interested in using $p(a \mid \mathrm{do}(G = g))$ to measure the extent to which the grammatical gender of a noun influences which adjectives are used to modify that noun. In particular, we aim to measure how different the adjective choice would be if the noun had a different grammatical gender. Because $p(a \mid \mathrm{do}(G = g))$ is a distribution over $\mathcal{A}$, we measure the causal effect by the weighted Jensen–Shannon divergence (Lin, 1991), which we define as

$$\mathrm{JS}_\pi(p_1 \,||\, p_2) \stackrel{\text{def}}{=} \pi_1 \mathrm{KL}(p_1 \,||\, m) + \pi_2 \mathrm{KL}(p_2 \,||\, m) \tag{3}$$

where $\pi_1, \pi_2 \geq 0$, $\pi_1 + \pi_2 = 1$ and $m = \pi_1 p_1 + \pi_2 p_2$ is a convex combination of p_1 and p_2 weighted according to π.[2] Further, we note that the weighted Jensen–Shannon divergence is related to a specific mutual information between two random variables. We make this relationship formal in the following proposition.

Proposition 1. *Let A and G be $\mathcal{A}$-valued and $\mathcal{G}$-valued random variables, respectively. Further assume they are jointly distributed according to $p(a \mid \mathrm{do}(G = g)) p_G(g)$. Then,*

$$\mathrm{JS}_{p_G}\left(\left\{p(\cdot \mid \mathrm{do}(G = g))\right\}\right) = \mathrm{MI}_{\mathrm{do}}(A; G) \tag{4}$$

where $\mathrm{MI}_{\mathrm{do}}(A; G)$ is the mutual information computed under the joint distribution $p(a \mid \mathrm{do}(G = g))\, p_G(g)$.

Proof: See Appendix A for a proof. ■

[1] Sentential context can also influence adjective usage, e.g., the probability distribution over adjectives describing the noun *bagel* might differ between the sentences *After the flood, the rat discovered a ___ bagel dissolving in the sewer.*, and *She was craving a ___ bagel.* Our model does not aim to account for such contextual effects.

[2] The Jensen–Shannon divergence can also be generalized to operate on N distributions as $\mathrm{JS}_\pi(p_1, \ldots, p_N) = \sum_{n=1}^N \pi_n \mathrm{KL}(p_n \,||\, m)$, where $\sum_{n=1}^N \pi_n = 1, \pi_n \geq 0,\ \forall n \in [N]$, and $m = \sum_{n=1}^N \pi_n p_n$.

Relating the weighted Jensen–Shannon divergence to a specific mutual information provides a clear interpretation. This measure explains in bits how much the entropy of the language's distribution over adjectives is reduced when the grammatical gender of the noun being modified is known at the time of the adjective choice. For instance, if the language's distribution over adjectives has an entropy $\mathrm{H}(A)$ of 10 bits and the mutual information $\mathrm{MI}(A; G) \stackrel{\text{def}}{=} \mathrm{H}(A) - \mathrm{H}(A \mid G)$ is 1 bit, then knowing the gender allows us to reduce the uncertainty over which adjectives modify the nouns to $\mathrm{H}(A \mid G) = 9$ bits. However, the reduced uncertainty measured by $\mathrm{MI}(A; G)$ is purely associational; we cannot conclude that the gender of the noun actually causes the change in adjective distribution. Such a change could also be attributed to a confounding factor (like noun meaning). On the other hand, $\mathrm{MI}_{\mathrm{do}}(A; G) \stackrel{\text{def}}{=} \mathrm{H}_{\mathrm{do}}(A) - \mathrm{H}_{\mathrm{do}}(A \mid G)$ represents the amount of uncertainty in the adjective distribution *causally* reduced by the gender random variable. Intuitively, we can reason about $\mathrm{H}_{\mathrm{do}}(A)$ and $\mathrm{H}_{\mathrm{do}}(A \mid G)$ as the uncertainty of the adjective distribution in a world where we can counterfactually imagine that all nouns have the same gender g, and thus by setting all else equal, isolate the effect of knowing gender alone on the uncertainty of the adjective distribution. For a formal definition of $\mathrm{H}_{\mathrm{do}}(A)$ and $\mathrm{H}_{\mathrm{do}}(A \mid G)$, see Appendix A.

3.3 Intuition

We now explain the intuition behind the mutual information $\mathrm{MI}_{\mathrm{do}}(A; G)$.

Case 1: No edge from N to G. First, consider the case when there is no edge from N to G, indicating that there is no causal relationship between a noun's meaning and its grammatical gender. Under this condition, we have $p_G(g \mid \boldsymbol{n}) = p_G(g)\ \forall g \in \mathcal{G}, \boldsymbol{n} \in \boldsymbol{\mathcal{N}}$. Consequently, the interaction between the grammatical gender of a noun and the adjectives used to describe this noun is not mitigated by the meaning of this noun, and thus, $\mathrm{MI}_{\mathrm{do}}(A; G) = \mathrm{MI}(A; G)$.

Case 2: No edge from G to A. Second, consider the case when there is no edge from G to A, indicating that there is no causal relationship between a noun's gender and its adjective distribution. In this case, $\mathrm{MI}_{\mathrm{do}}(A; G) = 0$. Further, we can show that $\mathrm{MI}_{\mathrm{do}}(A; G) = 0$ if and only if the edge from G to A does not exist.

Case 3: All edges. Finally, when both edges from G to N and G to A exist, $\mathrm{MI}_{\mathrm{do}}(A; G)$ can vary. In particular, $\mathrm{MI}_{\mathrm{do}}(A; G)$ is non-negative (and indeed non-zero by case 2). However, we know of no relationship between $\mathrm{MI}_{\mathrm{do}}(A; G)$ and $\mathrm{MI}(A; G)$. In this case, the strength of the relationship between a noun's grammatical gender and adjective choice is regulated by the meaning.

3.4 Parameterization

We now discuss the parameterization of the conditional distributions given in §3.1: adjectives (p_A), gender (p_G), and vector representations of nouns (p_N). We model p_A using a logistic classifier where the probability of each adjective a is predicted given the column vector $[\mathbf{e}(a); \boldsymbol{n}; \mathbf{e}(g)]$, which is a stacking of a column vector representation $\mathbf{e}(a)$ of an adjective a, the meaning representation of the noun $\boldsymbol{n}$, and a column vector representation of gender g, respectively. The classifier's functional form is given as

$$p_A(a \mid g, \boldsymbol{n}) = \frac{\exp\left(\boldsymbol{w}^\top \tanh \boldsymbol{W}\left[\mathbf{e}(a); \boldsymbol{n}; \mathbf{e}(g)\right]\right)}{\sum_{b \in \mathcal{A}} \exp\left(\boldsymbol{w}^\top \tanh \boldsymbol{W}\left[\mathbf{e}(b); \boldsymbol{n}; \mathbf{e}(g)\right]\right)} \tag{5}$$

where the parameters $\boldsymbol{W}$ and $\boldsymbol{w}$ denote the weight matrix and weight column vector, respectively. We note that Equation 5 gives the probability of a single $a \in \mathrm{A}_{\boldsymbol{n}}$ that co-occurs with $\boldsymbol{n}$. The probability of the set $\mathrm{A}_{\boldsymbol{n}}$ is the product of generating each adjective independently. While $\mathbf{e}(a)$ and $\mathbf{e}(g)$ could be trainable parameters, for simplicity, we fix $\mathbf{e}(a)$ to be standard word2vec representations and $\mathbf{e}(g)$ to be a one-hot encoding with dimension $|\mathcal{G}|$. Representations for $\boldsymbol{n}$ are pretrained according to methods described in §4.2.

We opt to model $p_G(g \mid \boldsymbol{n})$ and $p_N(\boldsymbol{n})$ as the empirical distribution of nouns in the corpus.

4 Experimental Setup

In this section, we describe the data used in our experiments, and how we estimate non-contextual word representations as a proxy for a noun's lexical semantics.

	DE	ES	HE	PL	PT
word2vec					
# noun types	932	953	814	891	929
# adj types	109,549	61,839	29,855	42,271	30,004
# noun-adj types	486,647	581,589	208,202	223,774	176,995
# noun-adj tokens	5,966,400	7,523,601	2,413,546	4,040,464	1,543,563
WordNet					
# noun types	437	773	391	450	630
# adj types	78,585	58,536	26,278	38,427	26,112
# noun-adj types	272,511	513,905	145,542	178,049	134,923
# noun-adj tokens	3,606,909	6,912,761	1,978,561	3,493,547	1,243,506

Table 1: Data statistics in our Wikipedia corpora with retrieved word2vec and WordNet representations.

4.1 Data

We gather data in five languages that exhibit grammatical gender agreement: German, Hebrew, Polish, Portuguese, and Spanish. This is certainly not a representative sample of the subset of the world's languages that exhibit grammatical gender, but we are limited by the need for a large corpus to estimate a proxy for lexical meaning. Hebrew, Portuguese, and Spanish distinguish between two grammatical genders (masculine and feminine), while German and Polish distinguish between three genders (masculine, feminine, and neuter).[3]

We use the Wikipedia dump dated August 2022 to create a corpus for each of the five languages,[4] and preprocess the corpora with the Stanza library (Qi et al., 2020).[5] Specifically, we tokenize the raw text, dependency parse the tokenized text, lemmatize the data, extract lemmatized noun–adjective pairs based on an `amod` dependency label, and finally filter these pairs such that only those for inanimate nouns remain. To determine which nouns are inanimate, we use the NorthEuraLex dataset, a curated list of cross-linguistically common inanimate nouns (Dellert et al., 2020). Table 1 shows the counts for the remaining tokens for all analyzed languages for which we retrieved word representations. Next, we describe the procedure for computing the non-contextual word representations.

[3]Polish also includes an animacy distinction for masculine nouns.

[4]`https://dumps.wikimedia.org/`.

[5]`https://stanfordnlp.github.io/stanza/`.

4.2 Non-contextual Word Representations

The model described in §3 relies on a representation of nominal lexical semantics—specifically, a representation independent (in the probabilistic sense) of the distributional properties of the noun.[6]

Word2vec. We train word2vec (Mikolov et al., 2013) on modified Wikipedia corpora. As found in Omrani Sabbaghi and Caliskan (2022), word representations learn the association between a noun and its grammatical gender in grammatically gendered languages. Thus, we first lemmatize the corpus with Stanza as discussed above. This step should remove any spurious correlations between a noun's morphology and its meaning. Second, we remove all adjectives from the corpora. Because our goal is to predict the distribution over adjectives *from* a noun's lexical semantic representation, that distribution should not, itself, be encoded in the semantic representation. We construct representations of length 200 through the continuous skip-gram model with negative sampling with 10 samples using the implementation from `gensim`.[7] We train these non-contextual word representations on the Wikipedia data described above. We ignore all words with a frequency below five and use a symmetric context window size of five.

WordNet-based Representations. In addition to those representations derived from word2vec,

[6]We describe two ways in which we construct such representations. Similar to this approach, Kann (2019) trains a classifier to predict gender from word representations trained on a lemmatized corpus.

[7]`https://radimrehurek.com/gensim/models/word2vec.html`.

WordNet	Words	Senses	Synsets
ODENet 1.4 (de)	120,107	144,488	36,268
OpenWN-PT (pt)	54,932	74,012	43,895
plWordNet (pl)	45,456	52,736	33,826
MCR (es)	37,203	57,764	38,512
Hebrew WordNet (he)	5,379	6,872	5,448

Table 2: Summary statistics on the WordNets used for training representations in each language.

we also derive lexical representations using WordNet (Miller, 1994). Because WordNet is a lexical database that groups words into sets of synonyms (synsets) and links synsets together by their conceptual, semantic, and lexical relations, representations of meaning based on WordNet are unaffected by biases that might be encoded in a training corpus of natural language. Following Saedi et al. (2018), we create word representations by constructing an adjacency matrix of WordNet's semantic relations (e.g., hypernymy, meronymy) between words and compressing this matrix to have a dimensionality of 200 for each of the languages in this study: German, Hebrew, Spanish, Polish, and Portuguese (Siegel and Bond, 2021; Ordan and Wintner, 2007; Gonzalez-Agirre and Rigau, 2013; Piasecki et al., 2009; de Paiva and Rademaker, 2012). We access and process these WordNets using the Open Multilingual WordNet (Bond and Paik, 2012). We report statistics on these WordNets in Table 2.

Evaluating the Representations. We now discuss how we validate our lexical representations. Because we construct the word2vec representations using modified corpora, it is reasonable to fear that those modifications would hinder the representations' ability to encode an adequate approximation to nominal lexical semantics. Thus, for each language, we evaluate the quality of the learned representations by calculating the Spearman correlation coefficient of the cosine similarity between representations and the human-annotated similarity scores of word pairs in the SimLex family of datasets (Hill et al., 2015; Leviant and Reichart, 2015; Vulić et al., 2020). A higher correlation indicates a better representation of semantic similarity. We report the Spearman correlation of the representations for each language in Table 3. Especially for representations generated using WordNet for languages

	WordNet		word2vec	
Lang.	ρ	% of eval set	ρ	% of eval set
DE	0.360	86.9%	0.380	92.2%
ES	0.234	71.8%	0.419	89.3%
HE	0.104	11.6%	0.460	59.6%
PL	0.092	49.9%	0.418	76.5%
PT	0.283	94.7%	0.308	94.5%

Table 3: Spearman's ρ correlation coefficient between judgments in similarity datasets and representation cosine similarity for each language for both WordNet and word2vec representations.

with sparsely populated WordNets (see Table 2), the representational power is relatively low (as measured by the Spearman correlation), which may influence conclusions of downstream results for these languages. We note that if the representations are very bad—such that gender is completely unpredictable from the noun meaning representation and $p_G(g_{\boldsymbol{n}} \mid \boldsymbol{n}) = p_G(g_{\boldsymbol{n}})$—then $\mathrm{MI}(A; G) = \mathrm{MI}_{\mathrm{do}}(A; G)$ because the edge in the graphical model from N to G is effectively removed.

5 Methodology

We now outline the methodology of our study, starting with parameter estimation of the model in §5.1 and plug-in estimation of $\mathrm{MI}(A; G)$ in §5.2. We conduct two experiments: First, for a point of comparison, we replicate Williams et al.'s (2021) study to estimate $\mathrm{MI}(A; G)$ for each of the five languages (§5.3); and second, we produce a causal analog of Williams et al. (2021) (§5.4). Using the notation of §3, in the second experiment, we estimate $\mathrm{MI}_{\mathrm{do}}(A; G)$ for each of the five languages. Finally, in §5.5, we discuss our permutation testing methodology.

5.1 Parameter Estimation

To estimate the parameters of the graphical model given in Figure 1, we perform regularized maximum-likelihood estimation. Specifically, we maximize the likelihood the model assigns to a training set $\mathrm{D}_{\mathrm{trn}} = \{(\mathrm{A}_n, g_n, \boldsymbol{n}_n)\}_{n=1}^{N}$ where each distinct $\boldsymbol{n}_n$ occurs at most once. The log-likelihood is

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{a \in \mathrm{A}_n} \log p_A(a \mid g_n, \boldsymbol{n}_n) \qquad (6)$$

where $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{W}\}$. We define p_A using a multilayer perceptron (MLP) with the rectified each of the five languages for a maximum linear unit (ReLU; Nair and Hinton, 2010) and a final softmax layer. We estimate p_N and p_G from the empirical distributions derived from the corpus. We further apply L_1-regularization to impose sparsity and L_2-regularization to prevent a representation's dimension from dominating the model's predictions. The regularization coefficients are each set to 0.001. We train our models for each of the five languages for a maximum of 100 epochs using the Adam optimizer (Kingma and Ba, 2015) to predict the adjective given its representation, a noun's gender, and representation.

5.2 Plug-in Estimation of $\mathrm{MI}(A; G)$

The first estimator of $\mathrm{MI}(A; G)$ is the plug-in estimator considered by Williams et al. (2021). In this case, we compute the maximum-likelihood estimate (MLE) of the marginal $p(a, g)$ and plug it into the formula for mutual information:[8]

$$\mathrm{MI}(A; G) = \sum_{a \in \mathcal{A}} \sum_{g \in \mathcal{G}} p(a, g) \log \frac{p(a, g)}{p(g)p(a)} \tag{7}$$

Following Williams et al. (2021), we use empirical probabilities as the plug-in estimates.

5.3 Model-based Estimation of $\mathrm{MI}(A; G)$

In the first experiment, we replicate Williams et al.'s findings on different data and with a different method. Let $p(a, g, \boldsymbol{n}) = p_A(a \mid g, \boldsymbol{n}) p_G(g \mid \boldsymbol{n}) p_N(\boldsymbol{n})$ be an estimated model that factorizes according to the graph given in Figure 1, and let $\tilde{\mathrm{N}}$ be a set of gender–noun pairs where the nouns are *distinct* from those in the test set, $\mathrm{D}_{\mathrm{tst}}$. Let h and $\boldsymbol{m}$ be gender–noun pairs from this test set. Using $\tilde{\mathrm{N}}$, consider the following approximate marginal:

$$\tilde{p}(a, g) = \frac{1}{|\tilde{\mathrm{N}}|} \sum_{(h, \boldsymbol{m}) \in \tilde{\mathrm{N}}} p_A(a \mid h, \boldsymbol{m}) \mathbb{1}\{g = h\} \tag{8}$$

We then plug $\tilde{p}(a, g)$ into the formula for correlational $\mathrm{MI}(A; G)$ defined in Equation 7.

[8]While we opt for the MLE approach to maintain consistency with Williams et al. (2021), we note that the alternative entropy estimators instead might have yielded lower mutual information estimates, as suggested by Arora et al. (2022).

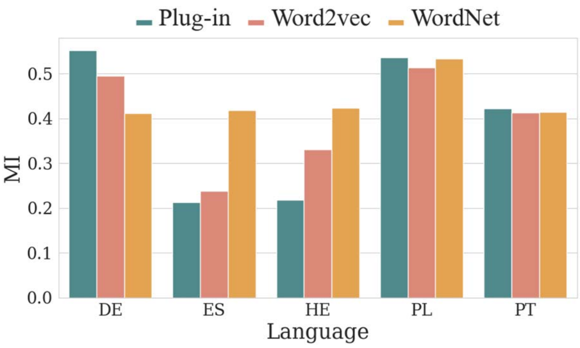

Figure 2: Results for the plug-in estimation of $\mathrm{MI}(A; G)$ and model-based estimations for $\mathrm{MI}(A; G)$.

5.4 Model-based Estimation of $\mathrm{MI}_{\mathrm{do}}(A; G)$

Here, in contrast to §5.3, we are interested in *causal* mutual information, which we take to be the mutual information as defined under $p(a \mid \mathrm{do}\,(G = g))p(g)$. We approximate the marginal $p(g)$ using a maximum-likelihood estimate on $\mathrm{D}_{\mathrm{trn}}$. We use $\tilde{\mathrm{N}}_g$, a set of gender–noun pairs *distinct* from those in $\mathrm{D}_{\mathrm{tst}}$ with a fixed gender g to compute the following estimate of the intervention distribution

$$\tilde{p}(a \mid \mathrm{do}(G = g)) = \frac{1}{|\tilde{\mathrm{N}}_g|} \sum_{(g, \boldsymbol{m}) \in \tilde{\mathrm{N}}_g} p_A(a \mid g, \boldsymbol{m}) \tag{9}$$

using the parameters of the model $p_A(a \mid G = g, \boldsymbol{n})$ estimated as described in §5.1. We use a permutation test to determine whether the estimate is significantly different than zero, as described in §5.5.

5.5 Permutation Testing

To do this, we train a model from scratch using 5-fold cross-validation on a subset of 100 adjectives to estimate $p_A(a \mid \boldsymbol{n}, g)$ with a random permutation of the gender labels and use that model to compute the pair-wise mutual information estimates between adjective distributions on the test set as described earlier for $k = 2,000$ times (a total of 10,000 runs). We design and run a permutation test to determine whether the mutual information between the adjective distributions conditioned on different genders is equal to the mutual information between the adjective distributions from a model trained on perturbed gender labels. We determine the significance of our result by evaluating the proportion of times that

Language	word2vec Model-based $\mathrm{MI}(A;G)$	word2vec Model-based $\mathrm{MI}_{\mathrm{do}}(A;G)$	word2vec Mean diff. Perturbed	WordNet Model-based $\mathrm{MI}(A;G)$	WordNet Model-based $\mathrm{MI}_{\mathrm{do}}(A;G)$	WordNet Mean diff. Perturbed
DE	0.526	1.24e−4	3.12e−4*	0.412	2.17e−5	1.03e−2*
ES	0.238	4.60e−5	4.85e−4*	0.418	1.24e−5	1.77e−3*
HE	0.331	8.03e−4	4.70e−3*	0.423	1.43e−5	1.11e−3*
PL	0.545	1.65e−4	6.67e−4*	0.533	8.68e−7	1.37e−4*
PT	0.413	1.72e−4	6.31e−3*	0.414	8.80e−5	1.76e−3*

Table 4: Results for the plug-in estimation of $\mathrm{MI}(A;G)$, model-based estimation for $\mathrm{MI}(A;G)$, and model-based estimation of $\mathrm{MI}_{\mathrm{do}}(A;G)$, mean difference between the model-based estimation of $\mathrm{MI}_{\mathrm{do}}(A;G)$ and a perturbed model with random gender labels for the causal model trained with word2vec and WordNet representations. Significant differences (p-value $< .05$) according to the permutation test are marked with an asterisk.

$\mathrm{MI}_{\mathrm{do}}(A;G)$, as computed using the non-permuted training set, is lower than one computed using randomly permuted genders during training. We choose the standard significance level of $\alpha = .05$; that is, when we observe a p-value lower than .05, we reject the null hypothesis, which posits no difference in mutual information between models trained on original and perturbed gender labels.

6 Results

First, we validate our model by comparing the model-based estimation of $\mathrm{MI}(A;G)$ to the method presented in Williams et al. (2021), the plug-in estimation of $\mathrm{MI}(A;G)$. Then, we employ our causal graphical model to investigate whether there is evidence for the neo-Whorfian claim that the grammatical gender of a noun influences the adjective chosen to modify this noun, even when we control for the meaning of the noun.

We first validate our model by comparing its results to Williams et al.'s (2021) plug-in estimate of $\mathrm{MI}(A;G)$. If the results of both of these estimates are comparable, we have evidence that our model indeed captures the relation between grammatical gender and adjective choice. We present the results in Figure 2. We observe a substantial relationship between grammatical gender and adjective usage based on the plug-in and model-based $\mathrm{MI}(A;G)$ estimates replicating the results of Williams et al. (2021). The estimates of the model-based $\mathrm{MI}(A;G)$ computed using both word2vec and WordNet representations, and the plug-in $\mathrm{MI}(A;G)$ lie between 0.2 and 0.5, with the estimates of the model-based approach being consistently higher (with the exception of German) than the estimates of the plug-in $\mathrm{MI}(A;G)$. Thus, the non-zero estimates of the model-based $\mathrm{MI}(A;G)$ indicate that some relationship exists between a noun's grammatical gender and adjective usage.

Given the above result, we are interested in whether the strength of this relationship is mitigated when controlling for the meaning of a noun. We present the estimates of the model-based $\mathrm{MI}_{\mathrm{do}}(A;G)$ in Table 4 and compare them to the model-based estimates of the $\mathrm{MI}(A;G)$. While we observe evidence for the influence of grammatical gender on adjective choice in a non-causal setup based on $\mathrm{MI}(A;G)$, this relationship shrinks to close to 0 when we control for noun meaning in our causal model trained using both word2vec and WordNet representations.

For completeness, we test for the presence of a difference between the size of the $\mathrm{MI}_{\mathrm{do}}(A;G)$ of our model and a model trained on randomly perturbed gender labels based on a subset of adjectives. We reject the null hypothesis that the distributions are exactly the same for all languages and representations' settings.

7 Discussion

Evidence Against the neo-Whorfian Hypothesis. We find that the interaction between the grammatical gender of inanimate nouns and the adjectives used to describe those nouns all but disappears when controlling for the meaning of those nouns, for all five analyzed gendered languages. While the order of magnitude of $\mathrm{MI}_{\mathrm{do}}(A;G)$ measured with our model is significantly different from that of a model trained on random gender labels, it remains minuscule in absolute terms. This minor difference points towards the absence of a meaningful causal relationship between a noun's gender and its modifiers in the languages studied. Thus, we provide an additional piece of evidence against the neo-Whorfian hypothesis in noun–adjective patterns.

A Possible Weakening of the Neo-Whorfian Hypothesis. Although the size of the overall effect is small, it is possible that the effect of gender on adjective choice is stronger for some words than others. Future work could explore whether there is evidence of a noticeable effect of gender on adjective choice for a more restricted set of inanimate nouns, e.g., referring to artifacts or body parts. Such evidence could perhaps support a weakened version of the neo-Whorfian hypothesis.

Comparing Results Between Word2vec and WordNet. Our results hold for both of the word representation conditions, word2vec, and WordNet. Notably, in comparing the two, we find that using WordNet representations consistently results in a lower $\mathrm{MI}_{\mathrm{do}}(A; G)$ than word2vec for all languages analyzed in this study. One possible explanation for this difference is that, despite our efforts to make non-contextual word2vec representations, these word2vec representations may still encode some signal regarding gender from the remaining context (such as verb choice or adjacent gendered pronouns in the corpora). If these word2vec representations contain unwanted context-based gender information in addition to the noun meaning, it could result in overestimating $\mathrm{MI}_{\mathrm{do}}(A; G)$. Furthermore, since WordNet representations are created independently from any context within a corpus, they should not contain any signal related to grammatical gender, which may therefore be reflected in the consistently lower $\mathrm{MI}_{\mathrm{do}}(A; G)$.

Design Choices and Limitations. We note several choices in the experimental setup which may influence this analysis. First, while we experiment with NorthEuraLex, which furnishes us with a list of inanimate nouns, the dataset excludes rarer nouns for which an effect might be observed.[9] Second, while non-contextual word representations are the current de facto proxy for lexical semantics, they remain a proxy and are fundamentally limited. Furthermore, in our effort to estimate word2vec representations for noun meaning without encoding gender-based context, we chose to remove some words in the context but not others. Specifically, while we remove adjectives which may carry signals of gender from the training corpora, we do not remove other parts of speech (e.g., verbs) under the reasoning that removing them may damage the training corpora too much for word2vec to effectively learn noun meanings.[10] Future work can also explore improved representation methods for noun meaning. For example, Recski et al. (2016) find that creating non-contextual word representations using a combination of word2vec, WordNet, and concept dictionaries can yield a better representation of meaning (i.e., achieving state-of-the-art correlation with the human-annotated similarity scores). Third, the corpus choice (and subsequently the noun–adjective pairs on which we conduct our analysis) may factor into the results. It is possible that when applied to other corpora (e.g., more colloquial ones like Reddit), this method may yield different results. Fourth, the choice of languages analyzed further limits this study to languages with up to three gender classes. Future work can investigate languages with more complex gender systems. Finally, our modeling approach assumes that the gender of a noun is influenced solely by its meaning. However, prior work has indicated that there are other factors that influence the grammatical gender of nouns such as their phonology and morphology (Corbett, 1991). Therefore, future work should investigate more complex graphical models in order to account for other confounding factors.

8 Conclusion

In this paper, we introduce a causal graphical model that jointly models the interactions between a noun's grammatical gender, its meaning, and adjective choice. We employ our model on five

[9] We note that the original laboratory experiments taken to be as evidence for the neo-Whorfian hypothesis (Boroditsky and Schimdt, 2000; Semenuks et al., 2017) also only used high-frequency nouns. Moreover, if an effect were observed mainly with respect to low-frequency nouns, this would further weaken the neo-Whorfian hypothesis.

[10] Verbs may carry less signal for gender regardless. For example, Hoyle et al. (2019) find fewer significant differences in the usage of verbs than of adjectives towards people, and Williams et al. (2021) also report that verbs yielded smaller gender effects than adjectives.

languages that exhibit grammatical gender to investigate the influence of nouns' gender on the adjectives chosen to describe those nouns. Replicating the findings of Williams et al. (2021), we find a substantial correlation between grammatical gender and adjective choice. However, taking advantage of our causal perspective, we show that when controlling for a noun's meaning, the effect of gender on adjective choice is marginal. Thus, we provide further evidence against the neo-Whorfian hypothesis.

Code Release

The code and data necessary to replicate our empirical findings may be found at `https://github.com/rycolab/neo-whorf`.

Acknowledgments

This work is mostly funded by Independent Research Fund Denmark under grant agreement number 9130-00092B. Thanks to Nadav Borenstein for proofreading an earlier manuscript draft.

References

Alexandra Y. Aikhenvald. 2000. *Classifiers: A Typology of Noun Categorization Devices*. Oxford University Press UK.

A. T. Aksenov. 1984. K probleme èkstralingvističeskoj motivacii grammatičeskoj kategorii roda [On extralinguistic motivation of the grammatical category of gender]. *Voprosy Jazykoznanija 33 (1)*, pages 14–25.

D. W. Arnott. 1967. Some reflections on the content of individual classes in fula and tiv. In *La classification nominale dans les langues négro-africaines*, pages 45–74, Paris. Éditions du Centre national de la recherche scientifique.

Aryaman Arora, Clara Meister, and Ryan Cotterell. 2022. Estimating the entropy of linguistic distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Dublin, Ireland. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.acl-short.20`

Leonard Bloomfield. 1935. *Language*, Unwin University Books edition. G. Allen & Unwin, Ltd, London.

Francis Bond and Kyonghee Paik. 2012. A survey of WordNets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

Lera Boroditsky. 2003. Linguistic relativity. In *Encyclopedia of Cognitive Science*. Wiley Online Library.

Lera Boroditsky and Lauren A. Schmidt. 2000. Sex, syntax, and semantics. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 22.

J. O. Breedveld. 1995. *Form and Meaning in Fulfulde: A Morphophonological Study of Maasinankoore*. Research School CNWS, Leiden.

Greville G. Corbett. 1991. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press. `https://doi.org/10.1017/CBO9781139166119`

Greville G. Corbett. 2013a. Number of genders (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Greville G. Corbett. 2013b. Systems of gender assignment (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Greville G. Corbett. 2015. *Gender Typology*, pages 87–130. De Gruyter Mouton Berlin & Boston, Berlin, Boston.

Greville G. Corbett and Norman M. Fraser. 2000. Gender assignment: A typology and model. In *Systems of Nominal Classification*. Cambridge University Press Cambridge.

Valeria de Paiva and Alexandre Rademaker. 2012. Revisiting a Brazilian WordNet. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2020. Northeuralex: A wide-coverage lexical database of northern eurasia. *Language Resources and Evaluation*, 54(1):273–301. https://doi.org/10.1007/s10579-019-09480-6, PubMed: 32214931

Harry E. Foundalis. 2002. Evolution of gender in Indo-European languages. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 24.

Aitor Gonzalez-Agirre and German Rigau. 2013. Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual central repository. *Linguamática*, 5(1):13–28.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695. https://doi.org/10.1162/COLI_a_00237

Charles F. Hockett. 1958. *A Course in Modern Linguistics*. The Macmillan Company, New York. https://doi.org/10.1111/j.1467-1770.1958.tb00870.x

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1167

Katharina Kann. 2019. Grammatical gender, neo-Whorfianism, and word embeddings: A data-driven approach to linguistic relativity.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*. San Diego, CA.

A. I. Koval'. 1979. O znachenii morfologicheskogo pokazatelja klassa v fula. In N. V. Oxotina, editor, *Morfonologija i morfologija klassov slov v jazykax Afriki*, pages 5–100. Nauka, Moscow.

Ruth Kramer. 2015. *The Morphosyntax of Gender*, volume 58. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199679935.001.0001

Ruth Kramer. 2020. Grammatical gender: A close look at gender assignment across languages. *Annual Review of Linguistics*, 6:45–66. https://doi.org/10.1146/annurev-linguistics-011718-012450

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151. https://doi.org/10.1109/18.61115

John A. Lucy. 2016. Recent advances in the study of linguistic relativity in historical context: A critical assessment. *Language Learning*, 66(3):487–515. https://doi.org/10.1111/lang.12195

Anne Mickan, Maren Schiefke, and Anatol Stefanowitsch. 2014. Key is a llave is a schlussel: A failure to replicate an experiment from Boroditsky et al 2003. *Yearbook of the German Cognitive Linguistics Association*, 2(1):39. https://doi.org/10.1515/gcla-2014-0004

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8–11, 1994*. https://doi.org/10.3115/1075812.1075938

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 807–814, Madison, WI, USA.

Don Nelson. 2005. French gender assignment revisited. *Word*, 56(1):19–38. https://doi.org/10.1080/00437956.2005.11432551

Shiva Omrani Sabbaghi and Aylin Caliskan. 2022. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pages 518–531, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3514094.3534176

Noam Ordan and Shuly Wintner. 2007. Hebrew WordNet: A test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

E. M. Parker and R. J. Hayward. 1985. *An Afar-English-French Dictionary. With Grammatical Notes in English*. School of Oriental and African Studies, University of London, London.

Judea Pearl. 1993. [Bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269. https://doi.org/10.1214/ss/1177010894

Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wroclaw University of Technology Press. (ISBN 978-83-7493-476-3).

Keith Plaster and Maria Polinsky. 2007. Women are not dangerous things: Gender and categorization. *Harvard Working Papers in Linguistics*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Gábor Recski, Eszter Iklódi, Katalin Pajkossy, and András Kornai. 2016. Measuring semantic similarity of words using concept networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 193–200, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-1622

Chakaveh Saedi, António Branco, João António Rodrigues, and João Silva. 2018. WordNet embeddings. In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 122–131, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-3016

Saumya Sahai and Dravyansh Sharma. 2021. Predicting and explaining French grammatical gender. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 90–96, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.sigtyp-1.9

Steven Samuel, Geoff Cole, and Madeline J. Eacott. 2019. Grammatical gender and linguistic relativity: A systematic review. *Psychonomic Bulletin & Review*, 26(6):1767–1786. https://doi.org/10.3758/s13423-019-01652-3, PubMed: 31429058

Arturs Semenuks, Webb Phillips, Ioana Dalca, Cora Kim, and Lera Boroditsky. 2017. Effects of grammatical gender on object description. *Cognitive Science*.

Melanie Siegel and Francis Bond. 2021. OdeNet: Compiling a GermanWordNet from other resources. In *Proceedings of the 11th Global Wordnet Conference*, pages 192–198, University of South Africa (UNISA). Global Wordnet Association.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897. https://doi.org/10.1162/coli_a_00391

Benjamin Lee Whorf. 1956. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press.

Adina Williams, Damian Blasi, Lawrence Wolf-Sonkin, Hanna Wallach, and Ryan Cotterell. 2019. Quantifying the semantic core of gender systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5734–5739, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1577

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2021. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*, 9:139–159. https://doi.org/10.1162/tacl_a_00355

A Proof of Proposition 1

Proposition 1. *Let A and G be $\mathcal{A}$-valued and $\mathcal{G}$-valued random variables, respectively. Further assume they are jointly distributed according to $p(a \mid \mathrm{do}(G=g))p_G(g)$. Then,*

$$\mathrm{JS}_{p_G}\left(\left\{p(\cdot \mid \mathrm{do}(G=g))\right\}\right) = \mathrm{MI}_{\mathrm{do}}(A; G) \tag{4}$$

where $\mathrm{MI}_{\mathrm{do}}(A;G)$ is the mutual information computed under the joint distribution $p(a \mid \mathrm{do}(G = g))\, p_G(g)$.

Proposition 2. *First, define the following distribution $m(a) \stackrel{\text{def}}{=} \sum_{g \in \mathcal{G}} p_G(g)p(a \mid \mathrm{do}(G=g))$. Now, the result follows by algebraic manipulation*

$$\begin{aligned}
\mathrm{JS}_{p_G}\left(\left\{p(\cdot \mid \mathrm{do}(G=g))\right\}\right) &= \sum_{g\in\mathcal{G}} p_G(g)\,\mathrm{KL}\Big(p(\cdot \mid \mathrm{do}(G=g)) \mid\mid m\Big) && \text{(10a)}\\
&= \sum_{g\in\mathcal{G}} p_G(g) \sum_{a\in\mathcal{A}} p(a \mid \mathrm{do}(G=g))\Big(\log p(a \mid \mathrm{do}(G=g)) - \log m(a)\Big) && \text{(10b)}\\
&= \sum_{g\in\mathcal{G}} p_G(g) \sum_{a\in\mathcal{A}} p(a \mid \mathrm{do}(G=g)) \log p(a \mid \mathrm{do}(G=g)) - && \text{(10c)}\\
&\qquad \sum_{g\in\mathcal{G}} p_G(g) \sum_{a\in\mathcal{A}} p(a \mid \mathrm{do}(G=g)) \log m(a) &&\\
&= \underbrace{-\sum_{g\in\mathcal{G}} p_G(g)\mathrm{H}\left(A \mid \mathrm{do}(G=g)\right)}_{\stackrel{\text{def}}{=}\mathrm{H}_{\mathrm{do}}(A\mid G)} - \sum_{g\in\mathcal{G}} p_G(g) \sum_{a\in\mathcal{A}} p(a \mid \mathrm{do}(G=g)) \log m(a) && \text{(10d)}\\
&= -\mathrm{H}_{\mathrm{do}}\left(A \mid G\right) - \sum_{g\in\mathcal{G}} p_G(g) \sum_{a\in\mathcal{A}} p(a \mid \mathrm{do}(G=g)) \log m(a) && \text{(10e)}\\
&= -\mathrm{H}_{\mathrm{do}}\left(A \mid G\right) - \sum_{a\in\mathcal{A}} \sum_{g\in\mathcal{G}} p_G(g)p(a \mid \mathrm{do}(G=g)) \log m(a) && \text{(10f)}\\
&= -\mathrm{H}_{\mathrm{do}}\left(A \mid G\right) - \underbrace{\sum_{a\in\mathcal{A}} m(a) \log m(a)}_{\stackrel{\text{def}}{=}-\mathrm{H}_{\mathrm{do}}(A)} && \text{(10g)}\\
&= -\mathrm{H}_{\mathrm{do}}\left(A \mid G\right) + \mathrm{H}_{\mathrm{do}}(A) = \mathrm{H}_{\mathrm{do}}(A) - \mathrm{H}_{\mathrm{do}}\left(A \mid G\right) = \mathrm{MI}_{\mathrm{do}}(A;G) && \text{(10h)}
\end{aligned}$$