Predicting and Explaining French Grammatical Gender

Saumya Yashmohini Sahai*
The Ohio State University, USA
sahai.17@osu.edu

Dravyansh Sharma*
Carnegie Mellon University, USA
dravyans@cs.cmu.edu

Abstract

Grammatical gender may be determined by semantics, orthography, phonology, or could even be arbitrary. Identifying patterns in the factors that govern noun genders can be useful for language learners, and for understanding innate linguistic sources of gender bias. Traditional manual rule-based approaches may be substituted by more accurate and scalable but harder-to-interpret computational approaches for predicting gender from typological information. In this work, we propose interpretable gender classification models for French, which obtain the best of both worlds. We present high accuracy neural approaches which are augmented by a novel global surrogate based approach for explaining predictions. We introduce *auxiliary attributes* to provide tunable explanation complexity.

1 Introduction

Grammatical gender is a categorization of nouns in certain languages which forms a basis for agreement with related words in sentences, and plays an important role in disambiguation and correct usage (Ibrahim, 2014). An estimated third of the current world population are native speakers of gendered languages, and over one-sixth are L2 speakers. Having a gender assigned to nouns can potentially affect how the speakers think about the world (Samuel et al., 2019). A systematic study of rules governing these assignments can point to the origin of and potentially help mitigate gender biases, and improve gender-based inclusivity (Sexton, 2020).

Grammatical gender (hereon referred to by gender) need not coincide with “natural gender”, which can make language acquisition more challenging. For example, Irish *cailín* (meaning "girl") is assigned a masculine gender. Works investigating the role of gender in acquiring a new language (Sabourin et al., 2006; Ellis et al., 2012) have found that the speakers of a language with grammatical gender have an advantage when acquiring a new gendered language. Automated generation of simple rules for assigning gender can be helpful for L2 learners, especially when L1 is genderless.

Tools for understanding predictions of statistical models, for example variable importance analysis of Friedman (2001), have been used even before the widespread use of black-box neural models. Recently the interest in such tools, reformulated as *explainability* in the neural context (Guidotti et al., 2018), has surged, with a corresponding development of a suite of solutions (Bach et al., 2015; Sundararajan et al., 2017; Shrikumar et al., 2017; Lundberg and Lee, 2017). These approaches typically *explain* the model prediction by attributing it to relevant bits in the input encoding. While faithful to the black box model’s “decision making”, the explanations obtained may not be readily intuited by human users. Surrogate models, which globally approximate the model predictions by a more interpretable model, or obtain prediction-specific explanations by perturbing the input in domain-specific ways, have been introduced to remedy this problem (Ribeiro et al., 2016; Molnar, 2019).

We consider a novel surrogate approach to explainability, where we map the feature embedding learned by the black box models to an *auxiliary* space of explanations. We contend that the best way to arrive at a decision (prediction) may not necessarily be the best way to explain it. While prior work is largely limited to the input encodings, by designing a set of auxiliary attributes we can provide explanations at desired levels of complexity, which could (for example) be made to suit the language learner’s ability in our motivating setting. Our techniques overcome issues in prior art in our setting and are completely language-independent, with potential for use in broader natural language processing and other deep learning explanations.

* Equal contribution

Proceedings of the Third Workshop on Computational Typology and Multilingual NLP, pages 90–96
June 10, 2021. ©2021 Association for Computational Linguistics

For illustration, we examine French in detail where the explanations require both meaning and form.

2 Related Work

We consider the problem of obtaining rules for assigning grammatical gender, which has been extensively studied in the linguistic context (Brugmann, 1897; Konishi, 1993; Starreveld and La Heij, 2004; Nelson, 2005; Nastase and Popescu, 2009; Varlokosta, 2011), but these studies are often limited to identifying semantic or morpho-phonological rules specific to languages and language families. In computational linguistics, prediction models have been discussed in contextual settings (Cucerzan and Yarowsky, 2003) and the role of semantics has been discussed (Williams et al., 2019). Williams et al. (2020) use information-theoretic tools to quantify the strength of the relationships between declension class, grammatical gender, distributional semantics, and orthography for Czech and German nouns. Classification of gender using data mining approaches has been studied for Konkani (Desai, 2017). In this work we look at explainable prediction using neural models.

The noun gender can be predicted better by considering the word form (Nastase and Popescu, 2009). Rule-based gender assignment in French has been extensively studied based on both morphonological endings (Lyster, 2006) and semantic patterns (Nelson, 2005). These studies carefully form rules that govern the gender, argue merits and demerits that often involve factors beyond what rules concisely explain the patterns. Further they are organized as tedious lists of dozens of rules, and evaluated only manually on smaller corpora (less than 8% the size of our dataset). Cucerzan and Yarowsky (2003) show that it is possible to learn the gender by using a small set of annotated words, with their proposed algorithm combining both contextual and morphological models. The encoding of grammatical gender in contextual word embeddings has been explored for some languages in Veeman and Basirat (2020). They find that adding more context to the contextualized word embeddings of a word is detrimental to the gender classifier's performance. Moreover these embeddings often learn gender from contextual agreement, like associated articles, which are not suitable for explanation (Lyster, 2006). In contrast, here we will study the role of semantics in gender determination by learning an encoding of the lexical definition of the word, along with the role of form.

In modern applications of machine learning, it is often desirable to augment the model predictions with *faithful* (accurately capturing the model) and *interpretable* (easily understood by humans) *explanations* of “why" an algorithm is making a certain prediction (Samek et al., 2019). This is typically formulated as an *attribution problem*, that is one of identifying properties of the input used in a given prediction, and has been studied in the context of deep neural feedforward and recurrent networks (Fong and Vedaldi, 2019; Arras et al., 2019). The *attributes* are usually just input features (encoding) used in training. By studying how these features, or perturbations thereof, propagate through a network, one obtains faithful explanations which may not necessarily be easy to interpret. In this work, we consider explanations obtained using *auxiliary attributes* which are not used in training, but correspond to a simpler and more intuitive space of interpretations. We learn a mapping of feature embedding (learned by the black-box neural model) to this space, to approximate faithfulness, at the profit of better explanations. A similar local surrogate based approach is considered by (Ribeiro et al., 2016), but it involves domain-specific input perturbations (e.g. deleting words in text, or pixels in image inputs) for explanation.

3 Dataset

We extract French words, their definitions and phonetic representations from Dbnary (Sérasset, 2015), a Wiktionary-based multilingual lexical database. The words are filtered so that only nouns tagged with a unique gender are retained (for example *voile* which has senses with both genders is removed). For words with multiple definitions but the same gender, we retain the one that appears first as the semantic feature. We retrieve 124803 words, which are split 90-10-10 into train, validation and test sets respectively. The class distribution of the resulting dataset is not skewed, with 58% masculine and 42% feminine words.

4 Methods

4.1 Models

Baselines. We consider two baselines. The *majority* baseline always predicts the masculine gender, while the *textbook* orthographic baseline is based on the following simple rules — predict masculine unless the word ends in `-tion`, `-sion`, `-té`,

-son, or -e, excepting -age, -me or -ège endings.

Semantic models (SEM). The definition of words is used to generate its semantic representation. These are tokenized on whitespace, and are then passed through a trainable embedding layer. These representations are passed through 2 layer bidirectional LSTM of size 25 each, with additive attention. The hidden representation is passed through fully connected layers, of sizes 1500, 1000 and 1. The last layer output is used to calculate cross entropy loss. The representations generated by the penultimate layer (size 1000) is the LSTM semantic embedding.

XLM-R semantic embedding is also generated for the defintion using XLM-R (Conneau et al., 2020). The [CLS] token is fine-tuned to predict the gender. The sequence of hidden states at the last layer represents the embedding.

Phonological model (PHON). To represent the phonology of a word, we use n-grams features, which are constructed by taking last n characters of the syllabized phoneme sequence (derived from Wiktionary IPA transcriptions) where n is in $\{1, 2, \ldots, k\}$ for an empirically set k. A logistic classifier is trained using these features to predict the gender.

Orthographic model (ORTH). To encode the orthography of a word, we use two models. As with phonology, we consider n-grams features, which are constructed here by taking last n characters of the word spelling where n is in $\{1, 2, \ldots, k\}$ for an empirically set k. A logistic classifier to predict the gender is trained using these features.

To generate dense representations for these features, the words are tokenized at character level. The tokens are passed through a 32 unit LSTM and then 2 fully connected layers of sizes 30 and 1. The output from the last layer is used to calculate cross entropy loss by comparing with the true gender labels. Once trained, the representation of penultimate layer (of size 30) is used as the orthographic embedding.

Combined models. A logistic classifier is trained on the concatenated orthographic and semantic features embeddings to discriminate between genders. This is done for both types of semantic embeddings, from LSTM and XLM-R models. We also add phonemic n-gram sequences (n is a hyperparameter set to a jointly optimal value here) as an additional model. All models and their test and validation accuracies are summarized in Table 1.

4.2 Explainability

For each word, we calculate a set of easy-to-interpret *auxiliary features*, with semantic or orthographic connotations. Orthographic features are the top 1000 n-grams in a logistic regression fit. For semantic features, we calculate the scores of the meanings of the words by using word vectors implemented in SEANCE (Crossley et al., 2017). The assignment of words to psychologically meaningful space can lead to increased interpretability. SEANCE package reports many lexical categories for words based on pre-existing sentiment and cognition dictionaries and has been shown by Crossley et al. (2017) to outperform LIWC (Tausczik and Pennebaker, 2010). As SEANCE is only available for the English language, we use translation[1] of the French definitions to English.

Global explanations. The global explanations are evaluated for *i*) masculine and feminine class predictions and for *ii*) classes generated by clustering the best performing combined model embeddings (Table 1). The embeddings are clustered using BIRCH (Zhang et al., 1996) into 10 clusters. The number of clusters are chosen to minimize the overall misclassification rate (calculated by assigning the majority predicted class to a cluster). Decision tree classifiers are fit using the interpretable features[2] of about 25k samples (including those for which an explanation is to be generated) to predict the black box model's gender prediction and the cluster of a word.

Local explanations. We extend the LIME approach of (Ribeiro et al., 2016) to our setting. A local decision tree classifier is trained on the k nearest neighbors of a given test point, to approximate the black box model on the neighborhood.

The size of the decision tree is a hyperparameter which may be reduced to improve interpretability (i.e. smaller, more easily understood explanations) at the cost of model faithfulness (Figure 3).

[1] azure.microsoft.com/en-us/services/cognitive-services/translator/. The authors manually verified the accuracy of translations, the word error rate was less than 2% on a sample of 250 words.

[2] Not to be confused with 'interpretable' and 'uninterpretable' features from formal linguistics (Svenonius, 2006).

5 Results and discussion

The best orthographic model achieves an accuracy of 92.5%, whereas the semantic model alone achieves only 77.23%. Combining the features from the two models leads to a gain in the accuracy of the classifier, to 94.01%. We can conclude that for French, the gender can be predicted robustly by the word orthography, but adding semantic information can further improve prediction. Adding phonology to the mix does not seem to help much. This may be attributed to the fact that phonological forms contain less information than the orthographical forms in French, e.g. *lit* /li/ (bed, m.) and *lie* /li/ (dregs, f.). Not only are the written forms phonetic here (i.e. pronunciation is typically unambiguous given spelling) but they often contain additional (e.g. etymological) information which may be missing in the spoken forms. A more detailed error analysis and comparison of model pairs is presented in Appendix A.

Model	**Test**	**Val**
Majority baseline	57.76	57.98
"Textbook" ORTH rules	83.69	84.10
LSTM (SEM)	76.30	77.13
XLM-R-base (SEM)	77.29	78.71
[N-grams(PHON)]logistic	81.67	81.24
[N-grams(ORTH)]logistic	86.30	86.28
[N-grams(ORTH+pos)]logistic	92.50	92.12
[LSTM(ORTH)]logistic	92.21	92.22
[LSTM(ORTH)+N-grams(PHON)]logistic	92.69	92.40
[LSTM(ORTH)+XLM-R(SEM)]logistic	93.84	93.82
[LSTM(ORTH)+LSTM(SEM)]logistic	94.01	94.00
[LSTM(ORTH)+N-grams(PHON)+LSTM(SEM)]logistic	94.09	93.73

Table 1: Accuracy results of various models on test and validation sets.

We define a 'good explanation' to be one with high model fidelity (measured by F1) and if it involves fewer rules (more easily interpretable). This can be quantified in the case of decision trees as the length of path from root to leaf node, when making a prediction. A class with higher average decision tree path length for its sample is less interpretable.

We observe the trade-off between achieving interpretability and model accuracy for masculine and feminine classes (Figure 1) and for clusters generated via embeddings (Figure 2). The clusters are generated so that within a gender class, a distinction could be made for nouns that could have different rules, so that easier explanations per class could be generated. Both Figures 1 and 2 show that increasing size of the tree, always increases F1 score, but that comes at the cost of interpretability due to higher number of decision rules. Some example features that distinguish the different clusters are noted in Appendix B.

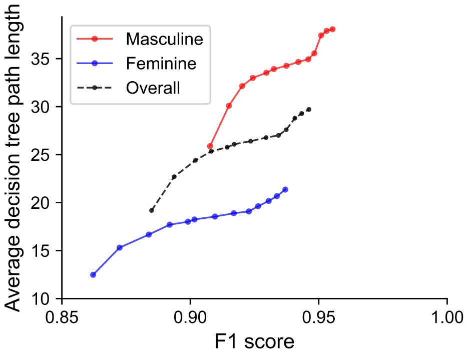

Figure 1: Class-specific/overall explainability (interpretability vs. fidelity) trade-off.

We see in Figure 1 that the explainability is higher for feminine nouns than masculine. This is consistent with the fact that there are many rules to indicate the feminine gender (such as words ending in *-ine, -elle, -esse*), whereas masculine gender is a default category leading to more complex, and harder to explain rules.

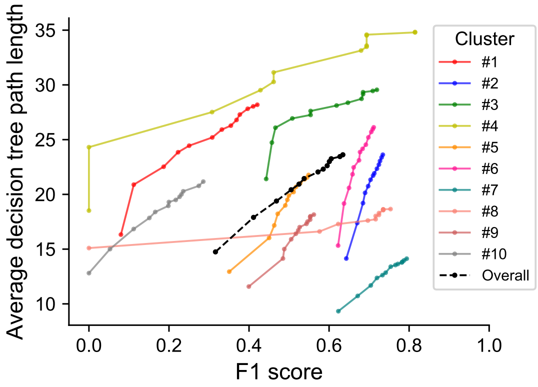

Figure 2: Cluster-specific explainability trade-off.

For the clusters, the misclassification rate for validation and testing set are 4.07% and 4.11% respectively, indicating that clusters mostly have one kind of gender. Figure 2 shows that some clusters (such as #2, #6, #7) are more explainable than the others (such as #1, #4), as latter show a poor F1 performance and low interpretability. Cluster #1 is majority feminine and #4 is majority masculine, indicating existence of exceptions in either gender. Identifying these clusters in the feature embedding can help in figuring out cases where the grammatical gender is assigned for formal reasons, in exception to semantic or morphonological rules. Moreover, these may be useful in designing a system with human-in-the-loop curation, for example by identifying relevant new auxiliary attributes.

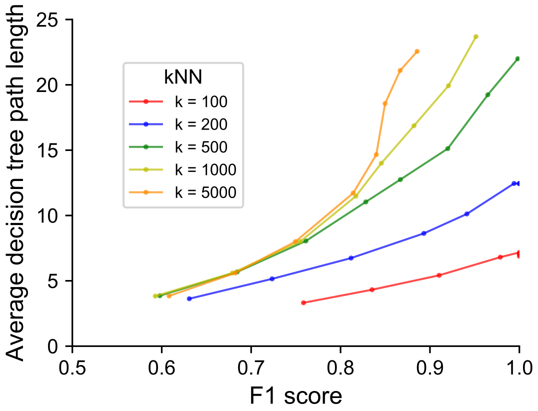

Figure 3: Explainability trade-off for local explanations for various neighborhood sizes.

The local explanations seem to outperform global ones, and the performance improves as we reduce the size of the local neighborhood considered. However, we note that this comes at some cost to consistency of explanations. For example, two local explanations for test points distant in the feature embedding may contain some contradictory rules. This is usually not an issue in typical applications of LIME which simply highlight part of the input as an explanation to provide some model justification. However, inconsistent rules can be of consequence in some applications considered here, for instance language learning where these contradictions are undesirable. Also, while per example explanations are larger on average for the global approach, we have the same rule for entire clusters, giving fewer rules overall.

6 Conclusion

Orthography predicts the grammatical gender in French with high accuracy, and adding semantic features can improve this prediction. The black-box embedding can be explained by simpler decision tree models over a given auxiliary explanation space, both locally and globally. Global explanations lead to fewer rules across examples but are more complex on individual instances. Explainable gender prediction can be useful to language learners and gender bias researchers. A cross-linguistic extension of our study is deferred to future work.

Acknowledgements

We thank the reviewers for their useful feedback.

References

Leila Arras, José Arjona-Medina, Michael Widrich, Grégoire Montavon, Michael Gillhofer, Klaus-Robert Müller, Sepp Hochreiter, and Wojciech Samek. 2019. Explaining and interpreting lstms. In *Explainable ai: Interpreting, explaining and visualizing deep learning*, pages 211–238. Springer.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Karl Brugmann. 1897. *The nature and origin of the noun genders in the Indo-European languages: A lecture delivered on the occasion of the sesquicentennial celebration of Princeton University*. C. Scribner's sons.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2017. Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49(3):803–821.

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

Ms Shilpa Desai. 2017. Data mining techniques for konkani grammatical gender identification. *Fr. Agnel College of Arts & Commerce Re-accredited by NAAC with "A" Grade Pilar-Goa*, page 38.

Carla Ellis, Simone Conradie, and Kate Huddlestone. 2012. The acquisition of grammatical gender in l2 german by learners with afrikaans, english or italian as their l1. *Stellenbosch Papers in Linguistics*, 41:17–27.

Ruth Fong and Andrea Vedaldi. 2019. Explanations for attributing deep neural network predictions. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 149–167. Springer.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

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

Toshi Konishi. 1993. The semantics of grammatical gender: A cross-cultural study. *Journal of psycholinguistic research*, 22(5):519–534.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774.

Roy Lyster. 2006. Predictability in french gender attribution: A corpus analysis. *Journal of French Language Studies*, 16(1):69.

Christoph Molnar. 2019. Interpretable machine learning.

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.

Don Nelson. 2005. French gender assignment revisited. *Word*, 56(1):19–38.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

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.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Steven Samuel, Geoff Cole, and Madeline J Eacott. 2019. Grammatical gender and linguistic relativity: A systematic review. *Psychonomic bulletin & review*, 26(6):1767–1786.

Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361.

Samantha R. Sexton. 2020. Cross linguistic analysis of grammatical gender: Implications for critical language pedagogy. *Thesis, Linguistics and Education departments of the University of Massachusetts Amherst.*

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR.

Peter Starreveld and Wido La Heij. 2004. Phonological facilitation of grammatical gender retrieval. *Language and Cognitive Processes*, 19(6):677–711.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Peter Svenonius. 2006. Interpreting uninterpretable features. *Linguistic Analysis.*

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Spyridoula Varlokosta. 2011. The role of morphology in grammatical gender assignment. *Morphology and its interfaces*, 178.

Hartger Veeman and Ali Basirat. 2020. An exploration of the encoding of grammatical gender in word embeddings. *arXiv preprint arXiv:2008.01946.*

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 5738–5743.

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

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114.

Appendix

A Error analysis

We examine in detail the errors of all our models. Some salient observations are noted below. The errors of our baselines indicate their insufficiency but are easier to understand in isolation. For our models, it is perhaps best to look at interesting pairs of models and compare their errors.

ORTH+SEM vs. ORTH: Adding phonology did not seem to help much in predicting gender beyond

Cluster	Majority gender	Error	Top-10 features
#1	Masculine	0.05	Role_GI, *sme*, *ien*, *n*, *sion*, *ade*, *che*, *nce*, *ue*, *ière*
#2	Feminine	0.08	*ien*, *sion*, *n*, *ade*, *r*, *che*, *ière*, *nce*, Role_GI, *ue*
#3	Masculine	0.00	*rice*, *sme*, *n*, *ien*, *age*, Ptlw_Lasswell, *tte*, *lle*, *té*, *ite*
#4	Masculine	0.00	Polit_2_GI, negative_adjectives_component, *age*, *ois*, *ne*, *se*, *ie*, *tion*, *ée*, *ite*
#5	Masculine	0.00	*sme*, *ien*, *n*, Social_GI, *sion*, *ade*, *che*, *ue*, *té*, *ite*
#6	Masculine	0.03	*l*, *sme*, *ien*, *n*, *sion*, *ade*, *che*, *ue*, *ite*, *té*
#7	Feminine	0.01	*ière*, Quan_GI, Fear_GALC, Role_GI, *ure*, *n*, Rctot_Lasswell, polarity_nouns_component, *ée*, *r*
#8	Feminine	0.00	*sme*, *ade*, *sion*, *ien*, *n*, *ière*, *che* , *nce*, *r*, *tte*
#9	Masculine	0.00	*ade*, *sion*, *che*, *nce*, *ue*, *ure*, *ée*, *ité*, *té*, *ite*
#10	Feminine	0.00	*ologie*, *n*, *r*, Fear_GALC, Anticipation_EmoLex, *ière*, *che*, Role_GI, *nce*, *ue*

Table 2: Top-10 features from decision tree with at most 500 leaf nodes for the clusters defined in Section 4.2.

orthography itself. Even though phonology alone (PHON) is more accurate than the best semantics (SEM) model in predicting gender (81% vs. 77%), semantics provide more useful additions over what orthography already encodes. For example, *poix* (meaning "pitch" or "tar"), *polio* ("polio") and *ardeur* ("ardor") are recognized as feminine with help from semantics (ORTH+SEM) but are classified incorrectly by the ORTH model. Similarly the meaning helps identify that *brais* ("crushed barley"), *polyane* ("plastic film") and *jurisconsulte* ("law expert") should be classified as masculine.

ORTH vs. PHON: Some examples which are correctly classified by the ORTH model but misclassified by the PHON model include *meringue* ("meringue", f.), *boulaie* ("birch grove", f.), *coccyx* ("coccyx", m.) and *explicit* ("end of a chapter or book", m.).

ORTH+SEM: Finally we look at errors of our best model (we consider ORTH+SEM as better than ORTH+SEM+PHON as it gets the same accuracy with fewer features). The list seems to include relatively rarer words, where it often seems hard to explain the gender assignment. Some examples are — *myrsite* ("Old medical wine", m.), *fomite* ("inanimate disease vector", m.) *cholestrophane* ("a chemical derived from caffeine", f.), *interpolateur* ("interpolator", f.).

B Auxiliary features for global explanations

For the 10 clusters described for global explainability in section 4.2, we show the top-10 important features in Table 2. These features are generated by training a decision tree classifier that could have at most 500 leaf nodes. The importance of a feature in each cluster was defined by the number of times it appeared on the decision path of the samples. The features are a mix of orthographic features (generated from word endings) and semantic features (generated from SEANCE) [3]. We emphasize that the features noted here are determined as the most common features for examples in the cluster, and are therefore more likely to appear in explanations of examples from that cluster — the exact explanation for an example is determined by the appropriate decision tree path.

The Table 2 also shows the error rates per clusters, which are fraction of misclassified labels per cluster with respect of predictions from the combined black-box model.

[3] Feature descriptions may be found at the following link: https://drive.google.com/file/d/1SUfSYNyuaWT2i4tQkiyr2rxVeqnh3cQe/view