Measuring the Similarity of Grammatical Gender Systems by Comparing Partitions

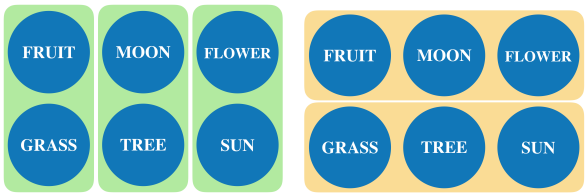
Arya D. McCarthy **Adina Williams** **Shijia Liu** **David Yarowsky** **Ryan Cotterell**
Johns Hopkins University, Facebook AI Research, ETH Zurich
arya@jhu.edu, adinawilliams@fb.com, liu.shij@northeastern.edu,
yarowsky@jhu.edu, ryan.cotterell@inf.ethz.ch

Abstract

A grammatical gender system divides a lexicon into a small number of relatively fixed grammatical categories. How similar are these gender systems across languages? To quantify the similarity, we define gender systems extensionally, thereby reducing the problem of comparisons between languages' gender systems to cluster evaluation. We borrow a rich inventory of statistical tools for cluster evaluation from the field of **community detection** (Driver and Kroeber, 1932; Cattell, 1945), that enable us to craft novel information-theoretic metrics for measuring similarity between gender systems. We first validate our metrics, then use them to measure gender system similarity in 20 languages. Finally, we ask whether our gender system similarities alone are sufficient to reconstruct historical relationships between languages. Towards this end, we make phylogenetic predictions on the popular, but thorny, problem from historical linguistics of inducing a phylogenetic tree over extant Indo-European languages. Languages on the same branch of our phylogenetic tree are notably similar, whereas languages from separate branches are no more similar than chance.

1 Introduction

As many as half the world's languages carve nouns up into classes (Corbett, 2013). In these languages, nouns are subdivided into **gender** categories, which together comprise the language's grammatical gender system. A gender system tends to use a small, fixed number of categories with fixed usage across speakers. Such categories, like 'feminine', can be defined extensionally,[1] and are reflected by agreement with other words within the noun phrase (i.e., **concord**). Gender exhaustively divides up the language's nouns; that is, the union of gender categories is the entire nominal lexicon. Taken this way, a gender system can be viewed as a **partition** of the lexicon into **communities** of same-gendered nouns. Given this, a lexical typologist might naturally wish to ask: how similar are two languages' gender systems?



(a) German, $K = 3$ (b) Spanish, $K = 2$

Figure 1: Two gender systems partitioning $N = 6$ concepts. German (a) has three communities: *Obst* (fruit) and *Gras* (grass) are neuter, *Mond* (moon) and *Baum* (tree) are masculine, *Blume* (flower) and *Sonne* (sun) are feminine. Spanish (b) has two communities: *fruta* (fruit), *luna* (moon), and *flor* are feminine, and *cesped* (grass), *arbol* (tree), and *sol* (sun) are masculine.

Using modern statistical and information-theoretic tools from the community detection literature, we offer the first cluster evaluation (Jardine et al., 1971) perspective on grammatical gender, and quantify the overlap of gender systems. We can compare the pairwise overlap of partitions of gender systems using a rich literature of measures, such as **mutual information** and several variants (Meilă, 2003; Vinh et al., 2010; McCarthy et al., 2019a), which we survey and contrast. Individual partitions of lexicons can also be framed as members of *distributions* over partitions—for instance, the distribution consisting of all partitions of N items, or of all partitions of N items into K gender clusters, as in Figure 1. For example, Spanish is bi-gendered (with masculine and feminine): a lexicon of Spanish nouns ($N = 1000$) and their genders would come from a distribution over partitions of $N = 1000$ items into $K = 2$ clusters.

[1] When we talk about the **extension** of a gender system, we refer to the set of nouns that belong to each gender. This stands in contrast to the **intension** of that gender system, which would be the governing dynamics that gave rise to the particular partitions observed. See §3.

The same lexicon translated into German, a tri-gendered language, would come from a distribution of $N = 1000$ items partitioned into $K = 3$ clusters. Indeed, languages needing different numbers of gender clusters makes this problem non-trivial. From this, we can compare the similarity to what we would expect for *the same lexica* if nouns were randomly supplied with gender specifications. That way, we can distinguish meaningful relationships from noise.

Armed with the first way to quantify community-wise similarity of gender systems, we ask: Do gender system similarities reflect linguistic phylogeny, or something else, like areal effects? Across 20 languages, we find that our pairwise overlap results measurably align with standard pairwise phylogenetic relationships. Zooming in on Indo-European, we find that we can recast pairwise similarities into an accurate phylogenetic tree, simply by measuring distance between gender systems and performing hierarchical agglomerative clustering (see §6.2).

The primary contribution of this work is *a novel metric for lexical typology* that measures the pairwise similarity of gender systems. We operationalize gender systems as partitions over a shared set of nouns (§3). We design and evaluate our measurements of gender system similarity under this formulation (§4), drawing on insights from community detection. Then we recover robust phylogenetic relationships between pairs of gender systems by applying these to 20 gendered languages (§6) and find that similarity between Slavic and Romance gender systems does not exceed chance levels. Finally, we show that our quantification of gender system similarity allows us to construct phylogenetic trees that closely resemble those posited for Indo-European in historical linguistics (e.g., Pagel et al. 2000; Gray and Atkinson 2003; Serva and Petroni 2008).

2 Background: Grammatical Gender

Grammatical gender is a highly fixed classification system for nouns. Native speakers rarely make errors in gender recall, which might tentatively argue against tremendous arbitrary variation (Corbett, 1991). Some regularity can surely be found in the associations between gender and various features of the noun, such as orthographic or phonological form, or semantics. With respect to form-based regularities, Cucerzan and Yarowsky (2003a) devise a system for inferring noun gender (masculine or feminine) from contextual clues and character representations, even in inflected forms of the noun. Nastase and Popescu (2009) also find that phonological form can lead to predictability of gender in two three-gender systems. With respect to word semantics, (Williams et al., 2019) quantify the relationship between the gender on inanimate nouns and their distributional word vectors.

We can't rely on form. Using phonological or orthographic form to derive gender is fraught with complications: particular to our study, nouns (i.e., words that can appear in multiple genders) can pose issues. In German, only gender concord on the definite article and adjectives can disambiguate the gender of some nouns; the same wordform *Band* means "volume" when masculine, but "ribbon" when neuter and "band, musical group" in feminine. Another complication with determining gender from the phonological or orthographic form of the noun is that correspondences between are rarely absolute. For example, even though nouns ending in *-e* are usually 'feminine' in German, this is not universally the case; for example *Affe*, and *Löwe* etc. are masculine. To sidestep these complications, we abstract away from particular word forms and observe the objective consequences of gender over *sets of cross-lingual concepts*, i.e., indices not word forms, and instead compare those across gender systems (see Figure 1).

Which gender systems are likely to be similar? Several accounts highlight similarities between the gender systems of phylogenetically-related languages (Fodor, 1959; Ibrahim, 2014) and argue that they are likely to be at least partially due to historical relations between communities and socio-political factors governing language use. Given this, can we recover phylogenetic similarities across gender systems using our methods? If so, this should provide validation that we are indeed measuring at least some of the genuine similarity that exists between gender systems.

3 Gender Systems as Partitions

Any concept can be related to its referents either intensionally or extensionally. While linguistic research has historically sought to uncover the rules for associating a noun with gender in terms of surface features or semantics (see Corbett 1991 for an overview), we take an extensional approach. That is, we treat a gender category in a language solely as the *set of words it covers*. This

maps directly to the notion of a **community** in the network science task of community detection: A community is defined by membership, not by other arbitrary properties, just as a gender here is defined by the union of all nouns it subsumes, not by its phonological realization or contributions to semantics. The disjoint set of communities forms a **partition** of the set of nouns: Each noun is a member of one and only one cluster.

Although some multi-gendered nouns are present in our investigated languages (see §2), these are very rare. We thus make the simplifying modeling assumption of identifying each word with only a single gender (in our case, the most frequent). This assumption is necessary for our reduction of gender system comparison to clustering evaluation. Without it, we would be forced for words like German *der/die/das Band* to consider overlapping or "fuzzy" partitions, which although an intriguing option, will be left for future work.

Notation. A language's gender system is a partition, named in sans serif (e.g., A). A gender system A has K components called gender classes (i.e., communities, e.g., $\{A_{\text{MSC}}, A_{\text{FEM}}, \ldots\}$); these are in turn sets whose members are items drawn from a finite base set $\mathcal{A} \subseteq \mathcal{L}$, where $\mathcal{A}$ is a sublexicon selected from the full lexicon $\mathcal{L}$. In our case, $\mathcal{A}$ holds all inanimate concepts in our data (see §5). We use Ω to name the set of all partitions of $N = |\mathcal{A}|$ items (in our case, inanimate nouns) into K communities. When comparing two languages' respective gender systems, we will use the letters A and B.

4 Comparing Partitions

A partition groups items into a set of disjoint categories. We could compare any two gender systems (i.e., partitions) which organize the same nouns by determining how similar their gender labelings are. A first pass at quantifying the similarity of two gender partitions would be to measure simple overlap. We could ask: What fraction of $\mathcal{A}$ agrees in gender across languages? That is, for each noun in our multilingual vocabulary, do both languages lexicalize it with the same gender? This is an easily interpretable, accuracy-like measure, bounded by 0 and 1. Still, it has no capacity for comparing systems with *different numbers of categories*; the measure would be handicapped when comparing two-gender systems to three-gender ones.

Comparing systems with different numbers of categories, though, is a well known problem in the field of community detection. While this looks insurmountable from the gender perspective, where gender categories refer to something we recognize, in community detection, the labels themselves are meaningless—there's no notion of a so-called "Cluster 2". The field has circumvented issues arising from comparing systems differing in number of categories by introducing information-theoretic measures to compare partitions. Cluster evaluation functions in community detection are, by and large, based on information-theoretic concepts.

We define a gender system A's entropy as:

$$\mathrm{H}(\mathsf{A}) \stackrel{\text{def}}{=} -\sum_{A \in \mathsf{A}} \frac{|A|}{N} \log \frac{|A|}{N} \tag{1}$$

where we observe the standard convention that $0 \log 0 \stackrel{\text{def}}{=} 0$. How is this notion of entropy for partitions related to the entropy of a probability distribution? These are connected through maximum-likelihood estimation (MLE). In our case, the maximum-likelihood estimate that an inanimate noun a is located in a given partition turns out to be the size of that partition divided by N, e.g. we have $p_{\text{MLE}}(\text{MSC}) = {}^{|A_{\text{MSC}}|}/_{N}$. Recall that the Shannon entropy of a distribution p is defined as

$$\mathrm{H}(p) \stackrel{\text{def}}{=} -\sum_{a \in \mathcal{A}} p(a) \log p(a) \tag{2}$$

We have equality between Eq. 1 and Eq. 2 when we plug the definition of p_{MLE} into Eq. 2, which is why Eq. 1 is considered the entropy of a partition.

4.1 Mutual information (MI)

Mutual information is a workhorse of quantifying similarity between two probability distributions, measuring how much information (in bits) is shared between two random variables. Now we consider the case of the similarity between two partitions. If we have two partition A and B, we may generalize the entropy of a single partition to the mutual information between two partitions as follows:

$$\begin{aligned} \mathrm{I}(\mathsf{A};\mathsf{B}) &\stackrel{\text{def}}{=} \sum_{A \in \mathsf{A}} \sum_{B \in \mathsf{B}} \frac{|A \cap B|}{N} \log \frac{N\,|A \cap B|}{|A|\,|B|} \\ &= \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p_{\text{MLE}}(a,b) \log \frac{p_{\text{MLE}}(a,b)}{p_{\text{MLE}}(a)\,p_{\text{MLE}}(b)} \end{aligned} \tag{3}$$

As the equality above shows, we find, again, that Eq. 3 has an interpretation as the standard definition of probabilistic mutual information applied to

the maximum-likelihood estimate of joint partition membership distribution. To foreshadow future discussion, we note the mutual information between any two clusterings on N items is bounded below by 0 and above by $\log N$. Beyond its interpretation as shared information, mutual information gives little in terms of interpretability: It has no consistent reference points, beyond that the minimum possible MI is zero. Therefore, several variants of MI are preferred in community detection.

Normalization. Furthermore, MI is often normalized to increase its interpretability, as:

$$\mathrm{NMI}(\mathsf{A},\mathsf{B}) \stackrel{\text{def}}{=} \frac{\mathrm{I}(\mathsf{A};\mathsf{B})}{\sqrt{\mathrm{H}(\mathsf{A})\,\mathrm{H}(\mathsf{B})}} \tag{4}$$

While our denominator is the geometric mean, any generalized mean of the partitions' entropies can be used as a bound to normalize MI (Yang et al., 2016). As we divide bits by bits (or nats by nats), normalized mutual information (NMI) is unitless, unlike entropy and MI. It expresses the amount of revealed information as a percentage. Unfortunately, NMI has both theoretical and empirical flaws (Peel et al., 2017; McCarthy, 2017; McCarthy et al., 2019b); namely, it suffers from the *finite-size effect*: the baseline rises as N increases. (Recall that MI is bounded above by $\log N$.) High reward for guessing *even the trivial partition* into singleton clusters rises, making the measure—like vanilla mutual information (as in Eq. 3)—difficult to interpret. For its flaws, we exclude NMI in favor of the following MI-based measures that are both more interpretable and more pertinent.

4.2 Adjusted mutual information (AMI)

Spurious correlations between two gender systems can mislead the results, showing a higher-than-deserved agreement. We select a measure which adjusts for these chance clusterings: the adjusted mutual information (AMI; Vinh et al., 2010). We employ a recent variant (Gates and Ahn, 2017; McCarthy et al., 2019b):

$$\mathrm{AMI}(\mathsf{A},\mathsf{B}) \stackrel{\text{def}}{=} \frac{\mathrm{I}(\mathsf{A};\mathsf{B}) - \mathbb{E}\left[\mathrm{I}(\mathsf{A}';\mathsf{B}')\right]}{\max \mathrm{I}(\mathsf{A}',\mathsf{B}') - \mathbb{E}\left[\mathrm{I}(\mathsf{A}';\mathsf{B}')\right]} \tag{5}$$

where the expectation is taken under the uniform distribution over Ω, all clusterings on N items with K_{A} and K_{B} clusters (Gates and Ahn, 2017). The maximum is also taken over Ω. This distinguishes it from the textbook form of AMI, where the expectation is over a subset of Ω—only those partitions whose community sizes match those of the arguments. As we have subtracted the mean, the expected numerator is centered at 0; the denominator serves to re-normalize the measure. The measure thus compares the mutual information for the observed pair of gender systems to all others within their family. Using AMI also lends some beneficial properties in cluster evaluation:

Remark 1. *AMI has a fixed maximum score* 1.0 *for exactly matching gender systems.*

Remark 2. *The mathematical expectation of AMI is* 0 *so spurious correlations are not rewarded.*

4.3 Variation of Information (VI)

Unlike MI and AMI, Variation of Information (Meilă, 2003) is a **distance** (metric), meaning each language becomes a point in this metric space, whose set is all possible partitions of N items. VI is useful because it satisfies the **triangle inequality** (Meilă, 2007). Additionally, as a metric, it guarantees identity of indiscernibles: if two partitions are at a distance 0, then they are identical. VI is defined as

$$\mathrm{VI}(\mathsf{A},\mathsf{B}) \stackrel{\text{def}}{=} \mathrm{H}(\mathsf{A} \mid \mathsf{B}) + \mathrm{H}(\mathsf{B} \mid \mathsf{A}) \tag{6}$$

and is the summation of two conditional entropies. It can also be normalized by dividing by the joint entropy, $\mathrm{H}(\mathsf{A},\mathsf{B})$. (This measure would be topologically equivalent to Eq. 6.) We do not adjust VI for chance. This would deprive it of its metric property, because of the subtraction in the numerator.

5 Data

Swadesh lists & NorthEuraLex. Our starting point is **Swadesh lists** (Buck, 1949; Swadesh, 1950, 1952, 1955, 1971/2006): concept-aligned minimal inventories of common, "core" or "basic" terminology thought to be "frequent, universal, and resistant to change over time" (Kaplan, 2017). For our purposes, concept-aligned sources are appealing, because they ensure a consistently present base set $\mathcal{A}$ across all our languages, maximizing comparability. We also use the NorthEuraLex dataset (Dellert and Jäger, 2017)—essentially, an extended Swadesh list covering 1016 concepts—to further validate our findings on the original Swadesh lists. Because grammatical gender on animate nouns has the added complication that it generally matches

"natural" gender (or expressed preference) of living creatures across languages (Corbett, 1991; Romaine, 1997; Kramer, 2015), we omit animate nouns to remove semantic confounds from our investigation of cross-lingual gender assignments. We now take the base set $\mathcal{A}$ from the larger concept list in a broader swath of languages. We have 69 inanimate nouns in the Swadesh lists and 387 in NorthEuraLex.

Gender dictionaries. We choose a corpus-based approach to identifying a word's gender. We study the gendered languages available in Universal Dependencies v2.3[2] (Nivre et al., 2018), resulting in a sample of 20 (Hebrew, Greek, Hindi, Lithuanian, Latvian, Polish, Croatian, Slovak, Ukrainian, Russian, Slovenian, Bulgarian, Swedish, Danish, Romanian, French, Catalan, Italian, Spanish, Portuguese). This sample is somewhat skewed based on family, with all but one language (Hebrew) belonging to Indo-European. All are members of the Standard Average European Sprachbund (Whorf, 1997; Haspelmath, 2001), except Hebrew, Hindi, and Greek, which are the only representatives of their groups. Why the Indo-European focus? First, we needed aligned concept lists *with gender and animacy annotations* in languages which possess a gender system. Second, it is natural to test unsupervised methods on a sample with a known *ground truth*. Indo-European phylogeny, while not without its debates, is relatively well studied, making it a strong testbed for verifying our methods. Future work can enable greater linguistic diversity by scraping annotated dictionaries.

Gender labels are drawn from the MarMoT contextual morphological tagger (Müller et al., 2013) trained on Universal Dependencies corpora (Nivre et al., 2018) in each language and applied to Wikipedia in that language. In the case of homophony (particularly with respect to multiple genders) and polysemy, we select the consensus gender (Cucerzan and Yarowsky, 2003b) for the character sequence—its most frequent gender label. We fill gaps manually using bilingual English-target language dictionaries. When multiple words are given to express a concept in a language, we select the most frequent.

[2] German and Arabic were excluded because of complications arising through alignment to annotated dictionaries.

6 Experiments

We apply each measure to the gender systems from our Swadesh lists, then validate our results on NorthEuraLex. We apply validation to ensure that they are picking up robust similarities as opposed to just reflecting properties of particular word lists. (See `github.com/aryamccarthy/gender-partitions`.) We then reconstruct phylogenetic trees of the languages involved. The trees show high agreement with ground truth, compared to random baselines.

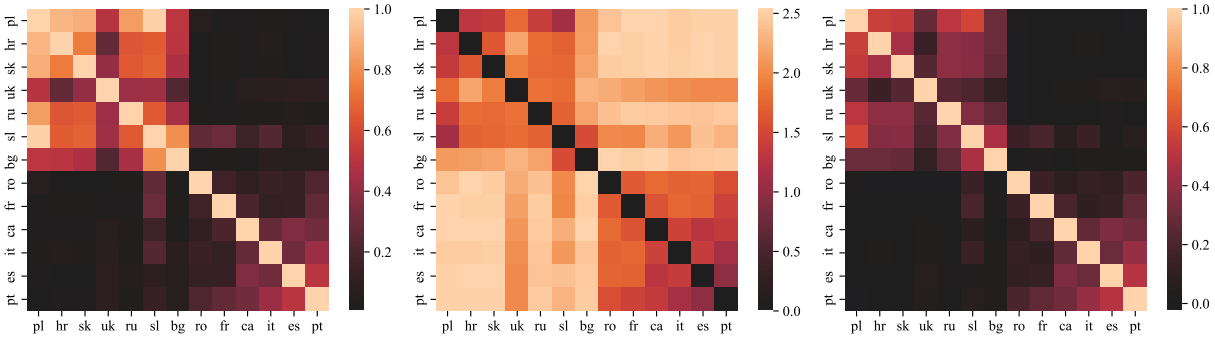
6.1 Similarity measures

We apply the three evaluation measures (§4) to the partitions computed for our languages over the common conceptual lexicon. Figure 2 shows the pairwise scores for languages' gender systems (on the Swadesh list) as partitions. The rows and columns have been reordered according to a "ground truth" of pairwise distances (Serva and Petroni, 2008), for reasons we will explain in the next subsection.[3] Regardless of measure, a few clusters emerge along the diagonal. The (Balto-)Slavic branch (i.e., Polish, Croatian, Slovene, Ukrainian, Slovenian, Russian, and Bulgarian) is present at the top left, and the Romance branch (i.e., French, Catalan, Italian, Spanish, and Portuguese) appears at the bottom right. Outside of these blocks, AMI shows us that the similarity of gender systems is *no better than a chance relationship*; at the whole-lexicon level, influence from the common Indo-European root is absent.

We also apply our measures to the wider swath of languages and larger aligned inventories of NorthEuraLex. The Romance languages again form a block, as do the Balto-Slavic languages. Figure 3 shows similar separation into families for both MI (a) and AMI (c), though this is less pronounced for Variation of Information (b). Variation of Information shows some surprising associations not present in AMI, such as associating Hebrew and Slovene highly with the Romance block.

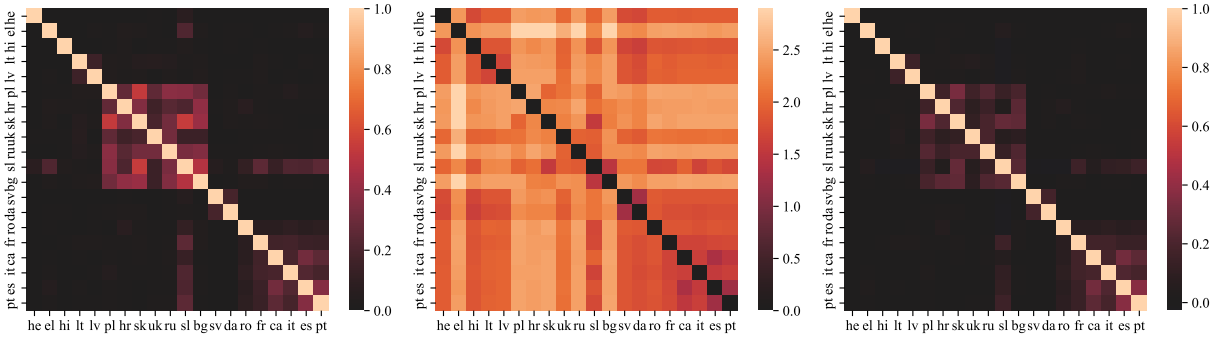
Romanian deserves particular note: It is a Romance language but has been geographically isolated from its family for over a millennium, instead sharing membership in the Balkan Sprachbund with Greek and Bulgarian. As such, we

[3] Selecting a ground truth hierarchy of languages is a contentious and sometimes political matter; even well-accepted trees suffer from criticism (Ringe et al., 2002; Gray and Atkinson, 2003; Greenhill, 2011; Pereltsvaig and Lewis, 2015).



(a) Mutual information (b) Variation of Information (c) Adjusted Mutual Information

Figure 2: Heatmaps uncovered in inanimate Swadesh list under each pairwise similarity measure, grouped by Levenshtein Distance ground-truth phylogenetic trees (Serva and Petroni, 2008). appendix A gives language codes.



(a) Mutual information (b) Variation of Information (c) Adjusted Mutual Information

Figure 3: Heatmaps uncovered in inanimate NorthEuraLex under each pairwise similarity measure, grouped by Levenshtein Distance ground-truth phylogenetic trees (Serva and Petroni, 2008).

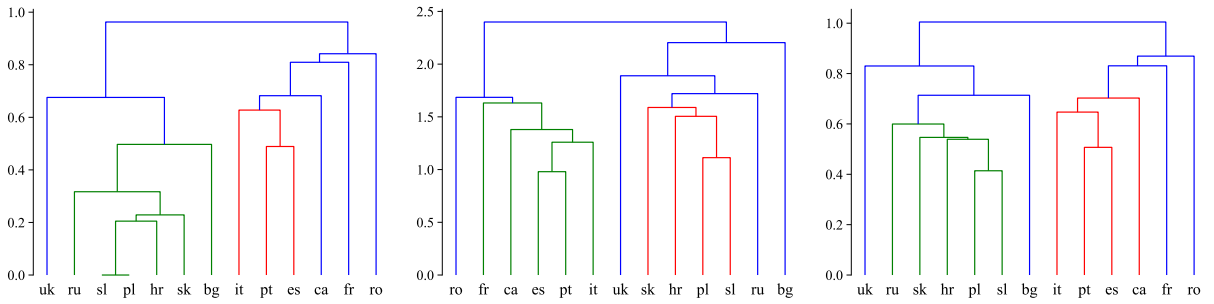
may ask whether its phylogeny or its areal effects are reflected in the gender similarity metrics. While Romanian differs from other Romance languages in many ways (Dinu and Dinu, 2005; Dobrovie-Sorin, 2011)—e.g., it possesses three genders instead of two[4]—it is still more similar to its phylogenetically related Romance relatives than to Balto-Slavic languages. This is easiest to discern in the Variation of Information plot: weak connections surface between Romanian and both Slovene and Ukrainian, but the majority of the Balto-Slavic languages are quite distant from it.

6.2 Phylogeny

Inspired by the findings in the previous section (especially the high similarity among Romance languages), we further validate our measure, asking whether the resulting similarities reflect known phylogenetic ground truth—namely, the developmental history of Indo-European languages. Obviously, there are many more facets to languages' relatedness than their gender systems, so it is interesting to find signal this strong from a single category. Rabinovich et al. (2017) cluster languages based on simple features of their translations into a common target language to craft phylogenetic trees. We take a similar approach, asking whether the pairwise similarities of gender systems are enough to reveal phylogenetic truth or some other relationship. We create phylogenetic trees through agglomerative hierarchical clustering, using both VI and one minus the AMI as distance measures. We use the weighted pair group method of averages (Sokal and Michener, 1958; Müllner, 2011) as implemented in the SciPy library (Jones et al., 2001).

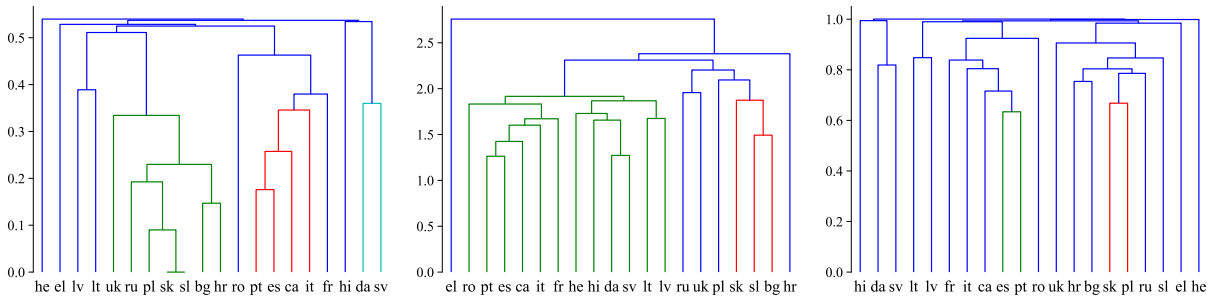
The resulting trees ("dendrograms") can be visualized showing the sequence of cluster formations during hierarchical clustering (Figure 4 and Figure 5). In a dendrogram, any ordering of the leaves maintains fidelity to the computed tree structure, so long as the branching is still correct. We choose to improve upon this by optimally ordering the leaves, swapping subtrees to convey similarity *both within and across* subtrees (Bar-Joseph et al., 2001). On

[4]This claim can be debated (Bateman and Polinsky, 2010): The neuter gender manifests as masculine when singular and feminine when plural (Corbett, 1991).



(a) Mutual information (b) Variation of Information (c) Adjusted Mutual Information

Figure 4: Phylogenies for inanimate Swadesh under each similarity measure. Colors label levels of similarity, with green being most similar, followed by red, then blue (e.g., blue is >70% of max value).



(a) Mutual information (b) Variation of Information (c) Adjusted Mutual Information

Figure 5: Phylogenies for inanimate NorthEuraLex under each similarity measure. Colors label levels of similarity, with green being most similar, followed by red, cyan, and dark blue (e.g., dark blue is >70% of max value).

the whole, our dendrograms recover known phylogenetic relationships between the languages we consider; this serves to largely validate our measures as having uncovered some meaningful similarity between the languages' gender system. Indeed, in every case, we reconstruct the subtree of Romance languages with high fidelity. The only difference is that on NorthEuraLex, Catalan is more similar to Portuguese and Spanish than Italian is. In all trees, Romanian is always grouped with the Romance languages, matching its ancestry. The Balto-Slavic subtree is less perfect. MI and AMI recover similarities between Russian and Ukrainian (Eastern Slavic), Slovak and Polish (Western Slavic), and Croatian and Bulgarian (South Slavic) fairly well. Further, the Slavic and Baltic languages are properly joined to form a Balto-Slavic group. We take this as validation of our method.

When measuring with Variation of Information, though, things go awry. While it correctly pairs Russian and Ukrainian and recreates the same Romance subtree as the other measures, there are some major discrepancies. Hebrew, the only non–Indo-European language, is found to be closer to the Romance languages than to the Balto-Slavic cluster. Hindi's closeness to others is similarly exaggerated. In fact, everything seems to be close for VI, except Greek! As the other measures better capture the phylogeny, we suggest that similarity measured with Variation of Information is ill suited to our main task.

6.3 Quantitative Evaluation

Our proposals to measure similarity of gender systems give rise to dendrograms that resemble phylogenetic trees. But how much so? We answer this by measuring the similarity to the ground truth tree. To measure the similarity of two trees T_1 and T_2, we use Rabinovich et al. (2017)'s extension of the L_2 norm to leaf pair distance. Here, we sum the number of edges on a path between two nodes to get their distance d. We then compute the total distance as the sum of squared distances: $\sum_{i \neq j} \left(d_{T_1}(\ell_i, \ell_j) - d_{T_2}(\ell_i, \ell_j)\right)^2$, where each ℓ_i identifies one language (or leaf).

We show that the distance according to any of our three measures is significantly more like the ground truth (from Serva and Petroni, 2008) than chance by comparing the computed trees to 1000 randomly generated trees on the same set of languages. (We report mean and standard deviation of

distance from the ground truth. We use Rabinovich et al. (2017)'s unweighted distance.) For each combination of dataset and measure, we use McNemar's test for significance and find $p < 0.0001$.

7 Related Work

There is a baffling dearth of work on quantifying similarity of gender systems. There is, however, ample work on characterizing intensional gender systems, i.e., sets of grammatical rules, that can be divided (Corbett, 1991) into sets of rules based on morphology (Tucker et al., 1977; Gregersen, 1967; Wald, 1975; Plank, 1986, i.a.) and on phonology (Bidot, 1925; Tucker et al., 1977; Newman, 1979; Hayward and Corbett, 1988; Marchese, 1988). Intensional approaches, particularly those with typological leanings, contribute very fine grained research on particular pairwise similarities for particular languages and dialects. Although we cannot survey these in detail here, we would love for our measures to contribute findings that can complement these approaches.

Relatedly, other recent works have investigated grammatical gender and other types of noun classification systems with information theoretic tools. For example, Williams et al. 2020b uses mutual information to quantify the strength of the relationships between declension class, grammatical gender, distributional semantics, and orthographic form respectively in several languages. Williams et al. 2020a, which is arguably closest to this work, measures the strength of semantic relationships between inanimate nouns and verbs or adjectives that takes those nouns as arguments, and that work can be seen as comparing the similarity of nouns clustered by their gender, with the same nouns clustered by the adjectives that modify them or the verbs that take them as arguments.

Although we adopt information theoretic measures, here there are two other major classes of cluster evaluation measures: set-matching measures, and pair-counting measures, which tally which pairs of items are in the same or different communities. One popular set-matching measure in information retrieval, *purity* (Manning et al., 2008), is asymmetric and biased by the size and number of communities (Danon et al., 2005). Its symmetric form, the F-measure (Artiles et al., 2007), has clear bounds but gives no indication of average-case performance.

The adjusted Rand index (ARI; Hubert and Arabie, 1985) is the preeminent pair-counting measure. It is related to AMI, adjusting the Rand index in the same way that AMI adjusts MI. ARI also computes an expectation, which can be computed over the proper distribution (Gates and Ahn, 2017), but it is empirically better suited to large, balanced clusters. In our case of small and uneven clusters, AMI should be preferred (Romano et al., 2016).

Dataset	Measure	Score	St. Dev.
Swadesh	MI	344	-
	VI	312	-
	AMI	344	-
	Random	1184	133.4
NorthEuraLex	MI	1231	-
	VI	1164	-
	AMI	1548	-
	Random	2531	209.6

Table 1: Distances of generated trees from gold tree.

We can only survey a representative handful of the numerous cluster evaluation measures in the limited space we have here. See McCarthy et al. (2019b) for an outline of desiderata for comparing partitions, as well as a general class of appropriate measures, and for further motivation for AMI using a different null model—languages have a fixed number of gender classes, so we select one over N items with K communities, rather than an arbitrary number of communities.

8 Conclusion

We have presented a clean method for comparing grammatical gender systems across languages: By defining gender classes extensionally, we reduced the problem to cluster evaluation from community detection. We validate three metrics by recovering known phylogenic relationships in our languages, with measurable success. Separate Indo-European branches are no more similar than chance.

We emphasize that our methods are not specifically tailored to gender systems. One could apply them more broadly other aspects of the lexicon, e.g. to Indo-European verb classes, Bantu noun classes, or diachronic time slices of a single language's gender system, data permitting. A related challenge is East and Southeast Asian numeral classifier systems, which associate nouns with classifiers based largely on the semantic properties of the nouns (Kuo and Sera, 2009; Zhan and Levy, 2018; Liu et al., 2019). They display more idiolectal variation, and often more than one classifier can accompany a given noun (Hu, 1993), unlike for gender (where

this is rare). We note that we could further extend our measures to fuzzy partitions, which remain less explored in community detection, but are a promising avenue for future work.

Acknowledgments

We thank Tongfei Chen for comments on the Slavic languages, Jean-Gabriel Young for suggesting that we consider Variation of Information, and Johannes Bjerva for providing us with code to compute the tree distance. We also thank Tiago Pimentel for his help with proofreading. Finally, we would like to thank Eleanor Chodroff for providing useful insights during the formulation of the problem.

References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The SemEval-2007 WePS evaluation: Establishing a benchmark for the Web people search task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 64–69. Association for Computational Linguistics.

Ziv Bar-Joseph, David K. Gifford, and Tommi S. Jaakkola. 2001. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17:S22–S29.

Nicoleta Bateman and Maria Polinsky. 2010. Romanian as a two-gender language. *Hypothesis A/Hypothesis B: Linguistic Explorations in Honor of David M. Perlmutter*.

Emile Bidot. 1925. *La clef du genre des substantifs français: méthode dispensant d'avoir recours au dictionnaire*. Imprimerie Nouvelle.

Carl Buck. 1949. *A Dictionary of Selected in the Principal Indo-European Languages*. University of Chicago Press.

Raymond B. Cattell. 1945. The description of personality: Principles and findings in a factor analysis. *The American Journal of Psychology*, 58(1):69–90.

Greville G. Corbett. 1991. *Gender*. Cambridge University Press., Cambridge.

Greville G. Corbett. 2013. Number of genders. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Silviu Cucerzan and David Yarowsky. 2003a. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Silviu Cucerzan and David Yarowsky. 2003b. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 40–47.

Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. 2005. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008–P09008.

Johannes Dellert and Gerhard Jäger. 2017. NorthEuraLex. Version 0.9.

Anca Dinu and Liviu P. Dinu. 2005. On the syllabic similarities of Romance languages. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 785–788. Springer.

Carmen Dobrovie-Sorin. 2011. *The syntax of Romanian: Comparative studies in Romance*, volume 40. Walter de Gruyter.

Harold Edson Driver and Alfred Louis Kroeber. 1932. *Quantitative expression of cultural relationships*, volume 31. University of California Press.

Istvan Fodor. 1959. The origin of grammatical gender. *Lingua*, 8:186–214.

Alexander J. Gates and Yong-Yeol Ahn. 2017. The impact of random models on clustering similarity. *Journal of Machine Learning Research*, 18(87):1–28.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the anatolian theory of Indo-European origin. *Nature*, 426:435.

Simon J. Greenhill. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, 37(4):689–698.

Edgar A. Gregersen. 1967. *Prefix and pronoun in Bantu*. Published at the Waverly Press by Indiana University, Bloomington.

Martin Haspelmath. 2001. The European linguistic area: Standard average European. In *Language typology and language universals. (Handbücher zur Sprach-und Kommunikationswissenschaft)*, pages 1492–1510. de Gruyter.

Richard J. Hayward and Greville G. Corbett. 1988. Resolution rules in Qafar. *Linguistics*, 26:259–279.

Qian Hu. 1993. *The Acquisition of Chinese Classifiers by Young Mandarin-speaking Children The Acquisition of Chinese Classifiers by Young Mandarin-speaking Children*. Ph.D. thesis, Boston University.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Muhammad Hasan Ibrahim. 2014. *Grammatical gender: Its origin and development*, volume 166. Walter de Gruyter.

N. Jardine, P. H. P. S. N. Jardine, and R. Sibson. 1971. *Mathematical Taxonomy*. Wiley Series in Probability and Mathematical Statistics. Wiley.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001. SciPy: Open source scientific tools for Python.

Judith Kaplan. 2017. From lexicostatistics to lexomics: Basic vocabulary and the study of language prehistory. *Osiris*, 32(1):202–223.

Ruth T. Kramer. 2015. *The Morphosyntax of Gender*, volume 58. Oxford University Press.

Jenny Y. Kuo and Maria D. Sera. 2009. Classifier effects on human categorization: the role of shape classifiers in Mandarin Chinese. *Journal of East Asian Linguistics*, 18:1–19.

Shijia Liu, Hongyuan Mei, Adina Williams, and Ryan Cotterell. 2019. On the idiosyncrasies of the Mandarin Chinese classifier system. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4100–4106, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Lynell Marchese. 1988. Noun classes and agreement systems in Kru: A historical approach. *Agreement in Natural Language: Approaches, Theories, Descriptions. Stanford: Center for the Study of Language and Information*, pages 323–341.

Arya D McCarthy. 2017. Gridlock in networks: The leximin method for hierarchical community detection. Master's thesis, Southern Methodist University.

Arya D. McCarthy, Tongfei Chen, and Seth Ebner. 2019a. An exact no free lunch theorem for community detection. In *Complex Networks and Their Applications VIII*, pages 176–187, Lisbon, Portugal. Springer International Publishing.

Arya D. McCarthy, Tongfei Chen, Rachel Rudinger, and David W. Matula. 2019b. Metrics matter in community detection. In *Complex Networks and Their Applications VIII*, pages 164–175, Lisbon, Portugal. Springer International Publishing.

Marina Meilă. 2003. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, pages 173–187, Berlin, Heidelberg. Springer Berlin Heidelberg.

Marina Meilă. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.

Vivi Nastase and Marius Popescu. 2009. What's in a name? In some languages, grammatical gender. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1377. Association for Computational Linguistics.

Paul Newman. 1979. Explaining Hausa feminines. *Studies in African Linguistics*.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, et al. 2018. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

M. Pagel, C. Renfrew, A. McMahon, and L. Trask. 2000. Time depth in historical linguistics. *C. Renfrew, A. McMahon, and L. Trask, editors*, pages 189–207.

Leto Peel, Daniel B. Larremore, and Aaron Clauset. 2017. The ground truth about metadata and community detection in networks. *Science Advances*, 3(5).

A. Pereltsvaig and M. W. Lewis. 2015. *The Indo-European Controversy*. Cambridge University Press.

Frans Plank. 1986. Paradigm size, morphological typology, and universal economy. *Folia Linguistica*, 20(1-2):29–48.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540. Association for Computational Linguistics.

Don Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Suzanne Romaine. 1997. Gender, grammar, and the space in between. *Pragmatics and Beyond: New Series*, pages 51–76.

Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. 2016. Adjusting for chance clustering comparison measures. *Journal of Machine Learning Research*, 17(1):4635–4666.

Maurizio Serva and Fabio Petroni. 2008. Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

Robert Reuven Sokal and Charles Duncan Michener. 1958. *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin. University of Kansas.

Morris Swadesh. 1950. Salish internal relationships. *International Journal of American Linguistics*, 16(4):157–167.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, 96(4):452–463.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137.

Morris Swadesh. 1971/2006. *The origin and diversification of language*. Chicago: Aldine.

G. R. Tucker, W. E. Lambert, and A. Rigault. 1977. *The French speaker's skill with grammatical gender: an example of rule-governed behavior*. Janua Linguarum: Series didactica. Mouton.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854.

Benji Wald. 1975. Animate concord in Northeast Coastal Bantu: Its linguistic and social implications as a case of grammatical convergence. *Studies in African linguistics*, 6(3):267–314.

Benjamin Lee Whorf. 1997. *The Relation of Habitual Thought and Behavior to Language*, pages 443–463. Macmillan Education UK, London.

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 5733–5738, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2020a. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*.

Adina Williams, Tiago Pimentel, Hagen Blix, Arya D. McCarthy, Eleanor Chodroff, and Ryan Cotterell. 2020b. Predicting declension class from form and meaning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6682–6695, Online. Association for Computational Linguistics.

Zhao Yang, René Algesheimer, and Claudio J. Tessone. 2016. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6(1):30750.

Meilin Zhan and Roger Levy. 2018. Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1997–2005, New Orleans, Louisiana. Association for Computational Linguistics.

A Languages

While there are over 70 languages in the Universal Dependencies treebanks, only a select handful possess grammatical gender. We use 20 languages in the Universal Dependencies corpora that have gender and also present in our concept lists. Below find their ISO 639-1 codes (used in the paper to conserve space), ISO 639-3 codes (widely preferred), and their major family (in the case of Hebrew) or subfamily (in the case of our Indo-European languages), and the number of grammatical genders they have:

Language	ISO 639-1	ISO 639-3	(Sub-)Family	Genders
Bulgarian	bg	bul	Balto-Slavic	3
Catalan	ca	cat	Romance	2
Danish	da	dan	Germanic	2
Greek	el	ell	Hellenic	3
Spanish	es	spa	Romance	2
9 French	fr	fra	Romance	2
Hebrew	he	heb	Semitic	2
Hindi	hi	hin	Indo-Iranian	2
Croatian	hr	hrv	Balto-Slavic	3
Italian	it	ita	Romance	2
Lithuanian	lt	lit	Balto-Slavic	2
Latvian	lv	lav	Balto-Slavic	2
Polish	pl	pol	Balto-Slavic	3
Portuguese	pt	por	Romance	2
Romanian	ro	ron	Romance	3
Russian	ru	rus	Balto-Slavic	3
Slovak	sk	slk	Balto-Slavic	3
Slovene	sl	slv	Balto-Slavic	3
Swedish	sv	swe	Germanic	2
Ukrainian	uk	ukr	Balto-Slavic	3

Table 2: Languages, with their subfamilies and ISO codes, used in this study.