Cross-lingual Embeddings Reveal Universal and Lineage-Specific Patterns in Grammatical Gender Assignment

Hartger Veeman, Marc Allassonnière-Tang, Aleksandrs Berdicevskis, Ali Basirat

Cross-lingual Embeddings Reveal Universal and Lineage-Specific Patterns in Grammatical Gender Assignment

Hartger Veeman
Uppsala University
Department of linguistics and philology
Box 635, 75126 Uppsala
hartger.veeman.7544@student.uu.se

Marc Allassonnière-Tang
University Lyon 2
Lab Dynamics of Language
14 avenue Berthelot 69363 Lyon
marc.tang@univ-lyon2.fr

Aleksandrs Berdicevskis
University of Gothenburg
Språkbanken
Box 200, 40530 Gothenburg
aleksandrs.berdicevskis@gu.se

Ali Basirat
Uppsala University
Department of linguistics and philology
Box 635, 75126 Uppsala
ali.basirat@lingfil.uu.se

Abstract

Grammatical gender is assigned to nouns differently in different languages. Are all factors that influence gender assignment idiosyncratic to languages or are there any that are universal? Using cross-lingual aligned word embeddings, we perform two experiments to address these questions about language typology and human cognition. In both experiments, we predict the gender of nouns in language X using a classifier trained on the nouns of language Y, and take the classifier's accuracy as a measure of transferability of gender systems. First, we show that for 22 Indo-European languages the transferability decreases as the phylogenetic distance increases. This correlation supports the claim that some gender assignment factors are idiosyncratic, and as the languages diverge, the proportion of shared inherited idiosyncrasies diminishes. Second, we show that when the classifier is trained on two Afro-Asiatic languages and tested on the same 22 Indo-European languages (or vice versa), its performance is still significantly above the chance baseline, thus showing that universal factors exist and, moreover, can be captured by word embeddings. When the classifier is tested across families and on inanimate nouns only, the performance is still above baseline, indicating that the universal factors are not limited to biological sex.

1 Grammatical gender assignment

Grammatical gender is one of the nominal classification systems found in natural languages (Seifart, 2010). In languages with grammatical gender, certain words agree in a specific form with the noun they modify depending on the gender of the modified noun (Corbett, 1991, 2001). For instance, Swedish has a binary gender system with the common/neuter values, in which the articles and adjectives must have grammatical gender agreement with the noun they are modifying, c.f., *ett stor-t äpple* (SG.NEUT big-SG.NEUT apple.SG.NEUT) 'a big apple' and *en stor-∅ häst* (a.SG.UTER big-SG.UTER horse.SG.UTER) 'a big horse'.

The most common gender distinctions are masculine/feminine (e.g., in French and Italian), masculine/feminine/neuter (e.g., in Russian and German), and common/neuter (e.g., in Swedish and Danish) (Corbett, 2013a).[1] Within these distinctions, grammatical gender does not necessarily fully agree with the biological sex. By way of illustration, the German word for 'girl', *Mädchen*, is a neuter noun. Moreover, nouns with the same meaning may belong to different grammatical gender in different languages. For example, the German noun for 'sun', *Sonne*, is feminine, but its French equivalent, *soleil*, is masculine.

Based on these observations, several questions have been developed in the literature. First of all, what are the main factors that influence gender assignment in individual languages? Second, are there any principles that are shared cross-linguistically or are they all language- and/or culture-specific? With regard to the first question, two main factors have been identified in the literature: formal features of the noun and its meaning (Corbett and Fraser, 2000; Rice, 2006; Corbett, 2013b; Fedden and Corbett, 2019). With regard to the second question, it is generally believed that

[1] More complex distinctions are also found. As an example, Swahili has a more complex system with 18 classes. These systems are generally referred to as 'noun classes' in the literature and are not covered by the term 'grammatical gender' in the current paper.

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 265–275

https://doi.org/10.18653/v1/P17

gender assignment is based on a mix of shared cognitive principles (Kemmerer, 2017) and linguistic/cultural idiosyncrasies (Takamura et al., 2016; Di Garbo et al., 2019).

One of the most common way to answer the second question is to measure the transferability of gender across languages. If the gender assignment rules of a language can be easily used to predict the gender of nouns in another language, it shows that the principles of gender assignment are partly shared and transferable between the two languages. If such a transfer is possible within most languages, one can then assume that gender assignment is to a large extent based on universal patterns. Most empirical studies that adopted this approach followed the perspective of language acquisition and analyzed how native speakers of language X could predict the gender of a selected amount of nouns from language Y (Sabourin et al., 2006; Jarvis and Pavlenko, 2010, p.132-136). No studies known to the authors investigated the transferability of grammatical gender for a large amount of nouns from a large sample of languages by using natural language processing methods, which is the gap we aim at filling.

We use a transfer learning setting to measure the transferability of grammatical gender across languages. In this setting, a neural classification model is trained to predict the grammatical gender of nouns in a source language. This model is then applied to a set of test nouns in a target language to predict their grammatical gender. The classifier's ability to classify the test nouns is interpreted as an indication of the transferability of grammatical gender system from the source language to the target language (i.e., the higher the accuracy is, the more transferable the gender systems are). The entire setting is founded on the cross-lingual representation of words, providing for knowledge transfer between the gender classification models across languages. The embeddings are used to represent nouns in both source and target languages. The use of word embeddings for the study of grammatical gender is based on the premise that they can capture linguistically-motivated information about words (Nastase and Popescu, 2009; Andreas and Klein, 2014; Artetxe et al., 2018; Basirat and Tang, 2019; Williams et al., 2019), including information about the gender of nouns within a language (Basirat and Tang, 2019; Williams et al., 2019; Nastase and Popescu, 2009; Basirat et al., in press).

We ask the following research questions. Is successful gender transfer possible between non-related languages (if yes, it means that there exist universal factors in gender assignment)? When the classifier is applied to related languages, does its success depend on how related they are (if yes, this in an indication that some factors are not universal)? Does gender transfer work in the same way for all nouns or are there differences between certain noun classes?

2 Experimental materials and settings

In this section, we present the languages involved in this study along with the source of our data. Then, we provide an overview of the cross-lingual word embedding method and the settings of the classifier used for gender transfer.

2.1 Materials

Two sources of data are selected for each language. First, a noun-gender dictionary is constructed from the morphological annotations of the Universal Dependencies 2.6 (Zeman et al., 2020). For each language, a dictionary is created by iterating through all available UD treebanks for the given language. For every unique downcased noun form in these treebanks, the grammatical gender is extracted from the treebank and labeled to the noun as one of four classes: neuter, feminine, masculine and common (underspecified values such as "Fem,Masc" were ignored). The same four-class label structure was used for all languages, to ensure compatibility of the models. Second, word embeddings are selected from pre-trained cross-lingual embeddings published on the fastText website (Joulin et al., 2018).[2] Further details about the embeddings are provided in the following subsection.

We selected all languages that have grammatical gender and are present in both data sources, with the exception of Albanian due to its small treebank size and Norwegian because in pilot experiments, our classifier showed unexpectedly poor performance for reasons we were not able to establish. This results in the selection of 24 languages that are shown in Table 1. Three types of gender systems are found: masculine/feminine (42%, 10/24), masculine/feminine/neuter (46%, 11/24), and common/neuter (12%, 3/24). Only a few languages belong to the third type, which is actually the result of a merge between the masculine and the feminine

[2] https://fasttext.cc/docs/en/aligned-vectors.html

categories existing originally in those languages (Enger, 2017, p.1439). The Indo-European language family is over-represented, which is due to practical limitations from the available resources. Nevertheless, our sample has its advantages. First, the Indo-European language family is considered to be one of the 'typical' grammatical gender language families (Audring, 2016, p.2), which represents an ideal starting point for a quantitative analysis. Second, comparing languages mostly from the same family allows us to address our second research question about the correlation between relatedness of the languages and the transferability of gender.

Language		m	f	n	c	Size
Arabic*	ar	33	67	-	-	3
Bulgarian	bg	24	33	43	-	9
Catalan	ca	49	51	-	-	9
Czech	cs	17	41	43	-	44
Danish	da	-	-	72	28	7
German	de	24	40	36	-	56
Greek	el	25	52	23	-	4
Spanish	es	44	56	-	-	1
French	fr	43	57	-	-	13
Hebrew*	he	44	57	-	-	6
Hindi	hi	33	67	-	-	8
Croatian	hr	17	39	45	-	12
Italian	it	45	55	-	-	13
Lithuanian	lt	38	62	-	-	7
Latvian	lv	49	51	-	-	13
Dutch	nl	-	-	72	28	9
Polish	pl	21	35	45	-	26
Portuguese	pt	45	55	-	-	8
Romanian[3]	ro	64	37	-	-	17
Russian	ru	17	34	49	-	44
Slovak	sk	18	39	43	-	9
Slovenian	sl	16	41	43	-	12
Swedish	sv	-	-	75	25	11
Ukrainian	uk	14	38	48	-	11

Table 1: Languages included in the data. Asterisk (*) denotes Afro-Asiatic languages, the rest are Indo-European. The gender distribution (in %) is shown in columns m = masculine, f = feminine, c = common (uter), n = neuter. The "Size" column indicates the number of nouns in thousand tokens (K).

[3]Traditionally, Romanian is considered to have three genders: masculine, feminine, and neuter, but an alternative two-gender analysis has also been proposed (Bateman and Polinsky, 2010). UD follows the two-gender annotation.

2.2 Cross-lingual Word Embeddings

Cross-lingual word embeddings aim at representing words of multiple languages in a joint embedding space such that similar words (in each language and across all languages) are clustered together. These resources provide a foundation for the cross-lingual study of words and the development of transfer learning models between languages.

The cross-lingual word embeddings can be trained in different ways (Ruder et al., 2019). One of the main approaches is to find a mapping between monolingual word embedding spaces using a seed dictionary that contains words and their translations in different languages. This approach is based on the observation made by Mikolov et al. (2013) that word embeddings exhibit similar structures across languages. Mikolov et al. (2013) formulate the mapping as a least-square linear regression between the monolingual embeddings of the seed lexicon to minimize the mean square error of the word translations. The mapping model is then generalized to all words in the languages. This approach is improved by Xing et al. (2015); Smith et al. (2017), imposing an orthogonal constraint on the transformation weights. Later attempts were made to reduce the need for the seed dictionary (Smith et al., 2017; Artetxe et al., 2017). Conneau et al. (2018); Zhang et al. (2017) leverage adversarial training to automatically produce the dictionary during training and completely eliminate its necessity as a supervision source. Joulin et al. (2018) further enhance the loss function of the regression model using the retrieval model of Conneau et al. (2018), providing for the representation of unseen words.

In this study, we use fastText cross-lingual word embeddings trained on the monolingual word embeddings of Bojanowski et al. (2017) using the mapping approach of Joulin et al. (2018). The monolingual embeddings are trained on Wikipedia data for words that appear at least five times. The embeddings encode information about the form and semantics of words from sub-word units and word co-occurrences, respectively. The information about the form and semantics plays a critical role in the assignment of grammatical gender to nouns (Corbett, 1991; Rice, 2006). This motivates us to use fastText embeddings for the study of cross-lingual grammatical gender transfer. The original embeddings are distributed in a 300-dimensional space and cover 44 languages belonging to differ-

ent language families. In the current study, we retrieved the embeddings for the 24 languages that have gender systems and a sufficiently large data size.

2.3 Settings

A multi-layer perceptron is used to predict the grammatical gender of nouns from their cross-lingual embeddings. The choice of a multi-layer perceptron instead of a recurrent model is motivated by 1) the fact that the grammatical gender is an inherent static property of a noun that does not change in different contexts, and 2) the proven ability of a multi-layer perceptron for the task (Basirat and Tang, 2019). The network has three layers, an input layer that reads the 300-dimensional word embeddings, a single hidden layer twice the size of the input layer with ReLu activation, and an output layer with softmax activation consisting of four neurons related to the four genders masculine, feminine, neuter, and common. This provides for modeling the three gender systems masculine/feminine, masculine/feminine/neuter, and common/neuter and analysing the extent to which these systems are transferable.

The classifier is trained on pairs of the noun embeddings and genders collected from the dictionary. The data is split into 80%, 10% and 10% for training, validation and testing. The data is randomly split in folds, of which the designations of training, test and validation are rotated between runs. The final results are the average of multiple runs covering a full rotation of the folds. We label the language the classifier is trained on *source* and the language to which it attempts predicting gender *target*. The train and the validation data is used for training a classification model on the source language and the test data is used for testing the model on the target language. We go through all possible source-target combinations, 576 (24×24) language pairs.

PyTorch (Paszke et al., 2019) is used to implement the classifier using the stochastic gradient descent optimizer with a learning rate of 0.1 and the cross-entropy loss function. Early stopping was employed if the model stopped improving over 20 epochs, with a minimum of 2200 epochs and a maximum of 25000 epochs. We ran the classifier ten times with different random seeds to measure the variability between training runs. For every language pair, we calculate Fleiss' kappa across the ten runs. The kappas vary from 0.73 (substantial agreement) to 0.99, the average value is 0.91 (almost perfect agreement), and the standard deviation is 0.04. We conclude that the results are robust with respect to random seed.

3 Gender transfer at the language level

In this section, we analyse the results of the experiment from two different perspectives. First, we consider the broad transferability of gender across languages by measuring the accuracy across all possible pairs of languages in the dataset (section 3.1, figures 1 and 3). While this step provides an overview of the accuracy of gender transfer, it is also extremely influenced by the different gender systems across languages. For instance, asking a language that has the masculine/feminine system to predict the categories on a language that has a masculine/feminine/neuter system is by definition going to result in a low accuracy since the source language does not have information about neuter nouns. To overcome this issue, we perform an additional analysis with narrower scope, where we compare pairs of only those languages that have isomorphic systems (section 3.2). As an example, we use languages that have masculine/feminine to predict the gender in languages that also have masculine/feminine. The rationale behind narrowing the scope is that it enables us to focus on the question of how similar the distributions of nouns across gender classes are, abstracting away from possible differences between the number of classes and their types.

We define a random guessing baseline for the transfer between each pair of languages. The baseline is the accuracy that would have been achieved by a classifier that makes a random guess based solely on gender probabilities in the source language. The accuracy it would achieve is

$$\sum_{g \in \{m,f,c,n\}} p(g_s)p(g_t)$$

where $p(g_s)$ is the probability of the given gender in the source language and $p(g_t)$ is the probability of the given gender in the target language. In all our experiments, we report the absolute improvement (or degradation) of the transfer learning accuracy from the random baseline accuracy. In this way, a negative value indicates that the transfer accuracy is below the baseline, a positive value indicates that the transfer accuracy is higher than the baseline,

and a zero value indicates that the transfer accuracy is only as good as the random baseline.[4]

3.1 Gender transfer between all systems

The mean accuracy from the ten different-seed runs of each transfer is compared with the random baseline and plotted in Figure 1. Each entry is the difference between the result and the baseline, i.e. an improvement (or degradation) from the baseline results.

We run three two-sided paired t-tests to check whether the accuracy is significantly different from the baseline: one for language pairs within the Indo-European family ($t(483) = 27.013$, *p-value* < 0.001), one for language pairs where the source language is from an Afro-Asiatic family and the target language is Indo-European ($t(43) = 10.454$, *p-value* < 0.001), one for language pairs where the source is Indo-European and the target is Afro-Asiatic ($t(43) = 9.5251$, *p-value* < 0.001), in all cases the average classifier accuracy is higher than the baseline.

Three main observations are worth noting. First, word embeddings do provide sufficient information to generate an accuracy significantly above the baseline for the majority of languages, as shown by the diagonal line in the plot and the output of the t-tests. Second, Danish, Dutch, and Swedish do not transfer well to other languages. These three languages are the only languages that have a common/neuter system, which explains the low accuracy of gender transfer. Third, even though Arabic and Hebrew are not related to the Indo-European language family, both as source and target languages they yield accuracy that is comparable to that yielded by Indo-European languages and is significantly higher than the baseline (see Section 4). These points imply that while the relatedness of languages affect the transferability of gender, we are also likely to find some shared principles of gender assignment across non-related languages.

Then, we compare the accuracy of the transfer with phylogenetic distance within the Indo-European language family. To do so, we extract the phylogenetic distance from the broad Indo-European tree published by Chang et al. (2015, tree A3), as shown in Figure 2. The branch lengths are annotated in terms of years, which allows a direct comparison of phylogenetic distance defined as the time depth of the first common ancestor shared by a pair of compared languages. The larger the distance, the less related is a pair of languages. The output of the comparison is shown in Figure 3.

A linear regression shows a significant relationship between accuracy and phylogenetic distance ($t(4838) = -60.12$, $p < 0.001$). The slope coefficient for phylogenetic distance is -0.00005, which means that the accuracy decreases by 5% for each 1000 years of phylogenetic distance. The R^2 value shows that 43% of the variation in accuracy can be explained by phylogenetic distance. A closer analysis indicates that the transfer accuracy of a pair of languages sharing the same system is generally higher than the transfer accuracy of pair of languages having different systems. For instance, the lower values between 1000 and 2000 years of phylogenetic distance are gender transfers between common/neuter and masculine/feminine/neuter languages. Further details are explained in subsection 3.2.

Finally, we performed a correlation test to see whether gender transferability in a language pair is affected by how different gender distributions in the two languages are. By gender distribution we mean a distribution of the marginal probabilities of seeing each gender over all nouns in the vocabulary set, and we measure the difference between two distributions as the KL-divergence. We find that the transferability correlates negatively with the KL-divergence both globally over all languages (Spearman $\rho = -0.7$, $p < 0.001$) and locally within each branch (Slavic: $\rho = -0.4$, $p = 0.002$, Germanic: $\rho = -0.9$, $p < 0.001$, and Romance: $\rho = -0.8$, $p < 0.001$), indicating that gender transfer becomes weaker as the marginal distributions of gender become more different.

3.2 Gender transfer between isomorphic systems

Three types of comparisons are made. First, the accuracy of transfer between languages with masculine/feminine systems is measured. Then, the same process is conducted for languages with masculine/feminine/neuter systems and common/neuter systems.

To estimate the combined effect of phylogenetic distance and system type (masculine/feminine,

[4]Note that our measure is not of course a perfect quantification of gender-system similarity, since it does not yield the accuracy of 1 for all the cases when source and target languages are the same (the diagonal in Figure 1). It can, however, be viewed as an approximation (in principle, the accuracy of transfer X $\rightarrow$ Y can be normalized by dividing it by the accuracy of X $\rightarrow$ X).

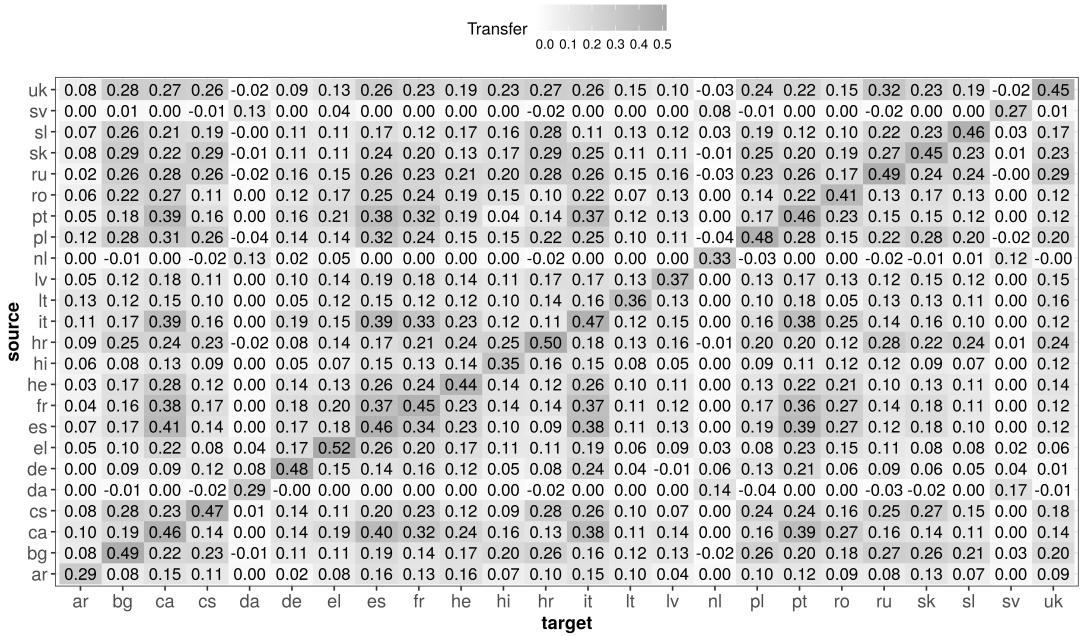

Figure 1: Average difference between accuracy and the random baseline. A positive value represents an accuracy above the baseline while a negative value indicates an accuracy below the baseline.

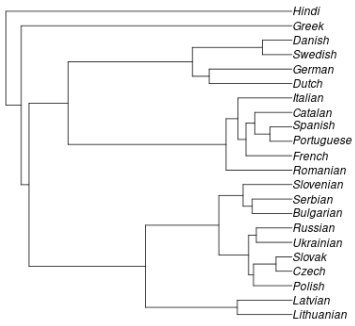

Figure 2: The phylogenetic tree of the Indo-European languages included in the analysis.

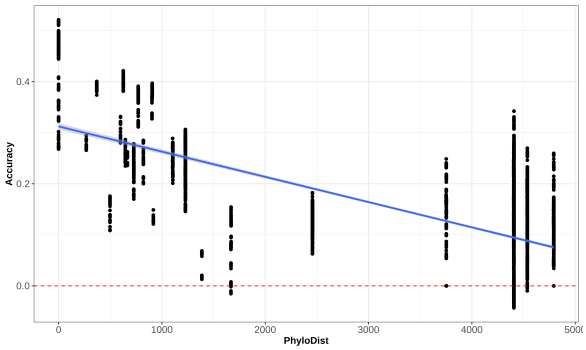

Figure 3: A comparison of the accuracy improvement (or degradation) of gender transfer (Y-axis) and the phylogenetic distance between each language pair (X-axis). The dashed line refers to the random baseline. The phylogenetic distance refers to the years separating each pair of languages in the Indo-European tree. Each point represents the average of the transfer accuracy over 10 runs.

masculine/feminine/neuter, common/neuter), we fit a linear regression model with the values of the accuracy improvement (or degradation) from the baseline as the dependent variable, phylogenetic distance a continuous predictor, and system type a categorical predictor (common/neuter is the reference level). The two-way interaction between the predictors is also included. The summary of the model is presented in Table 2. The R^2 value shows that 74% of the variation of accuracy in the sample can be explained by phylogenetic distance and gender system types.

The results again show a negative relationship between accuracy and phylogenetic distance. With regard to gender systems, having masculine/feminine or masculine/feminine/neuter has a positive effect on the accuracy, when considering common/neuter systems as the reference level. Within all three systems, masculine/feminine has the highest coefficient (0.12), which implies that transfers between languages having masculine/feminine gender systems generally result in higher accuracy than masculine/feminine/neuter and common/neuter. The interactions show that the negative effect of phylogenetic distance on accuracy is attenuated if both languages in the pair have masculine/feminine or

Predictor	Estimate	t(1894)	P value
PhyDis	-0.00009	-10.926	< 0.001
m/f	0.12251	12.053	< 0.001
m/f/n	0.06577	6.589	< 0.001
PhyDis:m/f	0.00003	4.104	< 0.001
PhyDis:m/f/n	0.00004	4.679	< 0.001

Table 2: Summary of the regression model: Accuracy as predicted by phylogenetic distance and gender system type (m = masculine, f = feminine, n = neuter).

masculine/feminine/neuter gender systems.

4 Gender transfer at the word level

In this section, we perform a finer-grained analysis: focus not on languages, but on individual nouns. Our main question is if there are any patterns in the distribution of errors. Is it random or are certain classes of nouns systematically more difficult to predict than others?

To obtain a single prediction for every noun from the 10 random-seed runs, we pick the gender which has the largest sum of confidence scores (softmax activation values). This is virtually equivalent to taking the gender that gets most votes across the runs, but has an advantage of avoiding ties.

We test whether the following factors play a role: how frequent a noun is, whether it is animate or not and whether its form is equivalent to lemma (citation form, baseform) or not.

It is reasonable to expect that embeddings of frequent nouns will capture more useful information and thus yield better accuracy. Note, however, that very infrequent nouns (frequency <5) have already been excluded from consideration, since for them the embeddings are not available. We calculate frequency of every noun form using the UD corpora.

Nouns denoting living beings, especially human beings, can be expected to yield higher accuracy, since for them the semantic motivation behind gender assignment is often more transparent (based on biological sex). That is not always the case (cf. the already-mentioned German *Mädchen* 'girl', which is neuter), and the proportion of nouns where grammatical gender is predicted by biological sex is likely to vary across languages. Furthermore, it is unknown to what extent the embeddings can actually capture the relevant semantics. Nonetheless, at least in some cases sex can predict gender (cf. French *garçon* 'boy' and *fille* 'girl', or Russian *kot* 'tomcat' and *koška* 'female cat' that are resp. masculine and feminine). For nouns that do not denote living things no such predictor is known.

As a proxy for "denoting a living thing" we use the animacy category available in some UD treebanks for Slavic languages (Czech, Slovak, Polish, Russian, Ukrainian, Slovenian and Croatian). In Slavic, animacy is manifested on the grammatical level, primarily through differential object marking (Janda, forthcoming). Animacy annotation is also available in the Hindi-PUD treebank, but that treebank lacks lemmas, which makes it unsuitable for our analysis; see below.

We extract animacy information in the following way: we go through all treebanks available for every language and calculate how often a form is annotated as "animate" or "inanimate". Polish has more detailed annotation: animate human and animate non-human, we collapse these two categories into "animate". Note that Slavic animacy is a formal feature that is not exactly isomorphic to living vs. non-living distinction.

Finally, at least in some languages it is easier to infer gender from the citation form (that is, the form which is equivalent to lemma) than from inflected forms (see (Berdicevskis, forthcoming) for an overview of Slavic languages). This presumably happens because cues are more transparent in the citation form. It can also be easier to learn the cues for the citation form, which is often the most frequent one. We test whether this tendency is observed in our data. Alternatively, we could have tested whether certain morphological features (number: singular vs plural, case: nominative vs oblique etc.) play a role, but the analysis we choose is more universal: it can be applied to any language without adjustments.

Other formal, semantic and historical properties can potentially affect how easy it is to infer the gender of a noun (e.g. whether a noun is a recent borrowing or not), but there is no straightforward way to reliably extract this information for all the nouns in our datasets.

We run two logistic regression models. Both include data only from those language pairs where the target language is one of the seven Slavic languages with detailed animacy information: Czech, Slovak, Polish, Russian, Ukrainian, Slovenian, Croatian. In both models, the dependent variable is whether a noun has its gender correctly predicted by the classifier. In Model 1, the independent variables are frequency, animacy, citation form and

phylogenetic distance between source and target languages. It is applied to all language pairs except those where source language is Arabic or Hebrew (since for them the distance cannot be estimated). To the remaining pairs we apply Model 2, which has only three independent variables: frequency, animacy, citation form. The results are reported in Table 3 and Table 4.

Predictor	Estimate	z	P value
Freq	-0.004	-0.95	0.340
Inan	-0.999	-13.4	< 0.001
Non-cit	-0.789	-10.4	< 0.001
PhyDis	-0.0004	-24.4	< 0.001
Freq:Inan	0.006	1.4	0.168
Freq:Non-cit	0.009	2.0	0.042
Inan:Non-cit	-0.311	-3.9	< 0.001
Freq:PhyDis	1e-06	1.0	< 0.302
Inan:PhyDis	5.7e-05	3.0	< 0.003
Non-cit:PhyDis	6.9e-05	3.6	< 0.001
Freq:Inan:Non-cit	-0.009	-2.1	0.034
Freq:Inan:PhyDis	-1.7e-06	-1.6	0.106
Freq:Non-cit:PhyDis	-2.1e-06	-1.9	0.059
Inan:Non-cit:PhyDis	6.5e-05	3.1	0.002
Freq:Inan:Non-cit:PhyDis	2.4e-06	2.2	0.030

Table 3: Summary of Model 1: Correctness of the guess as predicted by noun frequency, animacy (Anim vs Inan), citation form (Cit vs Non-Cit) and phylogenetic distance (PhyDis).

Coefficient	Estimate	z	P value
Freq	-0.003	-0.4	0.707
Inan	-1.871	-14.0	< 0.001
Non-cit	-1.455	-10.7	< 0.001
Freq:Inan	0.002	0.3	0.800
Freq:Non-cit	0.002	0.3	0.7744
Inan:Non-cit	0.852	6.0	< 0.001
Freq:Inan:Non-cit	-0.001	-0.1	0.897

Table 4: Summary of Model 2 (Arabic and Hebrew as source languages): Correctness of the guess as predicted by noun frequency, animacy (Anim vs Inan) and citation form (Cit vs Non-Cit).

Model 1 (Indo-European languages only) shows that the prediction accuracy is significantly lower for inanimate nouns than for animate nouns and for inflected forms than for citation forms. The following predictors also have significant negative effects: the interaction of inanimacy and non-citation form; the interaction of frequency, inanimacy and non-citation form; phylogenetic distance between source and target languages (the coefficient is small, but it shows change per year, and the distance in our dataset vary from approx. 300 to 5000 years). Interestingly, all other significant predictors (most of which all are interactions of distance with other predictors) are positive. It means that the negative effects (for instance, those of inanimacy and non-citation form) described above are smaller for less related languages. The negative effect of inflected forms is also smaller for more frequent nouns.

Model 2 (Afro-Asiatic languages as source) shows similar results: inanimacy and non-citation forms have significant and strong negative effect. This similarity with Model 1 provides further evidence in favor of the universal factors in gender assignment. What is different, however, is that the interaction of these two factors has a strong positive effect (that is, inflected forms of inanimate nouns have higher accuracy than can be expected).

Finally, we perform one more test. Our results indicate that gender systems are partly transferable across non-related languages (see Section 3.1), which suggests there are certain universalities in gender assignment. We want to test whether these universalities are limited to the aforementioned fact that grammatical gender for living creatures closely (even though not perfectly) matches biological sex. To investigate that, we focus on those language pairs where source language is Afro-Asiatic and target language is one of those for which the animacy information is available. From these pairs, we exclude those treebanks where animacy is annotated only for a small proportion of nouns (Slovenian and Croatian), and that leaves us with 10 pairs (Arabic and Hebrew as source, Russian, Czech, Polish, Slovak and Ukrainian as target). We focus on inanimate nouns only (thus eliminating any possible contribution of biological sex) and test whether the classifier still performs above the chance baseline. With only 10 datapoints, we cannot run a reliable t-test. Instead, we perform 10 simulation tests, running a naive classifier that guesses the gender relying solely on the source-language probabilities 10000 times for every language pair and taking the proportion of cases when it achieves the same accuracy as our classifier (or higher) as the p-value. The p-value is 0 in all pairs apart from four: Arabic $\rightarrow$ Czech: 0.009, Arabic $\rightarrow$ Polish: 0.718, Arabic $\rightarrow$ Slovak: 0.923, Hebrew $\rightarrow$ Polish: 0.003. In other words, in eight cases out of 10, the classifier, applied only to inanimate nouns, still performs significantly better than chance.

5 Conclusion

This study investigates how grammatical gender is transferable across languages from a transfer learning point of view. The cross-lingual word embeddings are considered as the source of knowledge shared between languages from which the grammatical gender of nouns are predicted using a multilayer perceptron. The empirical results reveals that there exist some universal and lineage-specific patterns in the grammatical gender assignment.

First, our analysis of gender transfer between Afro-Asiatic and Indo-European languages indicated that partly successful gender transfer is possible between non-related languages. This observation supports the existence of universal factors in gender assignment. The accuracy of the classifier is higher than the random baseline even when it is tested on inanimate nouns only, which means that the universal factors are not limited to biological sex.

Second, our analysis of gender transfer between Indo-European languages demonstrates that the phylogenetic distance between languages has a negative effect on the success of the transfer, which suggests that some factors of gender assignment are not universal. These results match with the literature by showing that gender assignment is a mixture of universal and idiosyncratic factors.

Third, we also found that gender transfer does not work in the same way for all nouns. The prediction accuracy is significantly lower for inanimate nouns than for animate nouns and for inflected forms than for citation forms. This effect is found when considering both family-internal and family-external transfers, which provides further evidence in favor of the universal factors in gender assignment.

We would like to make a few caveats and suggestions about the future development of the current study. While we address the universality of gender assignment cross-linguistically, our data is restricted to languages from two families and our word embeddings are trained on data from specific domains. Additional data from a more diverse sample is needed to further confirm our observations. Furthermore, we cannot fully exclude that the observed similarities are an areal effect caused by contact.

It should also be noted that we cannot identify which universal factors enable the classifier to perform above the baseline. A more fine-grained word-level analysis would be required to find the possible contributors to this. Linguistically, grammatical gender is strongly tied to the semantic and formal properties of nouns. Since the cross-lingual word embeddings used in this study encode both the formal and semantic information, we cannot disentangle the relative contributions of form and semantics to gender transfer.

Finally, it should be mentioned that an important line of research in modern NLP focuses on gender bias present in naturally occurring texts (Caliskan et al., 2017; Gonen et al., 2019). The combination of these questions and approaches with our perspective might become an interesting research direction.

Supplementary materials, including raw data and scripts for analysis are openly available.[5]

Acknowledgments

The authors are thankful for the constructive comments from the anonymous referees and editors, which helped to significantly improve the quality of the paper. The second author is thankful for the support of the IDEXLYON (16-IDEX-0005) Fellowship grant.

References

Jacob Andreas and Dan Klein. 2014. How much do word embeddings encode about syntax? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 822–827, Baltimore, Maryland. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 282–291, Brussels, Belgium. Association for Computational Linguistics.

Jenny Audring. 2016. Gender. In Mark Aronoff, editor, *Oxford research encyclopedia of linguistics*. Oxford University Press, Oxford.

[5] https://github.com/marctang/Cross-lingual-embeddings-Grammatical-gender

Ali Basirat, Marc Allassonnière-Tang, and Aleksandrs Berdicevskis. in press. An empirical study on the contribution of formal and semantic features to the grammatical gender of nouns. *Linguistics Vanguard*.

Ali Basirat and Marc Tang. 2019. Linguistic information in word embeddings. In *Agents and Artificial Intelligence*, pages 492–513, Cham. Springer International Publishing.

Nicoleta Bateman and Maria Polinsky. 2010. Romanian as a two-gender language. In *Hypothesis A/Hypothesis B*, pages 41–77. MIT Press.

Aleksandrs Berdicevskis. forthcoming. Gender and declension. In Neil Bermel and Jan Fellerer, editors, *Oxford Guides to the World's Languages: The Slavonic Languages*. Oxford University Press.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Will Chang, Chundra Cathcart, David Hall, and Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1):194–244.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data.

G. G. Corbett. 2001. Grammatical gender. In *International Encyclopedia of the Social Sciences*, pages 6335–6340.

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

Greville G Corbett. 2013a. 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.

Greville G Corbett. 2013b. Sex-based and non-sex-based gender systems. In Matthew S Dryer and Martin Haspelmath, editors, *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Greville G Corbett and Norman Fraser. 2000. Gender assignment: A typology and a model. In Gunter Senft, editor, *Systems of nominal classification*, pages 293–325. Cambridge University Press, Cambridge.

Francesca Di Garbo, Bruno Olsson, and Bernhard Wälchli. 2019. Grammatical gender and linguistic complexity : Volume ii: World-wide comparative studies.

Hans-Olav Enger. 2017. The Nordic languages in the 19th century II: Morphology. In Oskar Bandle, Kurt Braunmüller, Ernst Hakon Jahr, Allan Karker, Hans-Peter Naumann, Ulf Teleman, Lennart Elmevik, and Gun Widmark, editors, *The Nordic Languages, Part 2*, pages 1437–1442. De Gruyter, Berlin.

Sebastian Fedden and Greville G Corbett. 2019. The continuing challenge of the German gender system. *Paper presented at the International Symposium of Morphology*.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 463–471, Hong Kong, China. Association for Computational Linguistics.

Laura Janda. forthcoming. Gender and animacy. In Neil Bermel and Jan Fellerer, editors, *Oxford Guides to the World's Languages: The Slavonic Languages*. Oxford University Press.

Scott Jarvis and Aneta Pavlenko. 2010. *Crosslinguistic influence in language and cognition*, paperback ed edition. Routledge, New York, NY. OCLC: 845735473.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

David Kemmerer. 2017. Categories of object concepts across languages and brains: the relevance of nominal classification systems to cognitive neuroscience. *Language, Cognition and Neuroscience*, 32(4):401–424.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

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, Singapore. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang,

Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Curt Rice. 2006. Optimizing gender. *Lingua*, 116(9):1394–1417.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Laura Sabourin, Laurie A Stowe, and Ger J De Haan. 2006. Transfer effects in learning a second language grammatical gender system. *Second Language Research*, 22(1):1–29.

Frank Seifart. 2010. Nominal classification. *Language and Linguistics Compass*, 4(8):719–736.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 69–76, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, et al. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.