On the Relationships Between the Grammatical Genders of Inanimate Nouns and Their Co-Occurring Adjectives and Verbs

Adina Williams[*1] **Ryan Cotterell**[*,2,3] **Lawrence Wolf-Sonkin**[4]
Damián Blasi[5] **Hanna Wallach**[6]
[1]Facebook AI Research [2]ETH Zürich [3]University of Cambridge
[4]Johns Hopkins University [5]Universität Zürich [6]Microsoft Research
adinawilliams@fb.com ryan.cotterell@inf.ethz.ch lawrencews@jhu.edu
damian.blasi@uzh.ch wallach@microsoft.com

Abstract

We use large-scale corpora in six different gendered languages, along with tools from NLP and information theory, to test whether there is a relationship between the grammatical genders of inanimate nouns and the adjectives used to describe those nouns. For all six languages, we find that there is a statistically significant relationship. We also find that there are statistically significant relationships between the grammatical genders of inanimate nouns and the verbs that take those nouns as direct objects, as indirect objects, and as subjects. We defer deeper investigation of these relationships for future work.

1 Introduction

In many languages, nouns possess grammatical genders. When a noun refers to an animate object, its grammatical gender typically reflects the biological sex or gender identity of that object (Zubin and Köpcke, 1986; Corbett, 1991; Kramer, 2014). For example, in German, the word for a boss is grammatically feminine when it refers to a woman, but grammatically masculine when it refers to a man—*Chef**in*** and *Chef*, respectively. But inanimate nouns (i.e., nouns that refer to inanimate objects) also possess grammatical genders. Any German speaker will tell you that the word for a bridge, *Brücke*, is grammatically feminine, even though bridges have neither biological sexes nor gender identities. Historically, the grammatical genders of inanimate nouns have been considered more idiosyncratic and less meaningful than the grammatical genders of animate nouns (Brugmann, 1889; Bloomfield, 1933; Fox, 1990; Aikhenvald, 2000). However, some cognitive scientists have reopened this discussion by using laboratory experiments to test whether speakers of gendered languages reveal gender stereotypes (Sera et al., 1994)—for example, and most famously, when choosing adjectives to describe inanimate nouns (Boroditsky et al., 2003).

Although laboratory experiments are highly informative, they typically involve small sample sizes. In this paper, we therefore use large-scale corpora and tools from NLP and information theory to test whether there is a relationship between the grammatical genders of inanimate nouns and the adjectives used to describe those nouns. Specifically, we calculate the mutual information (MI)—a measure of the mutual statistical dependence between two random variables—between the grammatical genders of inanimate nouns and the adjectives that describe them (i.e., share a dependency arc labeled AMOD) using large-scale corpora in six different gendered languages (specifically, German, Italian, Polish, Portuguese, Russian, and Spanish). For all six languages, we find that the MI is statistically significant, meaning that there is a relationship.

We also test whether there are relationships between the grammatical genders of inanimate nouns and the verbs that take those nouns as direct objects, as indirect objects, and as subjects. For all six languages, we find that there are statistically significant relationships for the verbs that take those nouns as direct objects and as subjects. For five of the six languages, we also find that there is a statistically significant relationship for the verbs that take those nouns as indirect objects, but

[*]Equal contribution in this scientific whirlwind.

Transactions of the Association for Computational Linguistics, vol. 9, pp. 139–159, 2021. https://doi.org/10.1162/tacl_a_00355
Action Editor: Sebastian Padó. Submission batch: 3/2020; Revision batch: 7/2020; Published 3/2021.
© 2021 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

because of the small number of noun–verb pairs involved, we caution against reading too much into this finding.

To contextualize our findings, we test whether there are statistically significant relationships between the grammatical genders of inanimate nouns and the cases and numbers of these nouns. A priori, we do not expect to find statistically significant relationships, so these tests can be viewed as a baseline of sorts. As expected, for each of the six languages, there are no statistically significant relationships.

To provide further context, we also repeat all tests for *animate* nouns—a ''skyline'' of sorts—finding that for all six languages there is a statistically significant relationship between the grammatical genders of animate nouns and the adjectives used to describe those nouns. We also find that there are statistically significant relationships between the grammatical genders of animate nouns and the verbs that take those nouns as direct objects, as indirect objects, and as subjects. All of these relationships have effect sizes (operationalized as normalized MI values) that are larger than the effect sizes for inanimate nouns.

We emphasize that the practical significance and implications of our findings require deeper investigation. Most importantly, we do not investigate the characteristics of the relationships that we find. This means that we do not know whether these relationships are characterized by gender stereotypes, as argued by some cognitive scientists. We also do not engage with the ways that historical and sociopolitical factors affect the grammatical genders possessed by either animate or inanimate nouns (Fodor, 1959; Ibrahim, 2014).

2 Background

2.1 Grammatical Gender

Languages lie along a continuum with respect to whether nouns possess grammatical genders. Languages with no grammatical genders, like Turkish, lie on one end of this continuum, while languages with tens of gender-like classes, like Swahili (Corbett, 1991), lie on the other. In this paper, we focus on six different gendered languages for which large-scale corpora are readily available: German, Italian, Polish, Portuguese, Russian, and Spanish—all languages of Indo-European descent. Three of these languages (Italian, Portuguese, and Spanish) have two grammatical genders (masculine and feminine), while the other two have three grammatical genders (masculine, feminine, and neuter).

All six languages exhibit gender agreement, meaning that words are marked with morphological suffixes that reflect the grammatical genders of their surrounding nouns (Corbett, 2006). For example, consider the following translations of the sentence, *''The delicate fork is on the cold ground.''*

(1) ***Die*** *zierlich****e*** *Gabel steht auf* ***dem*** *kalt****en*** ***Boden***.
the.F.SG.NOM delicate.F.SG.NOM fork.F.SG.NOM stands on the.M.SG.DAT cold.M.SG.DAT ground.M.SG.DAT
The delicate fork is on the cold ground.

(2) ***El*** *tenedor delicad****o*** *está en* ***el*** *suel****o*** *frí****o***.
the.M.SG fork.M.SG delicate.M.SG is on the.M.SG ground.M.SG cold.M.SG
The delicate fork is on the cold ground.

Because the German word for a fork, *Gabel*, is grammatically feminine, the German translation uses the feminine determiner, *die*. Had *Gabel* been masculine, the German translation would have used the masculine determiner, *der*. Similarly, because the Spanish word for a fork, *tenedor*, is grammatically masculine, the Spanish translation uses the masculine determiner, *el*, instead of the feminine determiner, *la*. As we explain in Section 3, we lemmatize each corpus to ensure that our tests do not simply reflect the presence of gender agreement.

2.2 Grammatical Gender & Meaning

Although some scholars have described the grammatical genders possessed by inanimate nouns as ''creative'' and meaningful (Grimm, 1890; Wheeler, 1899), many scholars have considered them to be idiosyncratic (Brugmann, 1889; Bloomfield, 1933) or arbitrary (Maratsos, 1979, p. 317). In an overview of this work, Dye et al. (2017) wrote, ''As often as not, the languages of the world assign [inanimate] objects into seemingly arbitrary [classes]... William of

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

Ockham considered gender to be a meaningless, unnecessary aspect of language.'' Bloomfield (1933) shared this viewpoint, stating that ''[t]here seems to be no practical criterion by which the gender of a noun in German, French, or Latin [can] be determined.'' Indeed, adult language learners often have particular difficulty mastering the grammatical genders of inanimate nouns (Franceschina 2005, Ch. 4, DeKeyser 2005; Montrul et al. 2008), which suggests that their meanings are not straightforward.

Even if the grammatical genders possessed by inanimate nouns are meaningless, ample evidence suggests that gender-related information may affect cognitive processes (Sera et al., 1994; Cubelli et al., 2005, 2011; Kurinski and Sera, 2011; Boutonnet et al., 2012; Saalbach et al., 2012). Typologists and formal linguists have argued that grammatical genders are an important feature for morphosyntactic processes (Corbett, 1991, 2006; Harbour et al., 2008; Harbour, 2011; Kramer, 2014, 2015), while some cognitive scientists have shown that grammatical genders can be a perceptual cue—for example, human brain responses exhibit sensitivity to gender mismatches in several different languages (Osterhout and Mobley, 1995; Hagoort and Brown, 1999; Vigliocco et al., 2002; Wicha et al., 2003, 2004; Barber et al., 2004; Barber and Carreiras, 2005; Bañón et al., 2012; Caffarra et al., 2015), and the grammatical genders of determiners and adjectives can prime nouns (Bates et al., 1996; Akhutina et al., 1999; Friederici and Jacobsen, 1999). However, the precise nature of the relationship between grammatical gender and meaning remains an open research question.

In particular, the grammatical genders possessed by inanimate nouns might affect the ways that speakers of gendered languages conceptualize the objects referred to by those nouns (Jakobson, 1959; Clarke et al., 1981; Ervin-Tripp, 1962; Konishi, 1993; Sera et al., 1994, 2002; Vigliocco et al., 2005; Bassetti, 2007)—although we note that this viewpoint is somewhat contentious (Hofstätter, 1963; Bender et al., 2011; McWhorter, 2014). Neo-Whorfian cognitive scientists hold a particularly strong variant of this viewpoint, arguing that the grammatical genders possessed by inanimate nouns prompt speakers of gendered languages to rely on gender stereotypes when choosing adjectives to describe those nouns (Boroditsky and Schmidt, 2000; Boroditsky et al., 2002; Phillips and Boroditsky, 2003; Boroditsky, 2003; Boroditsky et al., 2003; Semenuks et al., 2017). Most famously, Boroditsky et al. (2003) claim to have conducted a laboratory experiment showing that speakers of German choose stereotypically feminine adjectives to describe, for example, bridges, while speakers of Spanish choose stereotypically masculine adjectives, reflecting 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. Boroditsky et al. (2003) took these findings to be a relatively strong confirmation of the existence of a stereotype effect—that is, that speakers of gendered languages reveal gender stereotypes when choosing adjectives to describe inanimate nouns. That said, the experiment has not gone unchallenged. Indeed, Mickan et al. (2014) reported two unsuccessful replication attempts.

2.3 Laboratory Experiments vs. Corpora

Traditionally, studies of grammatical gender and meaning have relied on laboratory experiments. This is for two reasons: 1) laboratory experiments can be tightly controlled, and 2) they enable scholars to measure speakers' immediate, real-time speech production. However, they also typically involve small sample sizes and, in many cases, somewhat artificial settings. In contrast, large-scale corpora of written text enable scholars to measure even relatively weak correlations via writers' text production in natural, albeit less tightly controlled, settings. They also facilitate the discovery of correlations that hold across languages with disparate histories, cultural contexts, and even gender systems. As a result, large-scale corpora have proven useful for studying a wide variety of language-related phenomena (e.g., Featherston and Sternefeld, 2007; Kennedy, 2014; Blasi et al., 2019).

In this paper, we assume that a writer's choice of words in written text is as informative as a speaker's choice of words in a laboratory experiment, despite the obvious differences between these settings. Consequently, we use large-scale corpora and tools from NLP and information theory, enabling us to test for the presence of even relatively weak relationships involving the grammatical genders of inanimate nouns across multiple different gendered languages. We

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

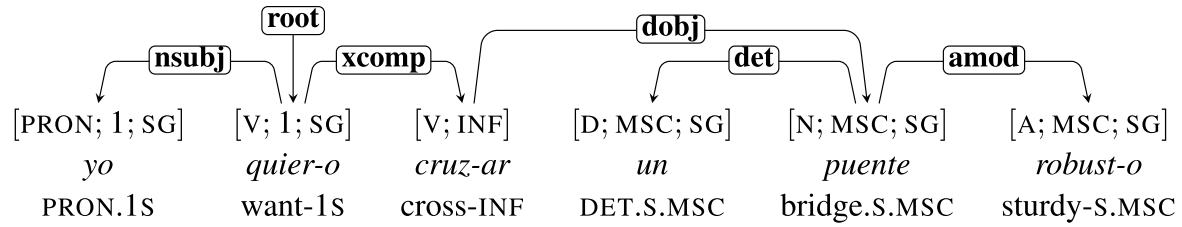

Figure 1: Dependency tree for the sentence, *''Yo quiero cruzar un puente robusto.''*

therefore argue that our findings complement, rather than supersede, laboratory experiments.

2.4 Related Work

Our paper is not the first to use large-scale corpora and tools from NLP to investigate gender and language. Many scholars have studied the ways that societal norms and stereotypes, including gender norms and stereotypes, can be reflected in representations of distributional semantics derived from large-scale corpora, such as word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Zhao et al., 2018). More recently, Williams et al. (2019) found that the grammatical genders of inanimate nouns in eighteen different languages were correlated with their lexical semantics. Dye et al. (2017) used tools from information theory to reject the idea that the grammatical genders of nouns separate those nouns into coherent categories, arguing instead that grammatical genders are only meaningful in that they systematically facilitate communication efficiency by reducing nominal entropy. Also relevant to our paper is the work of Kann (2019), who proposed a computational approach to testing whether there is a relationship between the grammatical genders of inanimate nouns and the words that co-occur with those nouns, operationalized via word embeddings. However, in contrast to our findings, they found no evidence for the presence of such a relationship. Finally, many scholars have proposed a variety of computational techniques for mitigating gender norms and stereotypes in a wide range of language-based applications (Dev and Phillips, 2019; Dinan et al., 2019; Ethayarajh et al., 2019; Hall Maudslay et al., 2019; Stanovsky et al., 2019; Tan and Celis, 2019; Zhou et al., 2019; Zmigrod et al., 2019).

3 Data Preparation

We use the May 2018 dump of Wikipedia to create a corpus for each of the six different gendered languages (i.e., German, Italian, Polish, Portuguese, Russian, and Spanish). Although Wikipedia is not the most representative data source, this choice yields language-specific corpora that are roughly parallel—that is, they refer to the same objects, but are not direct translations of each other (which could lead to artificial word choices). We use UDPipe to tokenize each corpus (Straka et al., 2016).

We dependency parse the corpus for each language using a language-specific dependency parser (Andor et al., 2016; Alberti et al., 2017), trained using Universal Dependencies treebanks (Nivre et al., 2017). An example dependency tree is shown in Figure 1. We then extract all noun–adjective pairs (dependency arcs labeled AMOD) and noun–verb pairs from each of the six corpora; for verbs, we extract three types of pairs, reflecting the fact that nouns can be direct objects (dependency arcs labeled DOBJ), indirect objects (dependency arcs labeled IOBJ), or subjects (dependency arcs labeled NSUBJ) of verbs. We discard all pairs that contain a noun that is not present in WordNet (Princeton University, 2010). We label the remaining nouns as ''animate'' or ''inanimate'' according to WordNet.

Next, we lemmatize all words (i.e., nouns, adjectives, and verbs). Each word is factored into a set of lexical features consisting of a lemma, or canonical morphological form, and a bundle of three morphological features corresponding to the grammatical gender, number, and case of that word. For example, the German word for a fork, *Gabel*, is grammatically feminine, singular, and genitive. For nouns, we discard the lemmas themselves and retain only the morphological

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

features; for adjectives and verbs, we retain the lemmas and discard the morphological features.

For adjectives and verbs, lemmatizing is especially important because it ensures that our tests do not simply reflect the presence of gender agreement, as we describe in Section 2.1. However, this means that if the lemmatizer fails, then our tests *may* simply reflect gender agreement despite our best efforts. To guard against this, we use a state-of-the-art lemmatizer (Müller et al., 2015), trained for each language using Universal Dependencies treebanks (Nivre et al., 2017). We expect that when the lemmatizer fails, the resulting lemmata will be low frequency. We try to exclude lemmatization failures from our calculations by discarding low-frequency lemmata. For each language, we rank the adjective lemmata by their token counts and retain only the highest-ranked lemmata (in rank order) that account for 90% of the adjective tokens; we then discard all noun–adjective pairs that do not contain one of these lemmata. We repeat the same process for verbs.

Finally, to ensure that our tests reflect the most salient relationships, we also discard low-frequency inanimate nouns and, separately, low-frequency animate nouns using the same process. We provide counts of the remaining noun–adjective and noun–verb pairs in Table 3 (for inanimate nouns) and Table 4 (for animate nouns).

4 Methodology

For each language $\ell \in \{de, it, pl, pt, ru, es\}$, we define $\mathcal{V}^{\ell}_{\text{ADJ}}$ to be the set of adjective lemmata represented in the noun–adjective pairs retained for that language as defined above. We similarly define $\mathcal{V}^{\ell}_{\text{VERB}}$ to be the set of verb lemmata represented in the noun–verb pairs retained for that language, as described above. We then define $\mathcal{V}^{\ell}_{\text{VERB-DOBJ}} \subset \mathcal{V}^{\ell}_{\text{VERB}}$, $\mathcal{V}^{\ell}_{\text{VERB-IOBJ}} \subset \mathcal{V}^{\ell}_{\text{VERB}}$, and $\mathcal{V}^{\ell}_{\text{VERB-SUBJ}} \subset \mathcal{V}^{\ell}_{\text{VERB}}$ to be the sets of verbs that take the nouns as direct objects, as indirect objects, and as subjects, respectively. We also define $\mathcal{G}^{\ell}$ to be the set of grammatical genders for that language (e.g., $\mathcal{G}^{es} = \{\text{MSC}, \text{FEM}\}$), $\mathcal{C}^{\ell}$ to be the set of cases (e.g., $\mathcal{C}^{de} = \{\text{NOM}, \text{ACC}, \text{GEN}, \text{DAT}\}$), and $\mathcal{N}^{\ell}$ to be the set of numbers (e.g., $\mathcal{N}^{pt} = \{\text{PL}, \text{SG}\}$). Finally, we define fourteen random variables: A^{ℓ}_i and A^{ℓ}_a are $\mathcal{V}^{\ell}_{\text{ADJ}}$-valued random variables, D^{ℓ}_i and D^{ℓ}_a are $\mathcal{V}^{\ell}_{\text{VERB-DOBJ}}$-valued random variables, I^{ℓ}_i and I^{ℓ}_a are $\mathcal{V}^{\ell}_{\text{VERB-IOBJ}}$-valued random variables, S^{ℓ}_i and S^{ℓ}_a are $\mathcal{V}^{\ell}_{\text{VERB-SUBJ}}$-valued random variables, G^{ℓ}_i and G^{ℓ}_a are $\mathcal{G}^{\ell}$-valued random variables, C^{ℓ}_i and C^{ℓ}_a are $\mathcal{C}^{\ell}$-valued random variables, and N^{ℓ}_i and N^{ℓ}_a are $\mathcal{N}^{\ell}$-valued random variables. The subscripts ''i'' and ''a'' denote inanimate and animate nouns, respectively

To test whether there is a relationship between the grammatical genders of inanimate nouns and the adjectives used to describe those nouns for language ℓ, we calculate the MI (mutual information)—a measure of the mutual statistical dependence between two random variables—between G^{ℓ}_i and A^{ℓ}_i:

$$\text{MI}(G^{\ell}_i; A^{\ell}_i) = \sum_{g \in \mathcal{G}^{\ell}} \sum_{a \in V^{\ell}_{\text{ADJ}}} P(g, a) \log_2 \frac{P_i(g, a)}{P_i(g)\, P_i(a)}, \quad (1)$$

where all probabilities are calculated with respect to inanimate nouns only. If G^{ℓ}_i and A^{ℓ}_i are independent (i.e., there is no relationship between them) then $\text{MI}(G^{\ell}_i; A^{\ell}_i) = 0$; if G^{ℓ}_i and A^{ℓ}_i are maximally dependent then $\text{MI}(G^{\ell}_i; A^{\ell}_i) = \min\{\text{H}(G^{\ell}_i), \text{H}(A^{\ell}_i)\}$, where $\text{H}(G^{\ell}_i)$ is the entropy of G^{ℓ}_i and $\text{H}(A^{\ell}_i)$ is the entropy of A^{ℓ}_i. For simplicity, we use plug-in estimates for all probabilities (i.e., empirical probabilities), deferring the use of more sophisticated estimators for future work. We note that $\text{MI}(G^{\ell}_i, A^{\ell}_i)$ can be calculated in $\mathcal{O}\left(|\,\mathcal{G}^{\ell}| \cdot |V^{\ell}_{\text{ADJ}}|\,\right)$ time; however, $|\,\mathcal{G}^{\ell}|$ is negligible (i.e, two or three) so the main cost is $|V^{\ell}_{\text{ADJ}}|$.

To test for statistical significance, we perform a permutation test. Specifically, we permute the grammatical genders of the inanimate nouns 10,000 times and, for each permutation, recalculate the MI between G^{ℓ}_i and A^{ℓ}_i using the permuted genders. We obtain a p-value by calculating the percentage of permutations that have a higher MI than the MI obtained using the non-permuted genders; if the p-value is less than 0.05, then we treat the relationship between G^{ℓ}_i and A^{ℓ}_i as statistically significant.

Because the maximum possible MI between any pair of random variables depends on the entropies of those variables, MI values are not comparable across pairs of random variables. We therefore also calculate the normalized MI (NMI)

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

	de	*it*	*pl*	*pt*	*ru*	*es*
$\text{MI}(G_i^\ell, A_i^\ell)$	**0.0310**	**0.0500**	**0.0225**	**0.0400**	**0.0520**	**0.0664**
$\text{MI}(G_i^\ell, D_i^\ell)$	**0.0290**	**0.0232**	**0.0109**	**0.0129**	**0.0440**	**0.0090**
$\text{MI}(G_i^\ell, I_i^\ell)$	**0.0743**	0.6973	**0.0514**	**0.0230**	**0.0640**	**0.0184**
$\text{MI}(G_i^\ell, S_i^\ell)$	**0.0276**	**0.0274**	**0.0226**	**0.0090**	**0.0270**	**0.0090**
$\text{MI}(G_i^\ell, C_i^\ell)$	< 0.001	N/A	< 0.001	N/A	< 0.001	N/A
$\text{MI}(G_i^\ell, N_i^\ell)$	< 0.001	< 0.001	<0.001	<0.001	< 0.001	< 0.001

Table 1: The mutual information (MI) between the grammatical genders of inanimate nouns and a) the adjectives used to describe those nouns (top row), b) the verbs that take those nouns as direct objects, as indirect objects, and as subjects (rows 2–4, respectively), and c) the cases and numbers of those nouns (rows 5 and 6, respectively) for six different gendered languages. Statistical significance (i.e., a p-value less than 0.05) is indicated using bold. MI values are not comparable across pairs of random variables.

between G_i^ℓ and A_i^ℓ by normalizing $\text{MI}(G_i^\ell, A_i^\ell)$ to lie between zero and one. The most obvious choice of normalizer is the maximum possible MI—that is, $\min\{\text{H}(G_i^\ell), \text{H}(A_i^\ell)\}$—however, various other normalizers have been proposed, each of which has different advantages and disadvantages (Gates et al., 2019). We therefore calculate six different variants of $\text{NMI}(G_i^\ell, A_i^\ell)$ using the following normalizers:

$$\min\{\text{H}(G_i^\ell), \text{H}(A_i^\ell)\} \tag{2}$$

$$\sqrt{\text{H}(G_i^\ell)\text{H}(A_i^\ell)} \tag{3}$$

$$\frac{\text{H}(G_i^\ell) + \text{H}(A_i^\ell)}{2} \tag{4}$$

$$\max\{\text{H}(G_i^\ell), \text{H}(A_i^\ell)\} \tag{5}$$

$$\max\{\log|\mathcal{G}^\ell|, \log|\mathcal{V}_{\text{ADJ}}^\ell|\} \tag{6}$$

$$\log M_i^\ell, \tag{7}$$

where M_i^ℓ is the number of non-unique (inanimate) noun–adjective pairs retained for that language.

To test whether there are relationships between the grammatical genders of inanimate nouns and the verbs that take those nouns as direct objects, as indirect objects, and as subjects, we calculate $\text{MI}(G_i^\ell, D_i^\ell)$, $\text{MI}(G_i^\ell, I_i^\ell)$, and $\text{MI}(G^\ell, S_i^\ell)$. Again, all probabilities are calculated with respect to inanimate nouns only, and we perform permutation tests to test for statistical significance. We also calculate six NMI variants for each of the three pairs of random variables, using normalizers that are analogous to those in Eq. (2) through Eq. (7).

As a baseline, we test whether there are relationships between the grammatical genders of inanimate nouns and the cases and numbers of those nouns—that is, we calculate $\text{MI}(G_i^\ell, C_i^\ell)$ and $\text{MI}(G_i^\ell, N_i^\ell)$ using probabilities that are calculated with respect to inanimate nouns only. Again, we perform permutation tests (but we do not expect that there will be statistically significant relationships), and we calculate six NMI variants for each pair of random variables using normalizers that are analogous to those in Eq. (2) through Eq. (7).

Finally, we calculate $\text{MI}(G_a^\ell, A_a^\ell)$, $\text{MI}(G_a^\ell, D_a^\ell)$, $\text{MI}(G_a^\ell, I_a^\ell)$, $\text{MI}(G_a^\ell, S_a^\ell)$, $\text{MI}(G_a^\ell, C_a^\ell)$, and $\text{MI}(G_a^\ell, N_a^\ell)$) using probabilities calculated with respect to *animate* nouns only. The first five of these are intended to serve as a ''skyline,'' while the last two are intended to serve as a sanity check (i.e., we expect them to be close to zero, as with inanimate nouns). Again, we perform permutation tests to test for statistical significance, and we calculate six NMI variants for each pair of random variables.

5 Results

In the first row of Table 1, we provide the MI between G_i^ℓ and A_i^ℓ for each language $\ell \in \{de, it, pl, pt, ru, es\}$. For all six languages, $\text{MI}(G_i^\ell, A_i^\ell)$ is statistically significant (i.e., $p < 0.05$), meaning that there is a relationship between the grammatical genders of inanimate nouns and

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

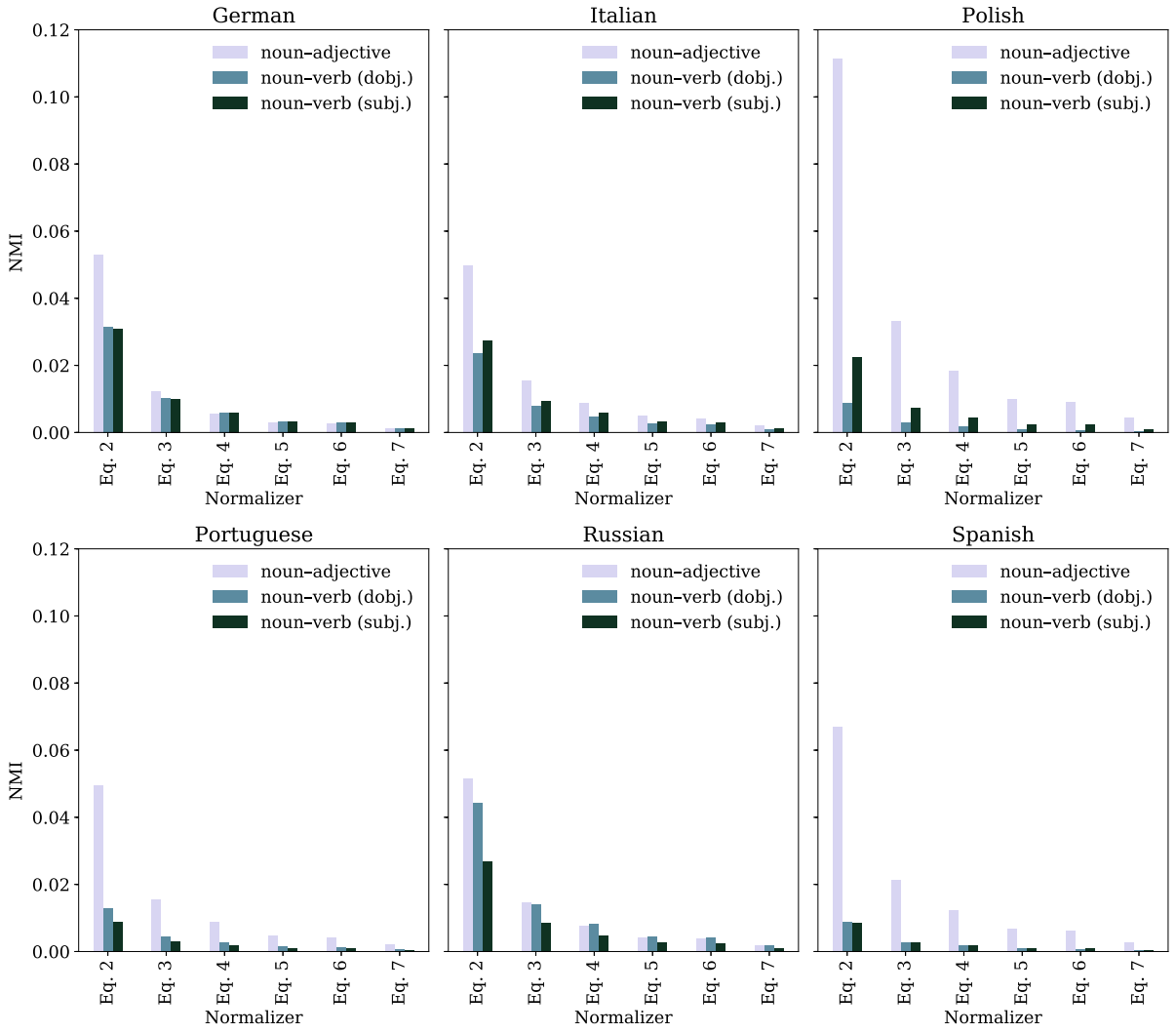

Figure 2: The normalized mutual information (NMI) between the grammatical genders of inanimate nouns and a) the adjectives used to describe those nouns and b) the verbs that take those nouns as direct objects and as subjects for six different gendered languages. Each subplot contains six variants of NMI(G_i^ℓ, A_i^ℓ), NMI(G_i^ℓ, D_i^ℓ), and NMI(G_i^ℓ, S_i^ℓ)—one per normalizer—for a single language $\ell \in \{de, it, pl, pt, ru, es\}$.

the adjectives used to describe those nouns. Rows 2–4 of Table 1 contain MI(G_i^ℓ, D_i^ℓ), MI(G_i^ℓ, I_i^ℓ), and MI(G^ℓ, S_i^ℓ) for each language. For all six languages, MI(G_i^ℓ, D_i^ℓ) and MI(G_i^ℓ, S_i^ℓ) are statistically significant (i.e., $p < 0.05$). For five of the six languages, MI(G_i^ℓ, I_i^ℓ) is statistically significant, but because of the small number of noun–verb pairs involved, we caution against reading too much into this finding. We note that direct objects are closest to verbs in analyses of constituent structures, followed by subjects and then indirect objects (Chomsky, 1957; Adger, 2003). Finally, the last two rows of Table 1 contain MI(G_i^ℓ, C_i^ℓ) and MI(G_i^ℓ, N_i^ℓ), respectively, for each language. We do not find any statistically significant relationships for either case or number.

To facilitate comparisons, each subplot in Figure 2 contains six variants of NMI(G_i^ℓ, A_i^ℓ), NMI(G_i^ℓ, D_i^ℓ), and NMI(G_i^ℓ, S_i^ℓ), calculated using normalizers that are analogous to those in Eq. (2) through Eq. (7), for a single language $\ell \in \{de, it, pl, pt, ru, es\}$. (We omit NMI$(G_i^\ell, I_i^\ell)$ from each plot because of the small number of noun–verb pairs involved.) For $\ell \in \{it, pl, pt, es\}$, NMI$(G_i^\ell, A_i^\ell)$ is larger than NMI(G_i^ℓ, D_i^ℓ) and NMI(G_i^ℓ, S_i^ℓ), regardless of the normalizer. For $\ell \in \{it, pl\}$, NMI(G_i^ℓ, S_i^ℓ) is larger than NMI(G_i^ℓ, D_i^ℓ); NMI(G_i^{pt}, D_i^{pt}) is larger than NMI(G_i^{pt}, S_i^{pt}); and NMI(G_i^{es}, D_i^{es}) and NMI(G_i^{es}, S_i^{es}) are roughly comparable—again, all regardless of the normalizer. Meanwhile, NMI(G_i^{de}, A_i^{de}) is larger than NMI(G_i^{de}, D_i^{de})

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

	de	*it*	*pl*	*pt*	*ru*	*es*
$\text{MI}(G_a^\ell, A_a^\ell)$	**0.0928**	**0.1316**	**0.0621**	**0.0933**	**0.0845**	**0.1111**
$\text{MI}(G_a^\ell, D_a^\ell)$	**0.0410**	**0.0543**	**0.0273**	**0.0320**	**0.0664**	**0.0091**
$\text{MI}(G_a^\ell, I_a^\ell)$	**0.0737**	**0.0543**	**0.0439**	**0.0687**	**0.0600**	**0.0358**
$\text{MI}(G_a^\ell, S_a^\ell)$	**0.0343**	**0.0543**	**0.0258**	**0.0252**	**0.0303**	**0.0192**
$\text{MI}(G_a^\ell, C_a^\ell)$	< 0.001	N/A	< 0.001	N/A	< 0.001	N/A
$\text{MI}(G_a^\ell, N_a^\ell)$	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table 2: The mutual information (MI) between the grammatical genders of animate nouns and a) the adjectives used to describe those nouns (top row), b) the verbs that take those nouns as direct objects, as indirect objects, and as subjects (rows 2–4, respectively), and c) the cases and numbers of those nouns (rows 5 and 6, respectively) for six different gendered languages. Statistical significance (i.e., a p-value less than 0.05) is indicated using bold. MI values are not comparable across pairs of random variables.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

and $\text{NMI}(G_i^{de}, S_i^{de})$ for the normalizer in Eq. (2), while $\text{NMI}(G_i^{de}, A_i^{de})$, $\text{NMI}(G_i^{de}, D_i^{de})$, and $\text{NMI}(G_i^{de}, S_i^{de})$ are all roughly comparable for the other five normalizers. Finally, $\text{NMI}(G_i^{ru}, A_i^{ru})$ and $\text{NMI}(G_i^{ru}, D_i^{ru})$ are roughly comparable and larger than $\text{NMI}(G_i^{ru}, S_i^{ru})$, regardless of the normalizer.

In other words, the relationship between the grammatical genders of inanimate nouns and the adjectives used to describe those nouns is generally stronger than, but sometimes roughly comparable to, the relationships between the grammatical genders of inanimate nouns and the verbs that take those nouns as direct objects and as subjects. However, the relative strengths of the relationships between the grammatical genders of inanimate nouns and the verbs that take those nouns as direct objects and as subjects vary depending on the language.

In Table 2, we provide $\text{MI}(G_a^\ell, A_a^\ell)$, $\text{MI}(G_a^\ell, D_a^\ell)$, $\text{MI}(G_a^\ell, I_a^\ell)$, $\text{MI}(G_a^\ell, S_a^\ell)$, $\text{MI}(G_a^\ell, C_a^\ell)$, and $\text{MI}(G_a^\ell, N_a^\ell)$ for each language $\ell \in \{de, it, pl, pt, ru, es\}$. As with inanimate nouns, we find that there is a statistically significant relationship between the grammatical genders of animate nouns and the adjectives used to describe those nouns. We also find that there are statistically significant relationships between the grammatical genders of animate nouns and the verbs that take those nouns as direct objects, as indirect objects, and as subjects. Again, the relationship for the verbs that take those nouns as indirect objects involves a small number of noun–verb pairs. As expected, we do not find any statistically significant relationships for either case or number.

Figure 3 is analogous to Figure 2, in that each subplot contains six variants of $\text{NMI}(G_a^\ell, A_a^\ell)$, $\text{NMI}(G_a^\ell, D_a^\ell)$, and $\text{NMI}(G_a^\ell, S_a^\ell)$, calculated using normalizers that are analogous to those in Eq. (2) through Eq. (7), for a single language $\ell \in \{de, it, pl, pt, ru, es\}$. (As with inanimate nouns, we omit $\text{NMI}(G_a^\ell, I_a^\ell)$ from each plot because of the small number of noun–verb pairs involved.) For $\ell \in \{de, it, pl, pt, es\}$, $\text{NMI}(G_i^\ell, A_i^\ell)$ is larger than $\text{NMI}(G_i^\ell, D_i^\ell)$ and $\text{NMI}(G_i^\ell, S_i^\ell)$, regardless of the normalizer. For $\ell \in \{it, pl\}$, $\text{NMI}(G_i^\ell, S_i^\ell)$ is larger than $\text{NMI}(G_i^\ell, D_i^\ell)$; for $\ell \in \{de, pt\}$, $\text{NMI}(G_i^\ell, D_i^\ell)$ is larger than $\text{NMI}(G_i^\ell, S_i^\ell)$; and $\text{NMI}(G_i^{es}, D_i^{es})$ and $\text{NMI}(G_i^{es}, S_i^{es})$ are roughly comparable—again, all regardless of the normalizer. Meanwhile, $\text{NMI}(G_i^{ru}, A_i^{ru})$ is larger than $\text{NMI}(G_i^{ru}, D_i^{ru})$ which is larger than $\text{NMI}(G_i^{ru}, S_i^{ru})$ for the normalizers in Eq. (2) and Eq. (3), while $\text{NMI}(G_i^{ru}, A_i^{ru})$ and $\text{NMI}(G_i^{ru}, D_i^{ru})$ are roughly comparable and larger than $\text{NMI}(G_i^{ru}, S_i^{ru})$ for the other five normalizers.

Finally, each subplot in Figure 4 contains $\text{NMI}(G_i^\ell, A_i^\ell)$ and $\text{NMI}(G_a^\ell, A_a^\ell)$, calculated using a single normalizer, for each for each language $\ell \in \{de, it, pl, pt, ru, es\}$. Each subplot in Figure 5 analogously contains $\text{NMI}(G_i^\ell, D_i^\ell)$ and $\text{NMI}(G_a^\ell, D_a^\ell)$, while each subplot in Figure 6 contains $\text{NMI}(G_i^\ell, S_i^\ell)$ and $\text{NMI}(G_a^\ell, S_a^\ell)$. The NMI values for animate nouns are generally larger

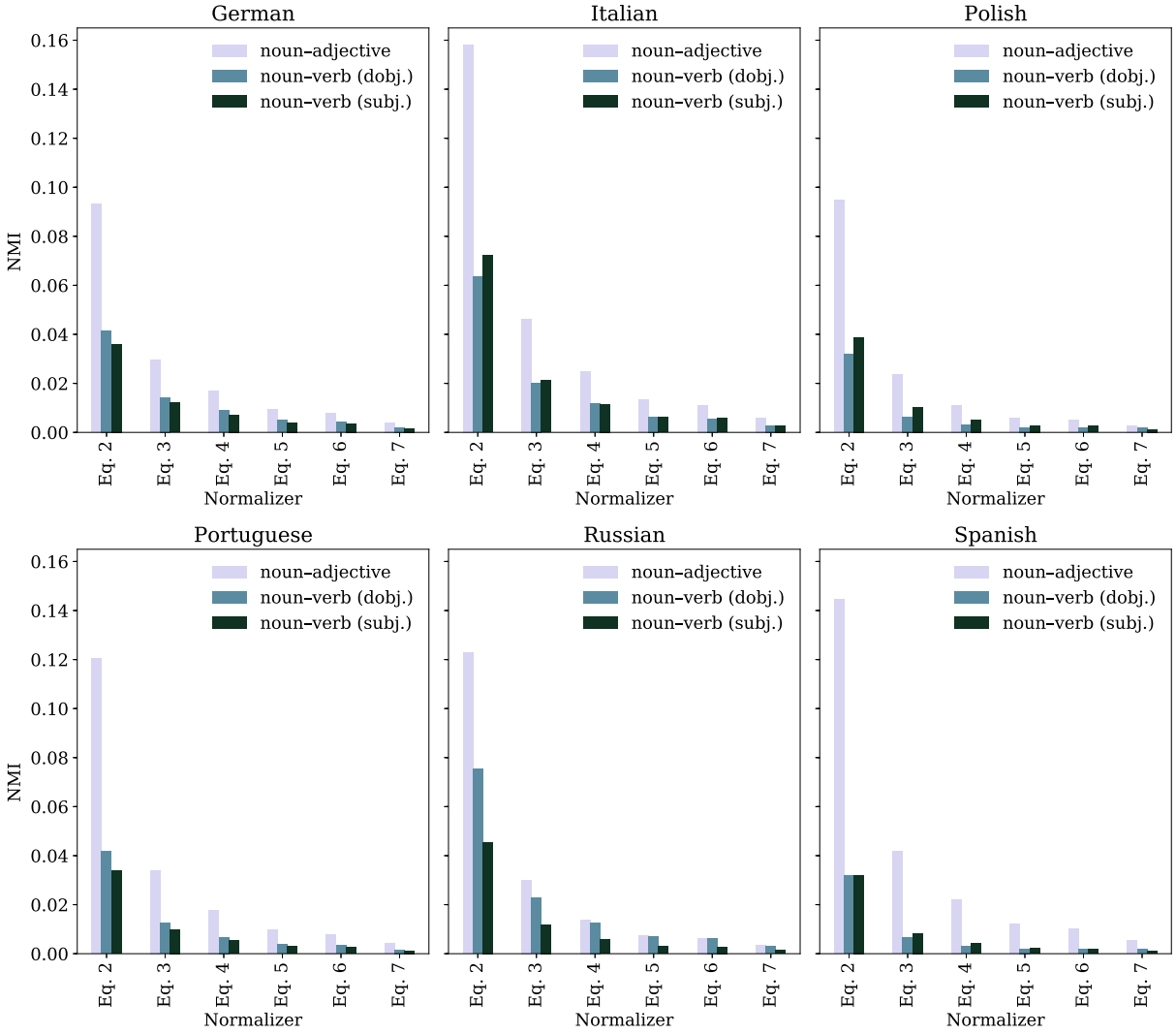

Figure 3: The normalized mutual information (NMI) between the grammatical genders of animate nouns and a) the adjectives used to describe those nouns and b) the verbs that take those nouns as direct objects and as subjects for six different gendered languages. Each subplot contains six variants of NMI(G_a^ℓ, A_a^ℓ), NMI(G_a^ℓ, D_a^ℓ), and NMI(G_a^ℓ, S_a^ℓ)—one per normalizer—for a single language $\ell \in \{de, it, pl, pt, ru, es\}$.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

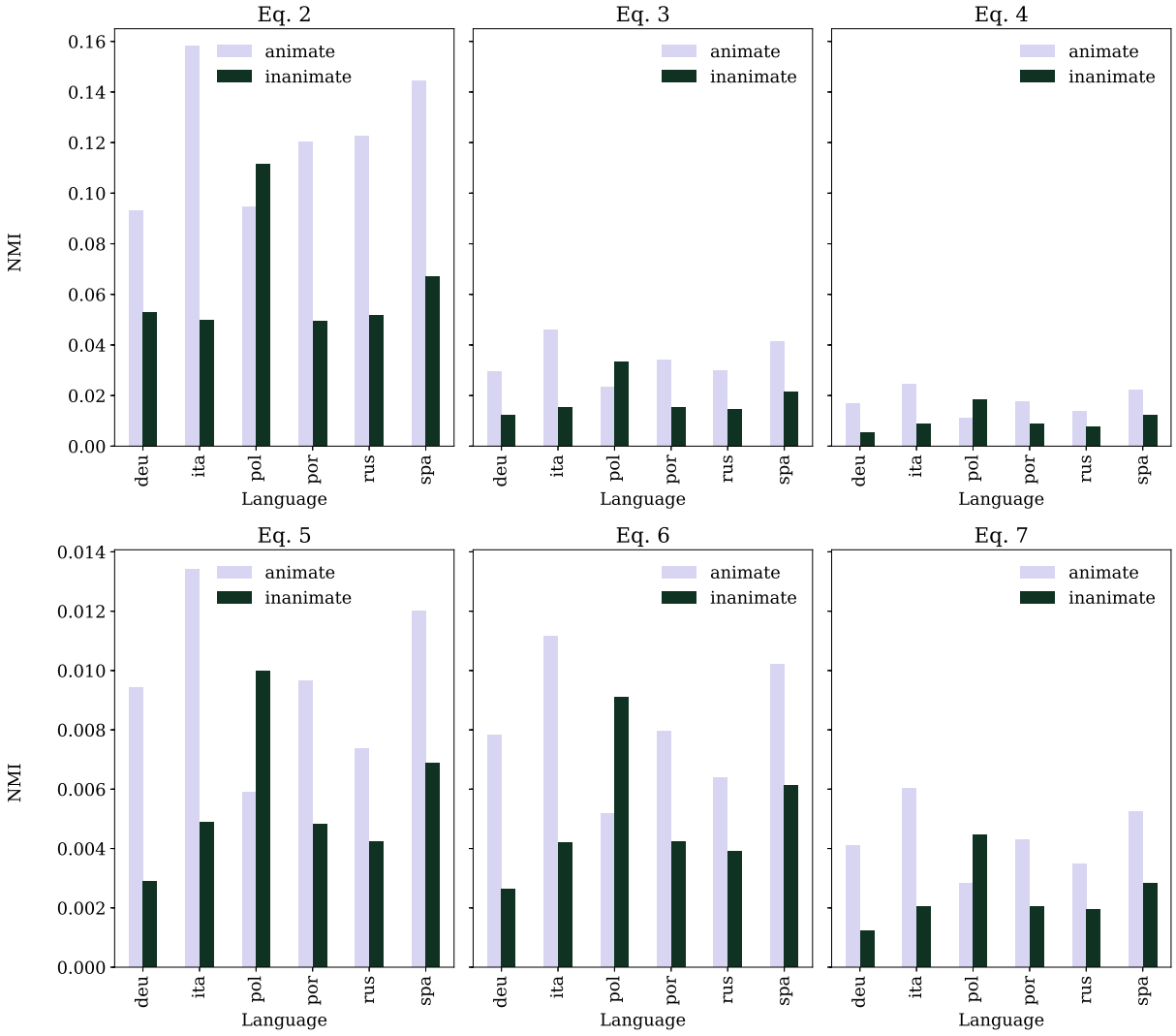

Figure 4: The normalized mutual information (NMI) between the grammatical genders of a) inanimate and b) animate nouns and the adjectives used to describe those nouns. Each subplot contains NMI(G_i^ℓ, A_i^ℓ) and NMI(G_a^ℓ, A_a^ℓ), calculated using a single normalizer, for each language $\ell \in \{de, it, pl, pt, ru, es\}$.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

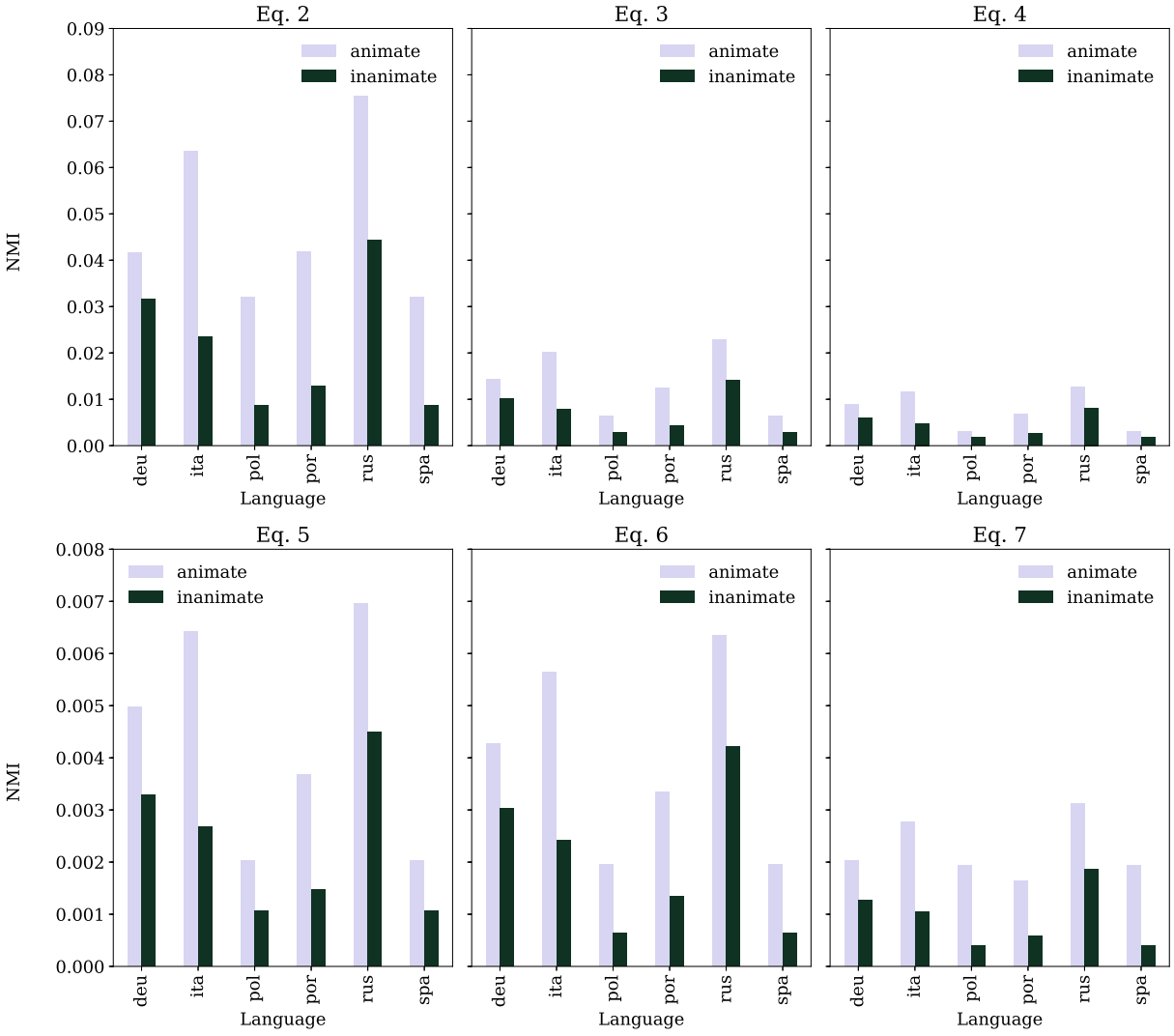

Figure 5: The normalized mutual information (NMI) between the grammatical genders of a) inanimate and b) animate nouns and the verbs that take those nouns as direct objects. Each subplot contains NMI(G_i^ℓ, D_i^ℓ) and NMI(G_a^ℓ, D_a^ℓ), calculated using a single normalizer, for each language $\ell \in \{de, it, pl, pt, ru, es\}$.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

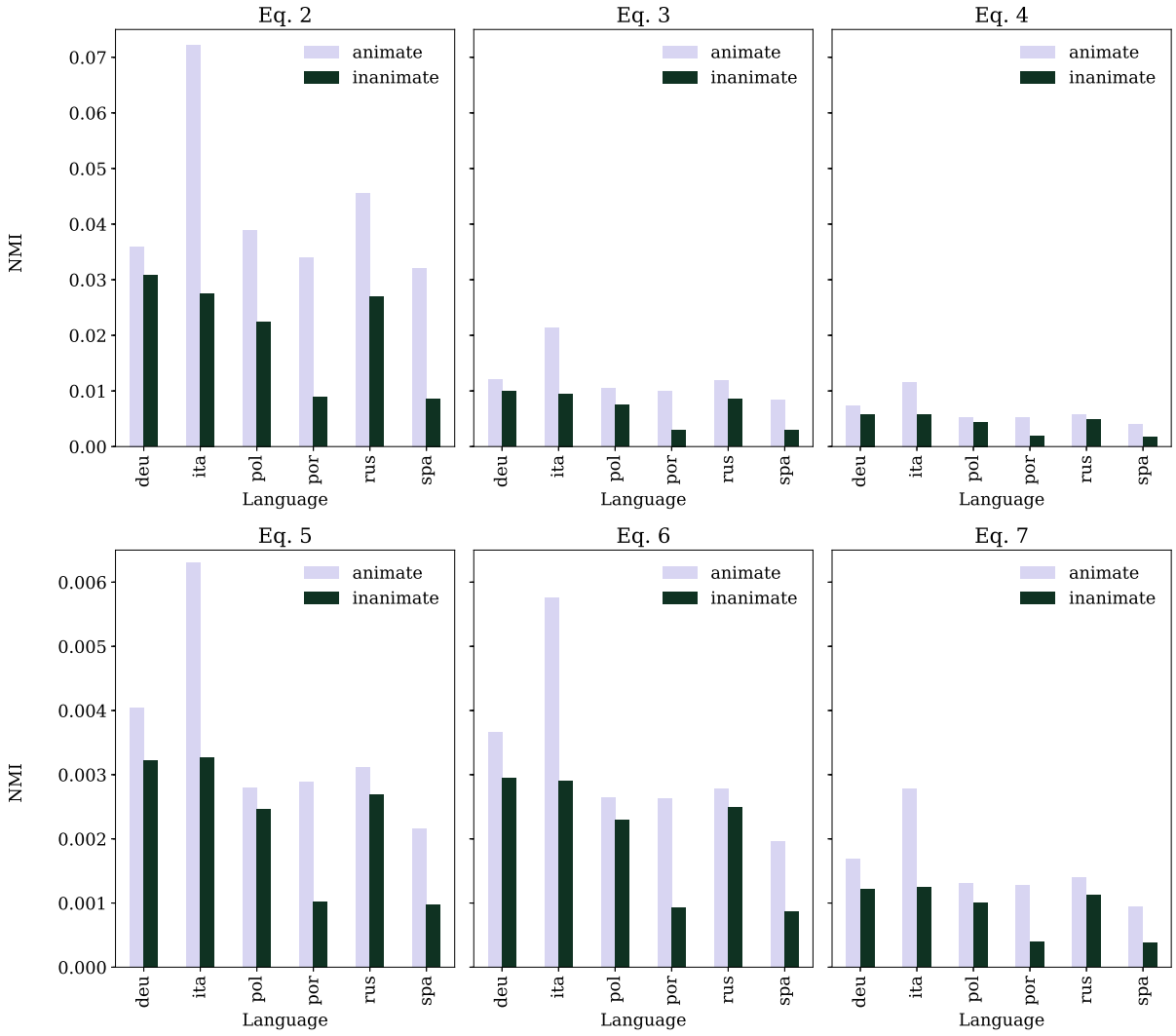

Figure 6: The normalized mutual information (NMI) between the grammatical genders of a) inanimate and b) animate nouns and the verbs that take those nouns as subjects. Each subplot contains NMI(G_i^ℓ, S_i^ℓ) and NMI(G_a^ℓ, S_a^ℓ), calculated using a single normalizer, for each language $\ell \in \{de, it, pl, pt, ru, es\}$.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

than the NMI values for inanimate nouns. The only exception is Polish, where NMI(G_i^{pl}, A_i^{pl}) is larger than NMI(G_a^{pl}, A_a^{pl}), regardless of the normalizer.

6 Discussion

We find evidence for the presence of a statistically significant relationship between the grammatical genders of inanimate nouns and the adjectives used to describe those nouns for six different gendered languages (specifically, German, Italian, Polish, Portuguese, Russian, and Spanish). We also find evidence for the presence of statistically significant relationships between the grammatical genders of inanimate nouns and the verbs that take those nouns as direct objects, as indirect objects, and as subjects. However, we caution against reading too much into the relationship for the verbs that take those nouns as indirect objects because of the small number of noun–verb pairs involved. The effect sizes (operationalized as NMI values) for all of these relationships are smaller than the effect sizes for animate nouns. As expected, we do not find any statistically significant relationships for either case or number.

We emphasize that our findings complement, rather than supersede, laboratory experiments, such as that of Boroditsky et al. (2003). We use large-scale corpora and tools from NLP and information theory to test for the presence of even relatively weak relationships across multiple different gendered languages—and, indeed, the relationships that we find have effect sizes (operationalized as NMI values) that are small. In contrast, laboratory experiments typically focus on much stronger relationships by tightly controlling experimental conditions and measuring speakers' immediate, real-time speech production. Moreover, although we find statistically significant relationships, we do not investigate the characteristics of these relationships. This means that we do not know whether they are characterized by gender stereotypes, as argued by some cognitive scientists, including Boroditsky et al. (2003). We also do not know whether the relationships that we find are causal in nature. Because MI is symmetric, our findings say nothing about whether the grammatical genders of inanimate nouns *cause* writers to choose particular adjectives or verbs. We defer deeper investigation of this for future work.

We note that each of our tests can be viewed as a comparison of the similarity of two clusterings of a set of items—specifically, a ''clustering'' of nouns into grammatical genders and a ''clustering'' of the same nouns into, for example, adjective lemmata. Although (normalized) MI is a standard measure for comparing clusterings, it is not without limitations (see, e.g., Newman et al. [2020] for an overview). For future work, we therefore recommend replicating our tests using other information-theoretic measures for comparing clusterings.

Acknowledgments

We thank Lera Boroditsky, Hagen Blix, Eleanor Chodroff, Andrei Cimpian, Zach Davis, Jason Eisner, Richard Futrell, Todd Gureckis, Katharina Kann, Peter Klecha, Zhiwei Li, Ethan Ludwin-Peery, Alec Marantz, Arya McCarthy, John McWhorter, Sabrina J. Mielke, Elizabeth Salesky, Arturs Semenuks, and Colin Wilson for discussions at various points related to the ideas in this paper. Katharina Kann approves this acknowledgment.

A Appendix A: Counts

Counts of the noun–adjective and noun–verb pairs for all six gendered languages are in Table 3 (for inanimate nouns) and Table 4 (for animate nouns).

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

	de	*it*	*pl*	*pt*	*ru*	*es*
# noun–adj. tokens	6443907	6246856	11631913	640558	32900200	3605439
# noun–adj. types	770952	666656	640107	638774	1633963	368795
# noun types	10712	6410	5533	5672	9327	6157
# adj. types	4129	3607	4080	3431	11028	1907
# noun–verb (subj.) tokens	3191030	1432354	2179396	1871941	6007063	1534211
# noun–verb (subj.) types	445536	292949	297996	337262	864480	376888
# noun (subj.) types	10741	6318	5522	5780	9129	7470
# verb types	707	702	874	758	1803	875
# noun–verb (dobj.) tokens	3440922	2855037	3964828	4850012	6738606	2859135
# noun–verb (dobj.) types	427441	393246	236849	541347	713703	576835
# noun (dobj.) types	10504	6407	4359	5896	8998	11567
# verb types	805	806	708	738	1539	9746
# noun–verb (iobj.) tokens	163935	71	54138	95009	1570273	56038
# noun–verb (iobj.) types	50133	53	18214	39738	300703	24830
# noun (iobj.) types	5520	59	2258	3757	8150	3574
# verb types	386	68	417	357	1816	464
# noun–case tokens	14681293	N/A	15300621	N/A	51641929	N/A
# noun–case types	2252632	N/A	1465314	N/A	5028075	N/A
# noun types	11989	N/A	5839	N/A	9692	N/A
# case types	4	0	7	0	6	0
# noun–number tokens	14681293	11588448	15300621	14631732	51641929	5672790
# noun–number types	2252632	1748927	1465314	2042626	5028075	1034307
# noun types	11989	7014	5839	6256	9692	1593
# number types	2	2	2	2	2	2

Table 3: Counts of the inanimate noun–adjective and noun–verb pairs for all six gendered languages.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

	de	*it*	*pl*	*pt*	*ru*	*es*
# noun–adj. tokens	662760	818300	1137209	712101	3225932	387025
# noun–adj. types	99332	92424	97847	90865	264117	50173
# noun types	1998	1078	954	1006	2098	1320
# adj. types	3587	3507	3836	3176	9833	1828
# noun–verb (subj.) tokens	637801	399747	526894	456349	1516740	310569
# noun–verb (subj.) types	113308	77551	89819	89959	253150	93586
# noun (subj.) types	2056	1066	969	1013	2020	1477
# verb types	707	702	874	758	1799	874
# noun–verb (dobj.) tokens	321400	388187	456824	527259	494534	850234
# noun–verb (dobj.) types	60760	55574	76348	92220	118818	85235
# noun (dobj.) types	1901	1025	867	1028	1912	1023
# verb types	804	805	724	737	1535	745
# noun–verb (iobj.) tokens	51359	7	43187	23139	518540	23955
# noun–verb (iobj.) types	17804	6	8440	110185	11353	9586
# noun (iobj.) types	1149	6	628	773	1858	947
# verb types	378	6	411	340	1769	456
# noun–case tokens	1926614	N/A	1907688	N/A	6357089	N/A
# noun–case types	390672	N/A	299511	N/A	987420	N/A
# noun types	2292	N/A	1024	N/A	2194	N/A
# case types	4	0	7	0	6	0
# noun–number tokens	1926614	1801285	1907688	1931315	6357089	786177
# noun–number types	390672	306968	299511	356352	987420	200785
# noun types	2292	1135	1024	1072	2194	1593
# number types	2	2	2	2	2	2

Table 4: Counts of the animate noun–adjective and noun–verb pairs for all six gendered languages.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

References

David Adger. 2003. *Core Syntax: A Minimalist Approach*, 33. Oxford University Press Oxford.

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

Tatiana Akhutina, Andrei Kurgansky, Maria Polinsky, and Elizabeth Bates. 1999. Processing of grammatical gender in a three-gender system: Experimental evidence from Russian. *Journal of Psycholinguistic Research*, 28(6): 695–713. **DOI:** https://doi.org/10.1023/A:1023225129058, **PMID:** 10510865

Chris Alberti, Daniel Andor, Ivan Bogatyy, Michael Collins, Dan Gillick, Lingpeng Kong, Terry Koo, Ji Ma, Mark Omernick, Slav Petrov, Chayut Thanapirom, Zora Tung, and David Weiss. 2017. SyntaxNet models for the CoNLL 2017 shared task. CoRR abs/1703.04929 arXiv preprint arXiv:1703.04929.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P16-1231

José Alemán Bañón, Robert Fiorentino, and Alison Gabriele. 2012. The processing of number and gender agreement in Spanish: An event-related potential investigation of the effects of structural distance. *Brain Research*, 1456:49–63. **DOI:** https://doi.org/10.1016/j.brainres.2012.03.057, **PMID:** 22520436

Horacio Barber and Manuel Carreiras. 2005. Grammatical gender and number agreement in Spanish: An ERP comparison. *Journal of Cognitive Neuroscience*, 17(1):137–153. **DOI:** https://doi.org/10.1162/0898929052880101, **PMID:** 15701245

Horacio Barber, Elena Salillas, and Manuel Carreiras. 2004. Gender or genders agreement. *On-line Study of Sentence Comprehension*, pages 309–328.

Benedetta Bassetti. 2007. Bilingualism and thought: Grammatical gender and concepts of objects in Italian–German bilingual children. *International Journal of Bilingualism*, 11(3): 251–273. **DOI:** https://doi.org/10.1177/13670069070110030101

Elizabeth Bates, Antonella Devescovi, Arturo Hernandez, and Luigi Pizzamiglio. 1996. Gender priming in Italian. *Perception & Psychophysics*, 58(7):992–1004. **DOI:** https://doi.org/10.3758/BF03206827, **PMID:** 8920836

Andrea Bender, Sieghard Beller, and Karl Christoph Klauer. 2011. Grammatical gender in german: A case for linguistic relativity? *The Quarterly Journal of Experimental Psychology*, 64(9):1821–1835. **DOI:** https://doi.org/10.1080/17470218.2011.582128, **PMID:** 21740112

Damian Blasi, Ryan Cotterell, Lawrence Wolf-Sonkin, Sabine Stoll, Balthasar Bickel, and Marco Baroni. 2019. On the distribution of deep clausal embeddings: A large cross-linguistic study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3938–3943. **DOI:** https://doi.org/10.18653/v1/P19-1384

Leonard Bloomfield. 1933. *Language*, London: Allen & Unwin.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Lera Boroditsky. 2003. Linguistic Relativity. *Encyclopedia of Cognitive Science*.

Lera Boroditsky and Lauren A. Schmidt. 2000. Sex, Syntax, and Semantics. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

Lera Boroditsky, Lauren A. Schmidt, and Webb Phillips. 2002. Can quirks of grammar affect the way you think? Spanish and German speakers' ideas about the genders of objects. https://escholarship.org/uc/item/31t455gf.

Lera Boroditsky, Lauren A. Schmidt, and Webb Phillips. 2003. Sex, Syntax, and Semantics. *Language in Mind: Advances in the Study of Language and Thought*, pages 61–79.

Bastien Boutonnet, Panos Athanasopoulos, and Guillaume Thierry. 2012. Unconscious effects of grammatical gender during object categorisation. *Brain Research*, 1479:72–79. **DOI:** https://doi.org/10.1016/j.brainres.2012.08.044, **PMID:** 22960201

Karl Brugmann. 1889. Das nominalgeschlecht in den indogermanischen sprachen. *Internationale Zeitschrift für allgemeine Sprachwissenschaft*, 4. **DOI:** https://doi.org/10.1111/psyp.12429, **PMID:** 25817315

Sendy Caffarra, Anna Siyanova-Chanturia, Francesca Pesciarelli, Francesco Vespignani, and Cristina Cacciari. 2015. Is the noun ending a cue to grammatical gender processing? An ERP study on sentences in Italian. *Psychophysiology*, 52(8):1019–1030. **DOI:** https://doi.org/10.1126/science.aal4230, **PMID:** 28408601

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Noam Chomsky. 1957. *Syntactic Structures (The Hague/Paris, Mouton)*, The Hague/Paris: Mouton.

Mark A. Clarke, Ann Losoff, Margaret Dickenson McCracken, and JoAnn Still. 1981. Gender perception in Arabic and English. *Language Learning*, 31(1):159–169. **DOI:** https://doi.org/10.1111/j.1467-1770.1981.tb01377.x

Greville G. Corbett. 1991. *Gender*, Cambridge University Press. **DOI:** https://doi.org/10.1017/CBO9781139166119

Greville G. Corbett. 2006. *Agreement*, Cambridge University Press.

Roberto Cubelli, Lorella Lotto, Daniela Paolieri, Massimo Girelli, and Remo Job. 2005. Grammatical gender is selected in bare noun production: Evidence from the picture–word interference paradigm. *Journal of Memory and Language*, 53(1):42–59. **DOI:** https://doi.org/10.1016/j.jml.2005.02.007

Roberto Cubelli, Daniela Paolieri, Lorella Lotto, and Remo Job. 2011. The effect of grammatical gender on object categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(2):449. **DOI:** https://doi.org/10.1037/a0021965, **PMID:** 21261427

Robert M. DeKeyser. 2005. What makes learning second-language grammar difficult? A review of issues. *Language Learning*, 55(S1):1–25. **DOI:** https://doi.org/10.1111/j.0023-8333.2005.00294.x

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*. **DOI:** https://doi.org/10.18653/v1/2020.emnlp-main.656

Melody Dye, Petar Milin, Richard Futrell, and Michael Ramscar. 2017. A functional theory of gender paradigms, In *Perspectives on Morphological Organization*, pages 212–239. Brill. **DOI:** https://doi.org/10.1163/9789004342934_011

Susan Ervin-Tripp. 1962. The connotations of gender. *Word*, 18249–261.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

pages 1696–1705, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1166

Sam Featherston and Wolfgang Sternefeld. 2007. *Roots: Linguistics in Search of its Evidential Base*, volume 96. Walter de Gruyter. **DOI:** https://doi.org/10.1515/9783110198621

Istvan Fodor. 1959. The origin of grammatical gender. *Lingua*, 8:186–214. **DOI:** https://doi.org/10.1016/0024-3841(59)90020-8

Anthony Fox. 1990. *The structure of German*, Oxford University Press.

Florencia Franceschina. 2005. *Fossilized Second Language Grammars: The Acquisition of Grammatical Gender*, volume 38. John Benjamins Publishing. **DOI:** https://doi.org/10.1075/lald.38

Angela D. Friederici and Thomas Jacobsen. 1999. Processing grammatical gender during language comprehension. *Journal of Psycholinguistic Research*, 28(5):467–484. **DOI:** https://doi.org/10.1023/A:1023243708702, https://doi.org/10.1023/A:1023264209610

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644. **DOI:** https://doi.org/10.1073/pnas.1720347115, **PMID:** 29615513, **PMCID:** PMC5910851

Alexander J. Gates, Ian B. Wood, William P. Hetrick, and Yong-Yeol Ahn. 2019. Element-centric clustering comparison unifies overlaps and hierarchy. *Scientific Reports*, 9(8574). **DOI:** https://doi.org/10.1038/s41598-019-44892-y, **PMID:** 31189888, **PMCID:** PMC6561975

Jacob Grimm. 1890. *Deutsche Grammatik*, C. Bertelsmann.

Peter Hagoort and Colin M. Brown. 1999. Gender electrified: ERP evidence on the syntactic nature of gender processing. *Journal of Psycholinguistic Research*, 28(6):715–728. **DOI:** https://doi.org/10.1023/A:1023277213129, **PMID:** 10510866

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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 5267–5275, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1530

Daniel Harbour. 2011. Valence and atomic number. *Linguistic Inquiry*, 42(4):561–594. **DOI:** https://doi.org/10.1162/LING_a_00061

Daniel Harbour, David Adger, and Susana Béjar. 2008. *Phi theory: Phi-features Across Modules and Interfaces*, 16, Oxford University Press.

Peter R. Hofstätter. 1963. Über sprachliche bestimmungsleistungen: Das problem des grammatikalischen geschlechts von sonne und mond. *Zeitschrift für experimentelle und angewandte Psychologie*.

Muhammad Hasan Ibrahim. 2014. *Grammatical Gender: Its Origin and Development*, 166, Walter de Gruyter.

Roman Jakobson. 1959. On linguistic aspects of translation. *On Translation*, 3:30–39. **DOI:** https://doi.org/10.4159/harvard.9780674731615.c18

Katharina Kann. 2019. Grammatical gender, neo-Whorfianism, and word embeddings: A Data-Driven Approach to Linguistic Relativity. *arXiv preprint arXiv:1910.09729*.

Graeme Kennedy. 2014. *An Introduction to Corpus Linguistics*, Routledge. **DOI:** https://doi.org/10.4324/9781315843674

Toshi Konishi. 1993. The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research*, 22(5):519–534. **DOI:** https://doi.org/10.1007/BF01068252, **PMID:** 8246207

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

Ruth Kramer. 2014. Gender in Amharic: a morphosyntactic approach to natural and grammatical gender. *Language Sciences*, 43: 102–115. **DOI:** https://doi.org/10.1016/j.langsci.2013.10.004

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

Elena Kurinski and Maria D. Sera. 2011. Does learning Spanish grammatical gender change English-speaking adults' categorization of inanimate objects? *Bilingualism: Language and Cognition*, 14(2):203–220. **DOI:** https://doi.org/10.1017/S1366728910000179

Michael Maratsos. 1979. How to get from words to sentences, Doris Aaronson and Rober W. Reiber, editors, *Psycholinguistic Research: Implications and Applications*, Psychology Press, Taylor & Francis Group, London and New York.

John H. McWhorter. 2014. *The Language Hoax: Why the world looks the same in any language*, Oxford University Press.

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

Silvina Montrul, Rebecca Foote, and Silvia Perpiñán. 2008. Gender agreement in adult second language learners and spanish heritage speakers: The effects of age and context of acquisition. *Language Learning*, 58(3): 503–553. **DOI:** https://doi.org/10.1111/j.1467-9922.2008.00449.x

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274. **DOI:** https://doi.org/10.18653/v1/D15-1272, **PMID:** 25768671

Mark E. J. Newman, George T. Cantwell, and Jean-Gabriel Young. 2020. Improved mutual information measure for classification and community detection. *Physical Review E*. **DOI:** https://doi.org/10.1103/PhysRevE.101.042304, **PMID:** 32422767

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, and Kaja Dobrovoljc. 2017. Universal dependencies 2.0.

Lee Osterhout and Linda A. Mobley. 1995. Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, 34(6):739–773. **DOI:** https://doi.org/10.1006/jmla.1995.1033

Webb Phillips and Lera Boroditsky. 2003. Can quirks of grammar affect the way you think? Grammatical gender and object concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25.

Princeton University. 2010. About WordNet. https://wordnet.princeton.edu/

Henrik Saalbach, Mutsumi Imai, and Lennart Schalk. 2012. Grammatical gender and inferences about biological properties in German-speaking children. *Cognitive Science*, 36(7): 1251–1267. **DOI:** https://doi.org/10.1111/j.1551-6709.2012.01251.x, **PMID:** 22578067

Arturs Semenuks, Webb Phillips, Ioana Dalca, Cora Kim, and Lera Boroditsky. 2017. Effects of grammatical gender on object description. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci 2017)*.

Maria D. Sera, Christian A. H. Berge, and Javier del Castillo Pintado. 1994. Grammatical and conceptual forces in the attribution of gender by English and Spanish speakers.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

Cognitive Development, 9(3):261–292. **DOI:** https://doi.org/10.1016/0885-2014(94)90007-8

Maria D. Sera, Chryle Elieff, James Forbes, Melissa Clark Burch, Wanda Rodríguez, and Diane Poulin Dubois. 2002. When language affects cognition and when it does not: An analysis of grammatical gender and classification. *Journal of Experimental Psychology: General*, 131(3):377. **DOI:** https://doi.org/10.1037/0096-3445.131.3.377

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1164

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-u files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA). http://ufal.mff.cuni.cz/udpipe

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, pages 13209–13220.

Gabriella Vigliocco, Marcus Lauer, Markus F. Damian, and Willem J. M. Levelt. 2002. Semantic and syntactic forces in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(1):46. **DOI:** https://doi.org/10.1037/0278-7393.28.1.46

Gabriella Vigliocco, David P. Vinson, Federica Paganelli, and Katharina Dworzynski. 2005. Grammatical gender effects on cognition: Implications for language learning and language use. *Journal of Experimental Psychology: General*, 134(4):501. **DOI:** https://doi.org/10.1037/0096-3445.134.4.501, **PMID:** 16316288

Benj Ide Wheeler. 1899. The origin of grammatical gender. *The Journal of Germanic Philology*, 2(4):528–545.

Nicole Y. Y. Wicha, Elizabeth A. Bates, Eva M. Moreno, and Marta Kutas. 2003. Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, 346(3): 165–168. **DOI:** https://doi.org/10.1016/S0304-3940(03)00599-8

Nicole Y. Y. Wicha, Eva M. Moreno, and Marta Kutas. 2004. Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, 16(7):1272–1288. **DOI:** https://doi.org/10.1162/0898929041920487, **PMID:** 15453979, **PMCID:** PMC3380438

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. **DOI:** https://doi.org/10.18653/v1/D19-1577

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D18-1521

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical*

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5276–5284, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1531, **PMID:** 31191883, **PMCID:** PMC6540912

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661. Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1161

David Zubin and Klaus-Michael Köpcke. 1986. Gender and folk taxonomy: The indexical relation between grammatical and lexical categorization. *Noun Classes and Categorization*, pages 139–180. **DOI:** https://doi.org/10.1075/tsl.7.12zub

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00355/1924142/tacl_a_00355.pdf by guest on 11 October 2024