An empirical study on the contribution of formal and semantic features to the grammatical gender of nouns

Ali Basirat, Marc Allassonnière-Tang, Aleksandrs Berdicevskis

Ali Basirat, Marc Allassonnière-Tang* and Aleksandrs Berdicevskis

An empirical study on the contribution of formal and semantic features to the grammatical gender of nouns

https://doi.org/10.1515/lingvan-2020-0048
Received December 10, 2019; accepted December 8, 2020; published online March 17, 2021

Abstract: This study conducts an experimental evaluation of two hypotheses about the contributions of formal and semantic features to the grammatical gender assignment of nouns. One of the hypotheses (Corbett and Fraser 2000) claims that semantic features dominate formal ones. The other hypothesis, formulated within the optimal gender assignment theory (Rice 2006), states that form and semantics contribute equally. Both hypotheses claim that the combination of formal and semantic features yields the most accurate gender identification. In this paper, we operationalize and test these hypotheses by trying to predict grammatical gender using only character-based embeddings (that capture only formal features), only context-based embeddings (that capture only semantic features) and the combination of both. We performed the experiment using data from three languages with different gender systems (French, German and Russian). Formal features are a significantly better predictor of gender than semantic ones, and the difference in prediction accuracy is very large. Overall, formal features are also significantly better than the combination of form and semantics, but the difference is very small and the results for this comparison are not entirely consistent across languages.

Keywords: formal features; gender; neural networks; semantics; word embeddings

1 Introduction

Grammatical gender (Corbett 2013a, 2013b) is a type of nominal classification system found in many natural languages of the world. The functions of such systems have been documented in the linguistic literature as referent identification and tracking (Allassonnière-Tang and Kilarski 2020; Contini-Morava and Kilarski 2013; Dye et al. 2017). However, it is not clear which principles govern the assignment of nouns to different gender categories in different languages (Comrie 1999; Kemmerer 2017). For example, in French, *table* is feminine and *book* is masculine. The reasons for such gender assignment are much more opaque than the functions of the gender systems.

Several theories have been put forward to explain which linguistic features affect the association between nouns and grammatical genders (Fedden and Corbett 2019). This work is limited to testing two influential theories of this kind that make contradicting claims. Corbett and Fraser (2000, p. 318) suggest that gender assignment is primarily affected by the semantic features of nouns rather than formal features. By semantic features they mean certain aspects of the noun's meaning, by formal features they mean the phonological form of the noun (all possible forms, if it is inflected). Rice (2006), however, argues in the *optimal gender assignment theory* that form and semantics contribute equally to the process of gender

***Corresponding author: Marc Allassonnière-Tang**, Lab Dynamics of Language CNRS UMR 5596, University Lyon 2, Lyon, France, E-mail: marc.tang@univ-lyon2.fr
Ali Basirat, Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden; Department of Computer and Information Science, Linköping University, Linköping, Sweden, E-mail: ali.basirat@liu.se
Aleksandrs Berdicevskis, Språkbanken Text, Department of Swedish, University of Gothenburg, Gothenburg, Sweden, E-mail: aleksandrs.berdicevskis@gu.se

assignment. Qualitative and small-scale quantitative evidence have been provided for both claims, e.g., Nesset (2006) and Kramer (2020) for the semantic-dominance hypothesis and Corteen (2019) for the equality hypothesis. Both hypotheses have different implications in linguistic and language acquisition observations, which show that learners tend to rely more on formal features rather semantic features (Culbertson et al. 2017, 2019). Some other studies investigate the effect of semantics (Williams et al. 2019) or form (Bonami et al. 2019; Nastase and Popescu 2009) individually. While these studies provide a valuable contribution to the discussion, neither of these studies provides extensive empirical evidence on the contributions of form and semantics to gender assignment, nor do they provide a formal metric to quantify the contribution of different features. Moreover, it is not clear what the claims - that one type of features is dominant or that two are equally important - actually mean.

In this paper, a systematic approach to assess the contribution of the formal and semantic features to gender assignment is proposed. To do that, we have to operationalize the theoretical claims about the dominance of certain types of features. We do this in the following way: how much a certain feature type affects gender is measured by how accurately gender can be predicted from the given feature type. Note that all our conclusions will apply to this operationalization of the two hypotheses we are discussing. Other operationalizations of "domination" are possible (for instance, that if there is a conflict of semantic and formal factors, semantics wins).

When considering formal features, this study is limited to the orthographical form of the nouns, ignoring the possible divergences between phonology and orthography. To capture the information about the forms and the meanings of nouns, we use distributional methods, a powerful, but (so far) underused test bed for linguistics hypotheses (Boleda 2020, p.321). We use character-based word embeddings (Bojanowski et al. 2017; Chen et al. 2015) to encode forms and context-based word embeddings (Mikolov et al. 2013; Pennington et al. 2014) to encode meanings. A classifier is trained on the embeddings with the task of predicting the gender of a given noun (Basirat and Tang 2018). The performance of the classifier trained on the respective embedding type serves as the measure of the contribution of this feature type to gender predictions.

The structure of the paper is as follows: Section 2 summarizes the existing studies and hypotheses on gender assignment. Section 3 provides an overview of the word embedding methods used in this research. It also outlines the architecture of the classifier[1] used to predict the grammatical gender of nouns. The experimental settings and results are explained in Sections 4 and 5.

2 Defining grammatical gender

Human beings constantly need to categorize the experiences that they encounter to facilitate storage and access to the information in the brain (Senft 2000; Lakoff and Johnson 2003, p. 162–163). Various mechanisms of nominal classification in language may be a manifestation of this need. One of the most prominent of these systems is grammatical gender, which is also referred to as noun classes in the literature (Corbett 2013a, 2013b). For instance, nouns in French are commonly assigned to either masculine (e.g., book) or feminine (e.g., table) classes, which triggers grammatical agreement on various elements in clauses, c.f. *un grand livre* (one.MASC big.MASC book) vs. *une grande table* (one.FEM big.FEM table). Figure 1 displays a simplified view of languages with gender. The data is a sample of 257 languages used to represent the distribution of gender languages worldwide. Within these 257 languages, 43.6% (112/257) have a gender system.

1 It is important to clarify that the term 'classifier' is used in the sense of a *computational classifier* in the current paper. In other words, it refers to a computational model that is used to classify tokens of a data into specific categories. We do not use the term 'classifier' to refer to a *numeral classifier*, which would be a morpheme or a word used to classify referents of nouns in languages such as Mandarin Chinese.

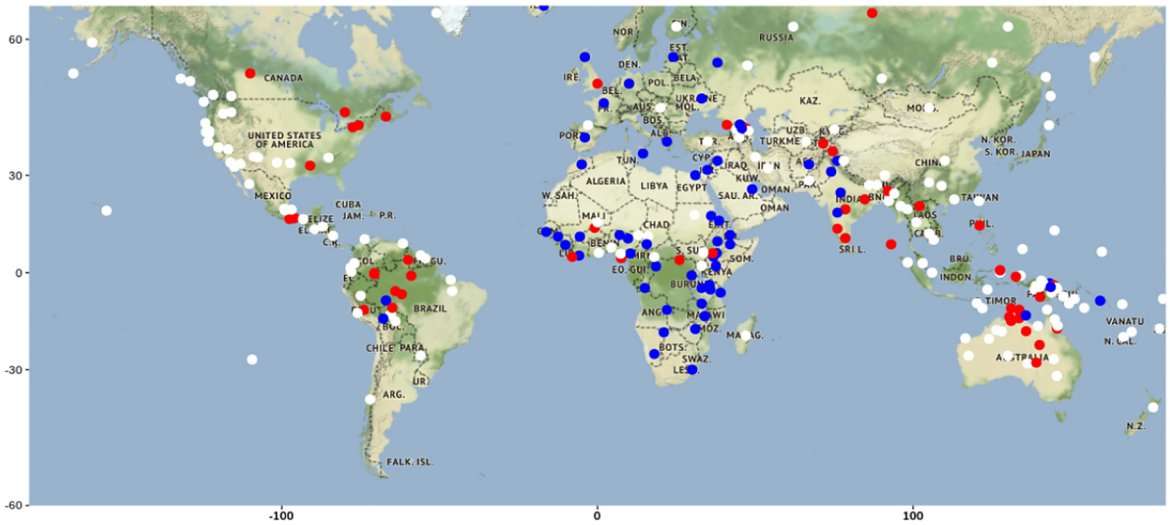

Figure 1: A representative sample of gender languages in the world. Each point represents a language. White represents languages without gender. Red indicates languages with a primarily semantic gender assignment system. Blue refers to languages that rely on formal and semantic features for gender assignment (Corbett 2013a, 2013b).

The assignment of nouns to gender is commonly considered as an interaction between the formal and the semantic features of nouns (Comrie 1999; Corbett 1991; Fedden and Corbett 2019). For instance, German divides nouns into masculine, feminine and neuter, whereas French categorizes nouns into either masculine or feminine. In both languages, nouns ending in *-e* (i.e., final schwa) tend to be feminine, c.f., *Schule* and *école* 'school'. Russian has three genders: masculine, feminine, and neuter. The assignment of nouns to these three genders is mostly predictable based on semantic information (animacy, biological sex), phonological and declensional features (Corbett 1982).

An influential line of research suggests that the contribution of semantic features generally outranks the contribution of formal ones (Corbett and Fraser 2000), but both types of factors are required for the most accurate identification of gender. Another approach (that uses the Optimality theory (Prince and Smolensky 1993) to evaluate the hierarchical interaction of form and semantics in gender assignment) makes diverging claims (Rice 2006). Evidence from languages such as German, Russian, French, Norwegian, and Dutch suggests that information about form and semantics (defined as constraints in the framework of the optimal gender assignment theory) operate together as a block and cannot be ranked. Instead of ruling out potential candidates for gender assignment by running through the constraints (i.e., features defined over semantics and word forms) one at a time, all the constraints about semantics and word forms are considered together. Each candidate is marked with the number of constraints violated. The candidate with the fewest violations is then selected as the optimal candidate. The contributions of each semantic and formal constraints are thus equal, and both are necessary for the correct identification of gender.

To sum up, the semantic-dominance hypothesis (Corbett and Fraser 2000), as operationalized by us, predicts the following: that the gender classification accuracy obtained from semantic features should be higher than the accuracy obtained from the formal features, but lower than the accuracy obtained from the combination of formal and semantic features. The equality hypothesis (Rice 2006) predicts the following: the accuracy obtained from the combination of formal and semantic features should be higher than the results obtained from each source individually; the individual accuracies should be equal. The studies addressing the two hypotheses typically rely on qualitative analyses. In some cases, a small sample of nouns is selected in different languages and the rules that govern their gender assignment are described. In this study, large-scale experiments are conducted to assess how well formal and semantic features can predict the gender of nouns.

3 Word embeddings for gender classification

Word embeddings are real-valued vectors associated with words in such a way that word similarities are reflected through vector similarities, i.e., similar words are associated with similar vectors (Boleda 2020). Word embeddings can be trained based on different aspects of words. In the following, we elaborate on two types of word embedding methods for encoding the formal and semantic features of word. Then, we propose a model to combine the two types of features.

3.1 Character-based embeddings

A character-based word embedding method (Bojanowski et al. 2017; Cao and Rei 2016; Chen et al. 2015;) captures information about the lemma (which is equivalent to "formal features" in the narrow definition of the current study). These methods map sequences of characters to vectors with respect to the similarities between the character sequences.

The character-based word embeddings are trained as part of a feed-forward neural network that takes a binary representation of characters forming a word as input and predicts the grammatical gender of the word on its output. Similar to previous studies (Yu et al. 2017), each character is associated with a one-hot vector, a binary vector whose elements correspond to all characters forming a language with all elements equal to zero except for the one that corresponds to the target character. In the current study, each noun is represented by the concatenation of one-hot character vectors associated with all characters of a word.

The one-hot character vectors are fed into an embedding layer. The character embeddings are then fed to a two-layer perceptron followed by a softmax layer whose outputs correspond to the grammatical gender of the word in the input layer. The dropout regularization (Hinton et al. 2012) is used for each layer to prevent overfitting to the training data.

3.2 Context-based embeddings

The context-based word embeddings (Mikolov et al. 2013; Pennington et al. 2014) trained on large unannotated corpora can capture semantic information about words (Schütze 1992; Sahlgren 2006) such as gender-related information (Gonen et al. 2019; Williams et al. 2019). In these methods, words are embedded into a vector space such that words with similar meanings are clustered together. Hence, content-based embeddings contain information about the semantics of words (Lebret and Collobert 2015). Among the commonly used tools to train context-based word embeddings are word2vec (Mikolov et al. 2013), and GloVe (Pennington et al. 2014).

We use a two-layer perceptron with a softmax layer, similar to what is used for the character-based word embeddings, to classify the context-based word embeddings with respect to their grammatical gender. The embeddings are fed into the classifier as input and their grammatical gender is predicted in the output of the softmax layer.

3.3 Combined embeddings

To study the contribution of both formal and semantic features on grammatical genders, we concatenate character-based and context-based embeddings and feed them to a multi-layer perceptron. Figure 2 shows the architecture of the classifier. For each word, it generates a character-based embedding on the output of the character-embedding layer and retrieves a pre-trained context-based word embedding. These embeddings are concatenated and fed into a two-layer perceptron whose output is a softmax layer that predicts the grammatical gender of the input word.

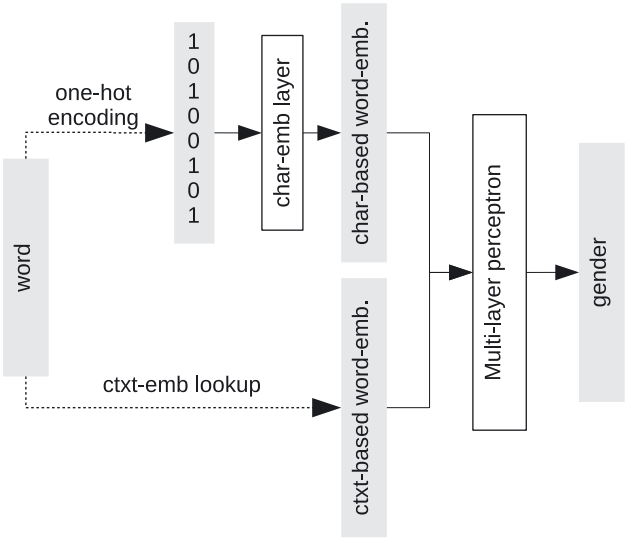

Figure 2: The architecture of the neural classifier used to predict the grammatical gender of a noun with character-based and context-based word embeddings.

The character-based embeddings are trained from scratch, and the pre-trained context-based embeddings are updated while training the entire network. The perceptron directly models all the interactions between the elements of the character-based and context-based embeddings. The weight vectors of the first layer of the perceptron can be considered as a dense representation that encodes both the formal and semantic information meaningful to the prediction of grammatical gender.

4 Experimental settings

This section provides an overview of the data sources and parameters used during the series of experiments. To facilitate reading, several additional experiments are reported in the appendices instead of the main text of the paper.

4.1 Languages selected for the experiments

The experiments are performed on a sample of three languages: French, German and Russian. As shown in Figure 1, the languages with gender are pretty much spread throughout the world. The gender systems found in South America, Oceania, and Africa are usually considered to belong to significantly different types than grammatical gender found in Indo-European languages (Corbett 2013a, 2013b; Grinevald 2015). This study is restricted to grammatical gender found in the latter type.

- French (Romance) has two genders: masculine and feminine, cf. *un grand livre* (one.MASC big.MASC book) ‘a big book’ vs. *une grande table* (one.FEM big.FEM table) ‘a big table’.
- German (West Germanic) has three genders: masculine, feminine, and neuter, cf. *der Hund* (the.MASC dog) ‘the dog’, *die Katze* (the.FEM cat) ‘the cat’, and *das Vögelein* (the.NEUT bird) ‘the bird’.
- Russian (Slavic) has three genders: masculine, feminine, and neuter, cf. *Èt-a spel-aja gruša* (this-FEM ripe-FEM pear) ‘this ripe pear’, *Èt-ot spel-yj persik* (this-MASC ripe-MASC peach) ‘this ripe peach’, and *Èt-o spel-oje jabloko* (this-NEUT ripe-NEUT apple) ‘this ripe apple’.

The current sample is too small and skewed for any quantitative typological comparison. Nonetheless, it contains languages with different configurations of gender systems.

4.2 Character-based embeddings

The character-based word embeddings are trained with all characters of each noun. We consider a maximum word length of 100 characters for all words. The characters of a word are placed into an array of 100 elements and the unused places are padded with 0. Each place is then mapped to the one-hot representation of the corresponding character and the zero elements are maps to a vector of zeros. The one-hot character vectors and the zero vectors are then concatenated in order to form the character representation of the word. In this representation, the order of characters is taken into account. We also tested if including fewer characters of each word (e.g., only counting the first three and the last three characters) has an effect on the performance of the model. Further details are found in Appendix A.

The gender and characters of nouns are extracted from corpora-based word lists and only include lemmas. The French data is collected from the Dicollecte lexical database[2] with 48,149 nouns. The DEMorphy database (Duygu 2018) with 988,040 nouns is used for German. The Russian data has 13,851 nouns extracted from the Syntagrus corpus.[3] The German word list is inflated since it includes an extensive set of compounds. However, this does not impact the results.

4.3 Context-based embeddings

The context-based word embeddings are trained by GloVe (Pennington et al. 2014). Word embeddings are generally sensitive to the size and the type of the context (Melamud et al. 2016). The model is trained with a symmetric bag-of-words context, which is expected to capture the semantic information about words (Lebret and Collobert 2015), and the context size is set to $l = 5$. Results on other context sizes and embedding methods are reported in Appendix B, in which we show that the optimal size of the context does not vary much between languages and methods.

To be consistent with the character-based word embeddings, we also set the dimensions of context-based word embeddings to 50. Other options of the embedding tool are set to their default values. The Wikipedia part of the raw data provided for the CoNLL 2017 shared task (Ginter et al. 2017) is used to train embeddings. All words are lemmatized and turned to their lowercase form. The embeddings are generated for words with the frequency higher than 50. This results in a vocabulary of size 243K for French, 479K for German, and 248K for Russian. The detailed statistics about the lemmatized corpora are shown in Table 1.

The choice of lemmatized corpora is motivated by the possibility that the embeddings based on raw corpora learn not only the semantics of the nouns, but also formal co-occurrences with modifiers of the certain gender, which is expected from languages with rich agreement both within the noun phrase and the verb phrase (Kann 2019). It is possible that the estimate of the contribution of the semantic features is somewhat diminished in embeddings based on lemmatized corpora. Nevertheless, the results based on raw corpora also

Table 1: The total number of types, tokens, and sentences in each of the lemmatized corpora; K:10^3, and M:10^6.

	Type	Token	Sentence
French	8,593K	787M	65M
German	16,859K	1070M	105M
Russian	9,046K	554M	54M

2 grammalecte.net

3 http://ruscorpora.ru/search-syntax.html.

match with the results from lemmatized corpora reported in the current paper (see Appendix C for further details).

4.4 Shared settings

An overview of the settings used in the paper is shown in Table 2. Both character-based and context-based word embeddings are trained on 50 dimensions and lemmatized word forms. The character-based word embeddings are fed with all characters in each noun. The context-based word embeddings are trained with the five preceding and five following words of each noun. After being evaluated individually in Study 1 and 2, both information sources are merged in Study 3.

To avoid the overfitting issue, each data set is divided into three parts (80, 10, and 10%) with no overlap between them to train, validate, and test the classifiers for each language. The classifiers are first trained on 80% on the data. Then, we use the development set of 10% to investigate the effect of parameters on the performance for the classifiers. That is to say, the results of the experiments reported in the appendices are assessed with the validation set. The final accuracy of the classifiers is obtained by asking the classifiers to predict the gender of the nouns in the test set of 10%. During this process, the test data is used only once to assess the two linguistic theories on grammatical gender.

5 Results

In this section, we report the results on the test sets and assess whether they support the semantic-dominance hypothesis or the equality hypothesis. The results are based on the settings reported reported in Table 2. Details on additional experiments such as the effect of lemmatized vs. inflected corpora and different embedding methods are available in Appendix C.

When assessing the final results, we also considered the fact that neural networks trained with the back-propagation algorithm are sensitive to their initial state (LeCun et al. 2012). That is to say, the initial weights of a neural network can have a significant effect on its performance and the results may vary across different experiments. To cover this possible variation of results, the experiments were run 100 times with different random seed values, and the mean and the standard deviation of the network's accuracy on the test set are reported in Table 3. The relatively small values of the standard deviation indicate that the neural network's performance is not very sensitive to the initial values.

Table 2: The different combinations of information sources tested in the experiments. We predict the gender of the *input form* based on the *character-based* and *context-based* word embeddings.

Study	Character-based	Context-based	Window size	Dimensions
1	Lemma	–	All characters	50
2	–	Lemmatized corpora	Five words	50
3	Lemma	Lemmatized corpora	Both of the above	50

Table 3: The mean and the standard deviation of the accuracy of 100 times trial of the grammatical gender classification.

	Form		Sem		Form + Sem	
	μ	Σ	M	σ	μ	σ
French	0.92	0.005	0.65	0.010	0.90	0.006
German	0.88	0.002	0.48	0.005	0.87	0.003
Russian	0.96	0.003	0.54	0.014	0.97	0.005

The semantic-dominance hypothesis (Corbett and Fraser 2000) predicts that the semantic features dominate formal features in gender assignment. Based on this theory, the results in the 'Sem' column should be on average higher than the results in the 'Form' column. A two-sided paired *t*-test indicates that the opposite is found in the data, the 'Form' column is significantly higher than the results in 'Sem' ($t = -89.64$, $df = 299$, $p<0.01$). In other words, the result does not support our operationalization of the semantic-dominance hypothesis.

The equality hypothesis (Rice 2006) claims that formal and semantic features contribute equally to gender assignment and contain complementary information. Under such an assumption, no differences between the results on columns 'Form' and 'Sem' are expected. The two-sided paired *t*-test reported in the previous paragraph indicates that the accuracy of 'Form' is larger than 'Sem'. Thus, the assumption of no difference does not hold. Both hypotheses predict that the highest accuracy should be obtained by combining the formal and the semantic features. Two individual two-tailed paired *t*-tests show that, on the one hand, the accuracy of 'Form + Sem' is significantly higher than 'Sem' ($t = 79.58$, $df = 299$, $p < 0.01$), but on the other hand, the accuracy of 'Form + Sem' is not significantly higher than 'Form'. On the contrary, the accuracy of 'Form' is significantly higher than 'Sem + Form' ($t = \mathbf{-9.93}$, $df = 299$, $p < 0.01$). In other words, combining the factors boosts the accuracy if the results are compared against semantics only, but not if compared against formal factors only, and the reason is not that the two information sources are complementary, but that formal factors are a better predictor. To sum up, none of the predictions made by the two hypotheses (in our operationalization) are supported by our data.

6 Conclusion

The experimental results show that both formal and semantic features are informative about the gender of nouns. The two hypotheses about gender assignment are assessed individually. First, semantic features do not yield higher accuracy than formal ones, which goes against the semantic-dominance hypothesis (Corbett and Fraser 2000). On the contrary, they yield significantly lower accuracy, which goes against the equality hypothesis as well, since the contributions of forms and semantics are not equal. Second, combining lemmas and semantics does not result in significantly higher accuracy. This observation supports neither the equality nor the semantic-dominance hypothesis (Rice 2006). The information about the grammatical gender of nouns is encoded in different ways across formal and semantic features, but the combination of the two sources of information does not yield a significant improvement over using formal features only.

Finally, we would like to make a few caveats and suggestions about the future development of the current study. First, the analysis attempts to evaluate only the two specific hypotheses (strictly speaking, our operationalizations of these hypotheses), not the whole gender assignment theories proposed by Corbett and Fraser, on the one hand, and Rice, on the other hand. The claims should not be extended to any parts of these theories beyond the two hypotheses. Second, the experiments are limited to only three languages, which belong to the same family and the same area. The small size of the sample means that the statistical analysis has low power, and the bias means that their results should be interpreted with caution. Third, the experiments are based on two static types of words embeddings that represent words as a single point in a vector space with fixed features. However, in practice, the meaning of the words may change depending on their contextual environments. This suggests the use of recently developed contextualized word embeddings (Devlin et al. 2019; McCann et al. 2017; Peters et al. 2018). Fourth, when we say that we estimate the contribution of formal and semantic features, we assume that our methods of capturing their predictive power are adequate. While there is little doubt that character-based embeddings capture the contribution of the phonological forms adequately, we cannot be equally certain that word embeddings fully capture those components of meaning that can affect gender (e.g. biological sex). It is thus possible that our quantification underestimates the role of semantic factors. However, given the extremely high performance of formal features, it seems unlikely that this potential underestimation skews any results in any significant way. Fifth,

the semantic space could have a more complex structure than the space of phonological form relevant to gender assignment. Semantics may thus require a more complex data representation in order to exploit all the information actually available. The effect of dimensionality was also tested to assess if different settings of the model would affect its performance, as the different numbers of dimensions can be expected to capture this additional complexity. The results show that using different numbers of dimensions does not have a big influence on the conclusion of this study (see Appendix D). Increasing the dimensionality results in slightly better performance for context-based embeddings at first, but not by much, as the accuracy plateaus rather quickly. Even with extremely high dimensionality, context-based embeddings do not outperform character-based embeddings.

Acknowledgement: The authors are thankful for the constructive comments from the anonymous referees and editors, which helped to significantly improve the quality of the paper. Special thanks to Niklas Edenmyr and Joakim Nivre for providing comments on earlier versions of the paper. The authors are fully responsible for all remaining errors.
Research funding: The second author expresses his gratitude for the support of the IDEXLYON Fellowship Grant (16-IDEX-0005), University of Lyon Grant NSCO ED 476 (ANR-10-LABX-0081), and French National Research Agency (ANR-11-IDEX-0007).

Appendix

A. The effect of window size for character-based word embeddings

The following results are based on the validation sets. Three values of character window size are tested: 3, 6, and ALL based on two types of forms: inflected and lemmatized. In this context, character window size refers to the number of characters at the beginning and the end of a word with regard to their order. The value of ALL covers the entire noun. Table 4 shows the accuracy of the classifiers trained with character-based word embeddings with different values of character window size. The results from lemmas and inflected forms are shown separately. Their respective performance is compared with the majority baselines, i.e., every noun is associated with the gender that has the largest size in the data. The performance on inflected forms and lemmas are rather similar. In both cases, the accuracy of the classifiers is significantly higher than the baseline.

Different window sizes yield nearly identical results, suggesting that most of the information about the grammatical gender of nouns tends to be present at the beginning or the end of the nouns. This is expected from a linguistic perspective since most nominal markers tend to be located either at the beginning or in the end of words in the three languages. In general, these results demonstrate that the formal features are very good predictors of the grammatical gender.

Table 4: The performance of the classifier for predicting the gender of nouns (inflected forms and lemmas) based on character-based word embeddings with character windows of size 3, 6, and *all* characters. The baseline refers to the majority baseline.

	Size 3		Size 6		Size all		Baseline	
	Inflected	Lemma	Inflected	Lemma	Inflected	Lemma	Inflected	Lemma
French	90.9	90.9	92.1	92.4	92.0	92.2	54.6	56.2
German	84.3	82.7	92.1	88.7	93.1	88.2	37.2	37.0
Russian	87.7	96.1	88.9	96.3	89.1	96.6	41.8	42.9

B. The effect of method for context-based word embeddings

The information captured by context-based word embeddings might be influenced by parameters such as the embedding method, corpus type, context size, and noun form. In this section, we study the effect of these parameters on the grammatical gender prediction.

Three different types of word embedding methods are used for this study, including word2vec (skip-gram) (Mikolov et al. 2013), GloVe (Pennington et al. 2014), and principal word embedding (PWE) (Basirat 2018; Basirat and Nivre 2019). Each method is trained with the symmetric bag-of-words content with different sizes, 1–5 on both inflected and lemmatized corpora. Furthermore, two forms of nouns are also distinguished: raw (i.e., lemma) versus inflected. That is to say, we also assess the performance of the classifier for predicting the gender of lemmatized noun forms and inflected noun forms. For instance, inflected noun forms may include noun forms with singular and/or plural case markings. To avoid confusion of terms, the terms 'lemmatized' and 'raw' are used to refer to the embeddings, while the terms 'lemma' and 'inflected' are used to refer to the noun forms in this section.

word2vec offers two types of embedding models: cbow and skip-gram. We use the skip-gram model. The skip-gram model is a two-layer neural network that takes the word in the middle of a sequence of words (context) as input and predicts surrounding words within a certain range before or after the input word. Unlike word2vec which relies on the local occurrences of words, GloVe and PWE rely on both local and global contexts of words. These methods use matrix factorization techniques to train word embeddings from a word co-occurrence matrix undergone a transformation function. GloVe applies a static transformation on the co-occurrence data and uses a regression model to factorize the transformed data. However, PWE applies an adaptive transformation and uses a randomized singular value decomposition to factorize the matrix.

The results obtained from Glove, word2vec, and PWE are shown in Figure 3. In terms of noun form, the performance does not vary much between inflected noun forms and lemmatized noun forms. The biggest drop in accuracy occurs with Russian, which is not surprising due to its complex case inflection systems that differentiates the nominative, the accusative, the genitive, the dative, the instrumental, and the locative cases. This is expected since raw noun forms represent a larger and more versatile data source than lemmatized noun forms.

The performance of the raw embeddings, i.e., the embeddings based on raw/inflected corpora, is higher than the performance of embeddings based on lemmatized corpora. This is expected since the inflected embeddings may include morphosyntactic features that provide obvious clues to the classifier. Nevertheless, the inflected embeddings trained with different methods act differently on the task of gender prediction of nouns. In general, the best results are obtained from the PWE embeddings with all context size for each language. Except for Glove with Russian and larger context size. word2vec does not perform as well as Glove and PWE, especially when the context size increases. The performance of the lemmatized embeddings, i.e., the embeddings based on lemmatized corpora are much lower than the performance of the inflected embeddings.

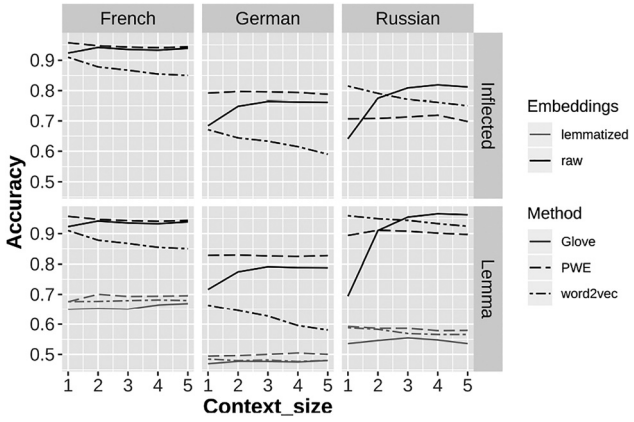

Figure 3: A cross-lingual comparison between the performance of the classifier for predicting the grammatical gender of nouns from context-based word embeddings trained with different methods and different context sizes.

Using lemmatized corpora thus seems to represent a more accurate picture of how helpful purely semantic information is in predicting the grammatical gender of nouns. Nevertheless, both types of embeddings result in accuracy higher than the baselines, which implies that semantics play an important role in the association between nouns and their grammatical gender.

Lemmatized embeddings result in a similar performance for word2vec, PWE, and Glove, regardless of the window size. However, inflected embeddings lead to different performance of the models. In general, the accuracy of PWE is stable at all window sizes with each language. However, Glove performs better with large window size while word2vec reaches a higher accuracy with small window size. Since the variation of accuracy based on different context size is not consistently found in each embedding method, it is more likely that the optimal size of the context does not vary much between languages. The small divergences in performance are more likely to occur due to the different structures of the embedding methods. In the current paper, we report the results from Glove since it is considered to be one of the most standard methods.

C. The effect of parameters and corpora on the final results

Each language has 180 classification accuracy results related to the two types of noun encoding (inflected forms and lemmas) and the two types of corpora (inflected and lemmatized) plus the three types of context-based word embeddings (word2vec, GloVe, and PWE) each trained with five values of context size and combined by character-based word embeddings trained with three values of character window size (3, 6, ALL). So in total, $180 \times 3 = 540$ results for the three languages. We found it more convenient not to represent the detailed results related to the character window size and only plot the results obtained from the character window size of six for which relatively good results were obtained for all languages.

Figure 4 summarizes the results obtained from the combination of the formal and semantic features using character-based and context-based word embeddings, respectively. As explained above, the character-based embeddings are trained with only 6 characters at the beginning and the end of words, but the context-based embeddings are trained with different values of context size along with inflected and lemmatized corpora. Context size is not an extremely important parameter on the classification performance. The optimal value of the context size is rather stable with lemmatized embeddings and fluctuates more with inflected embeddings. However, this fluctuation is more related to the embedding model rather than the data. In general, PWE and Glove result in relatively good results for all languages, with Glove requiring a larger context size for German. word2vec only performs as well as PWE and Glove for Russian.

In terms of languages, the performance on German is generally lower than the performance on French and Russian. When only using context-based embeddings, the performance on French is consistently the highest. When combining character-based embeddings and context-based embeddings, the performance is high for all three languages, especially for Russian, which almost reaches ceiling. This variation indicates that the level of

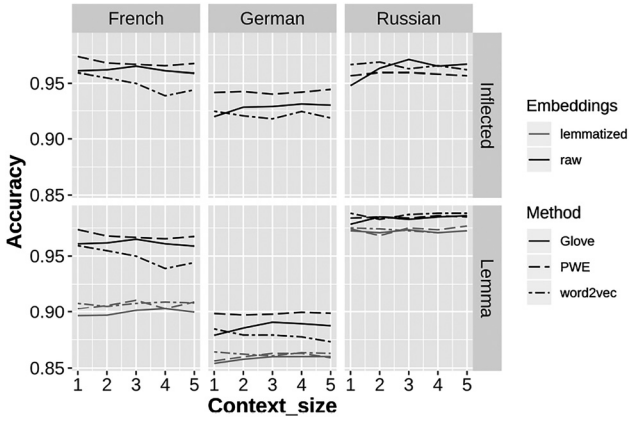

Figure 4: A cross-lingual comparison between the performance of the classifier for predicting the grammatical gender of nouns from both character-based and context-based word embeddings. The character-based word embeddings are trained with a character-window size of 6, but the context-based embeddings are trained with different context size with different embedding methods.

information encoded in form and semantics vary across languages, but it does not directly affect the current analysis, which focuses on the accuracy at the relative scale within each language.

In terms of semantics and form, due to the similarity of results between inflected forms and lemmas, only the results of the latter are shown in Table 5. The second column shows the accuracy of the classifier based on inflected embeddings and lemmatized embeddings. The third column refers to the performance of the classifier when both character-based embeddings and context-based embeddings are considered.

Formal features yield high accuracy on the classification task. In all cases, the accuracy based on character embeddings is higher than the accuracy based on lemmatized embeddings. The combination of lemmatized embeddings and character embeddings does not surpass the performance of character embeddings alone (except for Russian, but with a small magnitude). Inflected embeddings help the classifier to reach an accuracy higher than character embeddings. However, this high performance is very likely due to the morphosyntactic cues rather than the semantic information in inflected embeddings.

D. The effect of dimensionality for both types of embeddings

The number of dimensions of word embeddings has a vital role in the type and the amount of information captured (Yin and Shen 2018). A large number of dimensions may lead to the overfitting issue, and a small number may not be enough to capture the required information about words. We study the effect of dimensionality of the word embeddings on the grammatical gender classification as follows. We start with the content-based embeddings and train different sets of embeddings for each language with a different number of dimensions. Then, we fix the context-based embeddings to the one that results in the highest accuracy (i.e., 1,000-dimensional embeddings in this case) and study the effect of the dimensionality of character-based embeddings on the performance of the neural network. We report the average accuracy of 10 trial for each set of embeddings. Table 6 summarizes the average accuracy of the classifier trained with Glove word embeddings of a different number of dimensions.

Table 5: The accuracy of grammatical gender classification task based on the features defined over Lemmas, semantics (Sem.) and their combinations (Lemma + Sem.). The word embeddings representing semantics of words are trained on lemmas as noun forms with inflected and lemmatized (lem.) corpora. The results are obtained from the test sets with The PWE embedding method, context size five and character size 6.

Languages	Form	Sem.		Form + Sem.	
		Inf.	Lem.	Inf.	Lem
French	0.92	0.94	0.69	0.96	0.90
German	0.88	0.82	0.50	0.90	0.86
Russian	0.96	0.89	0.58	0.98	0.97

Table 6: The average accuracy of grammatical gender classification using word embeddings with different number of dimensions.

	Context-based							Character-based		
	50	100	500	750	1,000	1,250	1,500	50	100	500
French	0.65	0.68	0.72	0.72	0.73	0.73	0.73	0.91	0.91	0.91
German	0.48	0.49	0.51	0.51	0.52	0.52	0.52	0.88	0.88	0.89
Russian	0.54	0.57	0.59	0.59	0.60	0.60	0.60	0.97	0.97	0.97

On the left side, the results of the context-based embeddings show that the amount of gender-related information increases as the number of dimensions increases. The best results are obtained starting from the 500-dimensional word embeddings, which then reach plateau and keep the same level of accuracy. Additional experiments with higher dimensions indicate that the model starts overfitting and the accuracy drops. As an extreme example, the accuracy on Russian with 5,000 dimensions is 0.37. The results of the character-based embeddings do not show a strong relationship between the number of dimensions of the character-based embeddings and the performance of the classifier. This observation indicates that the dimensionality of the character-based embeddings is not an effective parameter for the task. These combined results show that even with higher dimensionality, context-based embeddings do not outperform character-based embeddings.

References

Allassonnière-Tang, Marc & Marcin Kilarski. 2020. Functions of gender and numeral classifiers in Nepali. *Poznan Studies in Contemporary Linguistics* 56(1). 113–168.

Basirat, Ali. 2018. *Principal word vectors*. Uppsala: Acta Universitatis Upsaliensis PhD thesis. Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-353866.

Basirat, Ali & Joakim Nivre. 2019. Real-valued syntactic word vectors. *Journal of Experimental and Theoretical Artificial Intelligence*. 32(4). 557–579.

Basirat, Ali & Marc Tang. 2018. Lexical and morpho-syntactic features in word embeddings: A case study of nouns in Swedish. *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, 2, 663–674. Setúbal: SciTePress.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5. 135–146.

Boleda, Gemma. 2020. Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics* 6(1). 213–234.

Bonami, Olivier, Matías Guzman Naranjo & Delphine Tribout. 2019. The role of morphology in gender assignment in French. *Paper presented at the International Symposium of morphology*. Paris: Laboratoire de linguistique formelle.

Cao, Kris & Marek Rei. 2016. A joint model for word embedding and word morphology. In *Proceedings of the 1st workshop on representation learning for NLP*, 18–26. Berlin, Germany: Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/W16-1603.

Chen, Xinxiong, Lei Xu, Zhiyuan Liu, Maosong Sun & Huanbo Luan. 2015. Joint learning of character and word embeddings. In *Proceedings of the 24th International Conference on artificial intelligence* (IJCAI'15), 1236–1242. Palo Alto: AAAI Press. Available at: http://dl.acm.org/citation.cfm?id=2832415.2832421.

Comrie, Bernard. 1999. Grammatical gender systems: A linguist's assessment. *Journal of Psycholinguistic Research* 28(5). 457–466.

Contini-Morava, Ellen & Marcin Kilarski. 2013. Functions of nominal classification. *Language Sciences* 40. 263–299.

Corbett, Greville. 1982. Gender in Russian: An account of gender specification and its relationship to declension. *Russian Linguistics* 6. 197–232.

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

Corbett, Greville G. 2013a. Number of genders. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Corbett, Greville G. 2013b. Systems of gender assignment. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

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

Corteen, Emma. 2019. *The assignment of grammatical gender in German: Testing optimal gender assignment theory*. Cambridge: Cambridge University PhD thesis.

Culbertson, Jennifer, Annie Gagliardi & Kenneth Smith. 2017. Competition between phonology and semantics in noun class learning. *Journal of Memory and Language* 92. 343–358.

Culbertson, Jennifer, Hanna Jarvinen, Frances Haggarty & Kenneth Smith. 2019. Children's sensitivity to phonological and semantic cues during noun class learning: Evidence for a phonological bias. *Language* 95(2). 268–293.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 4171–4186 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/N19-1423.

Duygu, Altinok. 2018. *DEMorphy: German language morphological analyzer*. Available at: arXiv.org. arXiv:1803.00902.

Dye, Melody, Petar Milin, Richard Futrell & Michael Ramscar. 2017. A functional theory of gender paradigms. In Ferenc Kiefer, James Blevins & Huba Bartos (eds.), *Perspectives on morphological organization*. Leiden: BRILL. Available at: https://brill.com/view/book/edcoll/9789004342934/B9789004342934_011.xml (accessed 29 March 2020).

Fedden, Sebastian & Greville G Corbett. 2019. The continuing challenge of the German gender system. *Paper presented at the International Symposium of morphology*. Paris: Laboratoire de linguistique formelle.

Ginter, Filip, Jan Hajič, Juhani Luotolahti, Milan Straka & Daniel Zeman. 2017. CoNLL 2017 shared task-automatically annotated raw texts and word embeddings. Available at: http://hdl.handle.net/11234/1-1989.

Gonen, Hila, Yova Kementchedjhieva & Yoav Goldberg. 2019. How does Grammatical Gender Affect Noun Representations in Gender-Marking Languages? In *Proceedings of the 2019 workshop on widening NLP*, 64–67. Florence, Italy: Association for Computational Linguistics.

Grinevald, Colette. 2015. Linguistics of classifiers. In James D. Wright (ed.), *International encyclopedia of the social and behavioral sciences*, 811–818. Oxford: Elsevier.

Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever & Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. CoRR. Available at: http://dblp.uni-trier.de/db/journals/corr/corr1207.html#abs-1207-0580.

Kann, Katharina. 2019. *Grammatical gender, neo-whorfianism, and word embeddings: A data-driven approach to linguistic relativity*. Ithaca: Cornell University.

Kemmerer, David. 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.

Kramer, Ruth. 2020. Grammatical gender: A close look at gender assignment across languages. *Annual Review of Linguistics* 6(1). 45–66.

Lakoff, George & Mark Johnson. 2003. *Metaphors we live by*. London: The University of Chicago Press.

Lebret, Rémi & Ronan Collobert. 2015. Rehabilitation of count-based models for word vector representations. In Alexander Gelbukh (ed.), *Computational linguistics and intelligent text processing*, 417–429. Cham: Springer International Publishing.

LeCun, Yann A, Léon Bottou, Genevieve B. Orr & Klaus-Robert Müller. 2012. Efficient backprop. In *Neural networks: Tricks of the trade*, 9–48. Berlin: Springer.

McCann, Bryan, James Bradbury, Caiming Xiong & Richard Socher. 2017. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (eds.), *Advances in neural information processing systems*, vol. 30, 6294–6305. New York: Curran Associates, Inc. http://papers.nips.cc/paper/7209-learned-in-translation-contextualized-word-vectors.pdf.

Melamud, Oren, David McClosky, Siddharth Patwardhan & Mohit Bansal. 2016. The Role of Context Types and Dimensionality in Learning Word Embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1030–1040. San Diego, California: Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/N16-1118.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st international conference on learning representations*, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings. Available at: http://arxiv.org/abs/1301.3781.

Nastase, Vivi & 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*, 1368–1377. Singapore: Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/D09-1142.

Nesset, Tore. 2006. Gender meets the usage-based model: Four principles of rule interaction in gender assignment. *Lingua* 116. 1369–1393.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543. Doha: Association for Computational Linguistics.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. *The 16th annual conference of the North American Chapter of the Association for Computational Linguistics*. New Orleans: Association for Computational Linguistics.

Prince, Alan & Paul Smolensky. 1993. *Optimality theory: Constraint interaction in generative grammar*. Boulder: Rutgers University and University of Colorado.

Rice, Curt. 2006. Optimizing gender. *Lingua* 116. 1394–1417.

Sahlgren, Magnus. 2006. *The word-space model*. Stockholm University PhD thesis.

Schütze, Hinrich. 1992. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE conference on supercomputing*, 787–796. IEEE Computer Society Press.

Senft, Gunter. 2000. What do we really know about nominal classification systems. In Gunter Senft (ed.), *Systems of nominal classification*, 11–49. Cambridge: Cambridge University Press.

Williams, Adina, Damian Blasi, Lawrence Wolf-Sonkin, Hanna Wallach & 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), 5734–5739. Hong Kong, China: Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/D19-1577.

Yin, Zi & Yuanyuan Shen. 2018. On the dimensionality of word embedding. *Advances in Neural Information Processing Systems*, 887–898.

Yu, Xiang, Agnieszka Falenska & Ngoc Thang Vu. 2017. A general-purpose tagger with convolutional neural networks. In *Proceedings of the first workshop on subword and character level models in NLP*, 124–129. Copenhagen, Denmark: Association for Computational Linguistics. Available at: https://www.aclweb.org/anthology/W17-4118.