The role of morphology in gender assignment in French

Olivier Bonami U. de Paris
Matías Guzman Naranjo U. Düsseldorf
Delphine Tribout U. de Lille

Corbett (1991) played a central role in establishing that grammatical gender assignment is far from being arbitrary. Although there are usually few categorical generalizations, semantic, morphological, and phonological properties of a noun typically contribute to predicting its grammatical gender. While this is clearly established, much remains to be explored on the exact nature of the predictors and the way they interact. As a case in point, consider the situation in French. At least since Tucker et al. (1977), it is firmly established that all three types of predictions have a role to play, as exemplified in (1).

(1) a. Semantic prediction: Nouns referring to properties are overwhelmingly feminine.
b. Morphological prediction: VN compounds denoting inanimates are all masculine.
c. Phonological prediction: nouns ending in /jɔ̃/ are overwhelmingly feminine

While such generalizations are clearly correct, there are often correlations between them that make it nontrivial to establish their exact status. For instance, property nouns are also overwhelmingly derived from adjectives by suffixes uniformly outputting feminine nouns (*-ité*, *-eur*, etc.), and nouns ending in the sequence /jɔ̃/ are overwhelmingly formed using the derivational suffix *-ion*.

This abstract presents a quantitative study of gender assignment in French that aims at disentangling the role of morphology and phonology in gender assignment. We start from the observation that about one third of French nouns with a unique gender end in a derivational suffix (see below), and that the relevant suffixes are for the most part compatible with only one gender (Bonami & Boyé, 2019).[1] This suggests that the predictability attributed to phonology since Tucker et al. (1977) could to a large extent be attributable to morphology. To investigate this issue, we annotated by hand a sample lexicon of 3,683 nouns for their phonological and morphological properties. We then expanded on Sokolik & Smith (1992) and Matthews (2005) by training neural networks to learn gender assignment on the basis of phonological and/or morphological predictors. We conclude that morphology plays a subtle role in gender assignment: while phonology is a very good predictor of gender on its own, this is to a large extent due to the way derivational morphology shapes the phonotactic properties of the lexicon.

1 Data collection and annotation

The sample of nouns we used in this study was randomly selected among the nouns contained in the *Lexique 3* database (New et al., 2007), limiting attention to wordforms found in only one gender and lemmas with a frequency above 0.3 per million words. The morphological annotation proceeded as follows. In a first step, we relied on previously published manually curated lexica: 760 nouns found in the lexicon of simplex nouns presented in (Tribout et al., 2014) were tagged as simplex, and 1,019 nouns were tagged as instances of conversion, either because they were listed in Tribout's (2010) database, or because they were homographic to a verb or adjective. We then proceeded to annotate manually the remaining 1,904 nouns. There

[1] According to Bonami & Boyé (2019), at least 25% of all French nouns are common gender, i.e. come in pairs of homophonous masculine and feminine nouns, and this proportion is rising steadily in recent years. We disregard such cases in the present study.

was a double morphological annotation. On the one hand, we noted the type of outermost word-formation process (prefixation suffixation, compounding, etc.), and, in the case of affixation processes, the identity of the affix. On the other hand, we noted the outermost suffix, if any was present (non-suffixed nouns were annotated as '0'). This was motivated by the presumption that a suffix might be relevant to gender assignment even where suffixation is not the last operation to have applied: e.g. the feminine gender of *contre-proposition* 'counter-proposal' can be tracked down to the presence of the suffix *-ion*, despite the fact that prefixation of *contre* is the outermost process. In the end, our dataset contains 1,222 nouns ending in a suffix (henceforth 'suffixed nouns') and 2,461 nouns not ending in a suffix (henceforth 'unsuffixed nouns'). Finally, we added to the database phonological transcriptions and syllable boundaries as documented in the GLÀFF lexicon (Hathout et al., 2014).

2 Modelling

We want to explore two main questions: (i) to what degree is gender predictable for French nouns, and (ii) what are the roles of phonological and morphological factors in gender assignment. To answer these questions we train several multilayer perceptrons to predict gender based on morphological and phonological predictors.[2] For the phonological predictors we extracted the last three segments of the noun, the number of syllables and number of segments. As morphological predictor we used the annotated suffix, or 0 for unsuffixed nouns.

We trained the models using `caret` (Kuhn, 2008) with `MxNet` (Chen et al., 2015). The perceptrons had 3 hidden layers with 128, 4, 2 neurons, respectively.[3] The results reported for each model are the aggregated accuracy and kappa scores[4] of 10-fold cross-validation. Our model choice obeyed two main reasons. First, we wanted to keep models consistent across datasets. Second, multilayer perceptrons have been shown to perform well in similar gender/class assignment tasks (Matthews, 2005).

First we consider the whole dataset. Table 1 shows the result of three models: one model with only morphological predictors, one with only phonological predictors, and one model with both phonological and morphological predictors. These initial results clearly show that gender is highly predictable in French nouns. The model trained on morphological predictors only shows that a large portion of the variation is due to morphology. On the other hand, the model trained only on phonological predictors reached the same accuracy as the model using both morphological and phonological predictors. This effect is likely due to the fact that phonological predictors are a good proxy for morphological markers.

Since morphological gender assignment mainly happens on suffixed nouns, we now fit similar models focusing exclusive on these. Table 2 shows the results for this set of models. For this set of nouns, the suffix overwhelmingly determines the gender of the noun, and adding phonological to morphological information does not lead to any increase in accuracy. Nonetheless, because the phonological predictors are a good proxy for the morphological predictors, the model with only phonological predictors reaches a similarly high accuracy.

Next we turn to nouns without a suffix. Table 3 shows the results for this group of nouns.[5]

[2]We also tried to use animacy information as an extra predictor, based on the intuition that this might help discriminate e.g. inanimate abstract feminine nouns in *-eur* such as *blancheur* 'whiteness' from animate masculine agent nouns such as *menteur* 'liar'. As it turns out, in none of the conditions described below does the addition of animacy lead to a discernible improvement of model accuracy. Hence we report results for the simpler models without animacy.

[3]We tweaked the momentum and learning rate of each network to ensure convergence. All layers had relu activation. The models were trained using an Nvidia Titan Xp (this card was donated to us by the NVIDIA Corporation).

[4]This metric measures how much better than random chance are the results of the model.

[5]For this dataset we did not train models with morphological predictors.

Predictors	Morphology		Phonology		Both	
	Reference		*Reference*		*Reference*	
Prediction	F	M	F	M	F	M
F	629	32	1092	264	1139	278
M	879	2143	416	1911	369	1897
Accuracy	0.75		0.82		0.82	
95% Accuracy's CI	(0.74, 0.77)		(0.80, 0.83)		(0.81, 0.84)	
Kappa	0.44		0.61		0.63	

Table 1: Three models for the whole dataset.

Predictors	Morphology		Phonology		Both	
	Reference		*Reference*		*Reference*	
Prediction	F	M	F	M	F	M
F	631	18	609	47	639	20
M	22	551	44	522	14	549
Accuracy	0.97		0.93		0.97	
95% Accuracy's CI	(0.96, 0.98)		(0.91, 0.94)		(0.96, 0.98)	
Kappa	0.93		0.85		0.94	

Table 2: Three models for suffixed nouns.

The results in this table show that gender is highly predictable for nouns without a suffix, but the error rate is, as expected, much higher than the error rate for nouns with a suffix.

	Reference	
Prediction	F	M
F	500	269
M	355	1337
Accuracy	0.75	
95% Accuracy's CI	(0.73, 0.76)	
Kappa	0.43	

Table 3: Phonological prediction of unsuffixed nouns.

3 Discussion

The results of our modelling experiments paint a subtle picture of the role of morphology in gender assignment. On the one hand, explicit use of morphological information plays a minor role in accurate prediction of gender: on the whole dataset, a model relying on both morphology and phonology does not outperform a model relying on phonology only; and even on the subset of suffixed nouns, the gain in accuracy of taking explicit morphological information into account is quite limited. This leads to the speculation that speakers may not need to attend to morphological information to correctly assign gender. Only psycholinguistic experimentation will be able to tell whether they do.

On the other hand, morphology plays a crucial role in shaping the phonotactic makeup of the lexicon, in such a fashion that phonological prediction of gender is much more accurate for suffixed than for unsuffixed nouns. Examination of conditional probability distributions between various variables estimated from our dataset help understand the causes of this situation.

Here we use conditional entropy as a rough indication of interpredictibility between variables. First, while the ultimate cause of predictability of gender in suffixed nouns is the fact that suffixes assign gender almost categorically ($H(\texttt{gender}|\texttt{suffix}) = 0.04$), the suffix itself is quite well predicted by the three last segments of the words ($H(\texttt{suffix}|\texttt{last\_3\_segments}) = 0.18$); hence, knowledge of the suffix adds little to knowledge of the last three segments when predicting gender ($H(\texttt{gender}|\texttt{last\_3\_segments}) = 0.06$; $H(\texttt{gender}|\texttt{last\_3\_segments}, \texttt{suffix}) = 0.02$), since phonology alone already approaches categorical prediction. Second, the final substrings of suffixed words and unsuffixed words are different enough on average that final substrings are a decent predictor of whether a word is suffixed ($H(\texttt{suffixed}|\texttt{last\_3\_segments}) = 0.30$). This indicates that unsuffixed nouns which happen to end in a sequence that could be a suffix are not frequent enough to strongly impair prediction of gender by phonology, and helps explain the absence of a contribution of morphological information to accuracy of prediction on the whole dataset.

One conclusion of this study is that explicit morphological knowledge plays a distinct but limited predictive role in the case of suffixed nouns. The limited amplitude of that role may be due to the fact that our models do not take into account any semantic information: it may be the case that semantic and phonological information are jointly sufficient to reach optimal accuracy. We will investigate that issue in the near future.

References

Bonami, Olivier & Gilles Boyé. 2019. Paradigm uniformity and the French gender system. In Matthew Baerman, Oliver Bond & Andrew Hippisley (eds.), *Perspectives on morphology: Papers in honour of Greville G. Corbett*, Edinburgh: Edinburgh University Press.

Chen, Tianqi, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang & Zheng Zhang. 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274* .

Corbett. 1991. *Gender*. Cambridge: Cambridge University Press.

Hathout, Nabil, Franck Sajous & Basilio Calderone. 2014. GLÀFF, a large versatile French lexicon. In *Proceedings of LREC 2014*, .

Kuhn, Max. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software, Articles* 28(5). 1–26.

Matthews, Clive A. 2005. French Gender Attribution on the Basis of Similarity: A Comparison Between AM and Connectionist Models. *Journal of Quantitative Linguistics* 12. 262–296.

New, Boris, Marc Brysbaert, Jean Veronis & Christophe Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* 28. 661–677.

Sokolik, M. E. & Michael E. Smith. 1992. Assignment of gender to french nouns in primary and secondary language: a connectionist model. *Second Language Research* 8(1). 39–58.

Tribout, Delphine. 2010. *Les conversions de nom à verbe et de verbe à nom en français*: Université Paris Diderot - Paris 7 dissertation.

Tribout, Delphine, Lucie Barque, Pauline Haas & Richard Huyghe. 2014. De la simplicité en morphologie. In *Actes de CMLF 2014*, 1879–1890.

Tucker, G Richard, Wallace E Lambert & André Rigault. 1977. *The French speaker's skill with grammatical gender: An example of rule-governed behavior*. Berlin: Mouton De Gruyter.