# The role of morphology in gender assignment in French

Olivier Bonami[1] Matías Guzman Naranjo[2] Delphine Tribout[3]

[1]Université de Paris, LLF, CNRS
[2]Heinrich-Heine-Universität Düsseldorf
[3]Université de Lille

ISMo 2019 — Paris — September 26, 2019

# Predicting gender in French

Gender assignment in French is highly predictable on the basis of phonology, morphology and/or semantics (Tucker, Lambert, and Rigault, 1977; Surridge, 1985, 1986, 1989; Corbett, 1991):

- Semantic prediction:
  - With human nouns: social gender tends to predict grammatical gender
  - Color terms are masculine
  - Property nouns are feminine
  - …
- Morphological prediction:
  - Inanimate VN compounds are masculine
  - Nouns suffixed in *-ion* are feminine
  - …
- Phonological prediction:
  - The last segment of a noun is often an excellent predictor of gender, e.g.
    - Nouns ending in /ɑ̃/ are overwhelmingly masculine
    - Nouns ending in /n/ strongly tend to be feminine
    - …
  - Feminine nouns are on average a bit longer than masculine nouns (6.6 vs. 6.2), hence word length partially predicts gender.

# Disentangling predictors

- Most previous computational/experimental studies of gender assignment either computationally or experimentally do not attempt to disentangle predictors (Tucker, Lambert, and Rigault 1977; Holmes and Segui 2006; Matthews 2005).
- Deciding what the ultimate predictors are is hard, because of strong correlations between phonology, morphology and semantics.
  1. Property nous are feminine, but property nouns are also overwhelmingly deadjectival nouns formed with a specific set of suffixes: *-erie*, *-esse*, *-eur*, *-ice*, *-ie*, *-ise*, *-(i)té*, *-itude* (Koehl, 2012).
     - Is the correct generalization in terms of the semantics of the output or properties of the affix producing this output?
  2. Most nouns ending in /ʒ/ are masculine, but this tendency is due to the prevalence of the suffixe *-age*.

| Gender | Simplex | *-age* |
|---|---|---|
| FEM | 14 | 0 |
| MAS | 14 | 75 |

Number of items ending in /ʒ/ in a sample of 3683 nouns

     - In such a situation, is it fair to say that phonology is predictive?

# Predicting from semantics? I

Two aspects:

1. Among human nouns, social gender correlates grammatical gender.
   - This is a lot more subtle than usually assumed.

     (1) a. la présidente ⇒ female

     b. le président ⇒ male or female (cf. *Madame le président*)

   - When referring to a woman, the choice of gender is imbued with social meaning (Burnett and Bonami, 2019a,b)
   - The structure of the French lexicon is in flux here, with a quick rise in both the type and token frequency of feminine nouns (Bonami and Boyé, 2019).
   - Given this, we do not have any clear expectation as to semantic predictability.

# Predicting from semantics? II

2. Among inanimate nouns, experimental (Boroditsky, Schmidt, and Phillips, 2003) and corpus (Dye et al., 2017; Williams et al., accepted) evidence of correlations between gender and semantics. However:
   - Effect sizes are small, hence the actual predictive power of semantics might be very low.
   - Unclear what semantic dimensions exactly play a role.
   - Although it is tempting to assume that semantic analogies to social gender and/or biological sex play a role (Damourette and Pichon, 1930; Boroditsky, Schmidt, and Phillips, 2003; Aikhenvald, 2016), various attempts to document it firmly have failed (Landor, 2014; Mickan, Schiefke, and Stefanowitsch, 2014).

For now we stick to prediction from phonology and morphology.

# Outline

1. Data collection and annotation
2. Modelling strategy
3. Results and discussion

# Data collection and annotation

# Data collection

- Random sample of 3750 nouns from the *Lexique 3* database (New et al., 2007) such that:
  - The form is gender-specific:
    - No common gender nouns (e.g. $artiste_M$ 'male artist', $artiste_F$ 'female artist')
    - No gender-constrasting homographs (e.g. $livre_M$ 'book', $livre_F$ 'pound')
  - Average of relative lemma frequencies in the two reference corpora (20th century literature, subtitles) is above 0.15 per million token.
- After elimination of tagging errors and missing annotation (see below), we ended up with 3683 data points.

# Annotation

**Phonology** Phonemic transcriptions taken from the (wiktionary-derived) GLÀFF lexicon (Hathout, Sajous, and Calderone, 2014).

**Animacy** Hand annotation of human/animal/inanimate reference taken over from a previous project.

**Morphology** New, large scale annotation by OB and DT:

1. Type of the outermost process (prefixed, suffixed, compounded, converted, clipped, …, native simplex, borrowed)
2. Identity of the last suffix, if any.
3. Identity of the first prefix, if any.
4. Compounding type, if any.
5. Conversion relation (disregarding directionality), if any.
6. …

# Type of outermost process

| Process | # |
|---|---|
| simplex | 1521 |
| suffixed | 1426 |
| converted | 380 |
| compounds | 174 |
| non concatenative | 78 |
| prefixed | 71 |
| syntactic constructs | 33 |
| total | 3683 |

(non concatenative : acronyms, blends, troncation, reduplication, verlan…)
(simplex: native simplex, antonomasia, borrowing, onomatopeic)

# Suffixed Nouns

- 1558 Ns ending with a suffix, although for 132, suffixation is not the last process (e.g. *survêtement* 'tracksuit', *porte-cigarettes* 'cigarette case')
- 102 distinct suffixes
- 20 most frequent suffixes (75% of the total set of suffixed nouns):

| Suffix | # | Suffix | # |
|---|---|---|---|
| *-ment* | 153 | *-ant* | 35 |
| *-ion* | 148 | *-euse* | 34 |
| *-eur*$_M$ | 136 | *-ière* | 33 |
| *-age* | 75 | *-isme* | 30 |
| *-ier* | 73 | *-et* | 25 |
| *-ette* | 70 | *-eux* | 23 |
| *-on* | 66 | *-ard* | 21 |
| *-ite* | 62 | *-ance* | 19 |
| *-erie* | 52 | *-eur*$_F$ | 17 |
| *-ure* | 50 | *-ade* | 16 |

- Note that:
  - We take gender-contrasting pairs of suffixes (e.g. *-ier* vs. *-ière*, *-eur* vs. *-euse*) to be two distinct but closely related (Bonami and Boyé, 2019).
  - There is some suffix homonymy (*eur*$_M$ for agent and instrument nouns, *eur*$_F$ for property nouns)

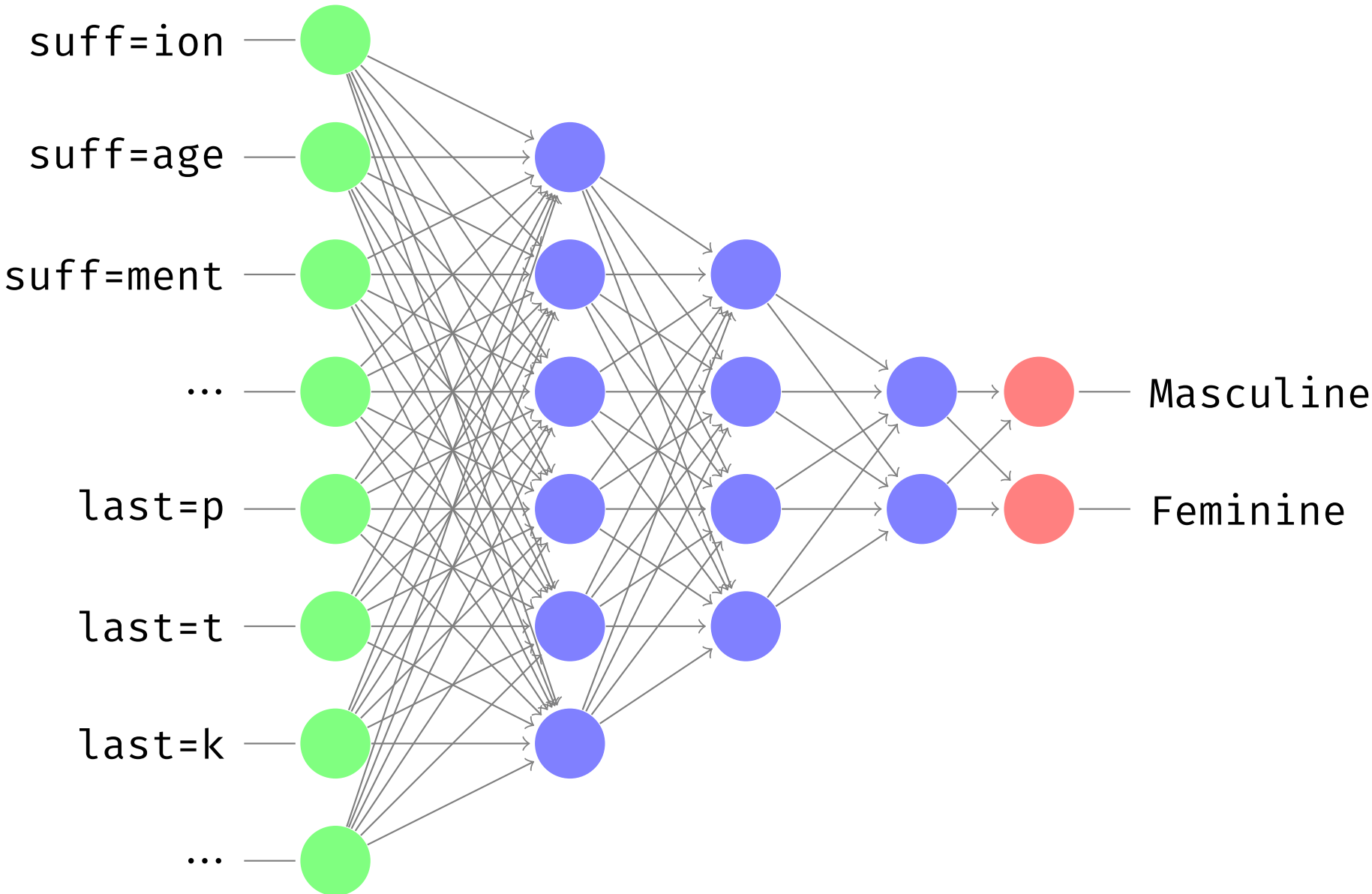
# Modelling strategy

# Gender assignment as classification

- Our goal is to see how well gender is predicted by various combinations of predictors applied to various subsets of our data.
- We could use many types of classifiers to that end—we are not interested in performance of classification *per se*, but in making meaningful comparisons between classifiers with the same architecture trained on different datasets.
- Simple neural networks have proven effective for prediction of lexical classes in general (Guzman Naranjo, 2019) and French gender in particular (Matthews, 2005)

# Neural Networks

- Specifically: perceptrons with 3 hidden layers of 128, 4 and 2 units.



# Confusion Matrix

| | observed | |
|---|---|---|
| predicted | M | F |
| M | $a$ | $b$ |
| F | $c$ | $d$ |

Accuracy: proportion of correct classification

$$\frac{a+d}{a+b+c+d}$$

No Information Rate (NIR): frequency of the largest class

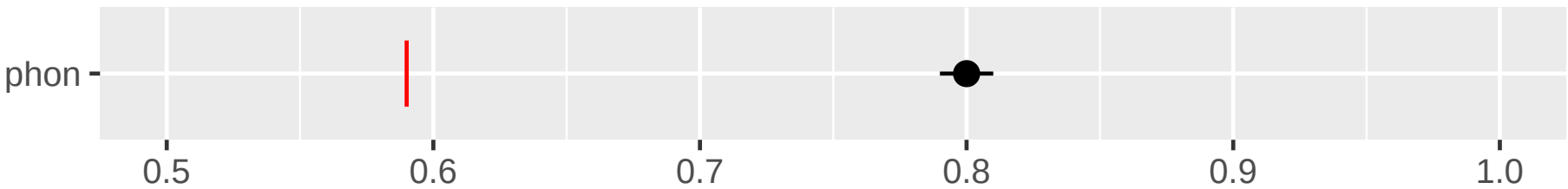
$$\max\left(\left\{\frac{a+c}{a+b+c+d}, \frac{b+d}{a+b+c+d}\right\}\right)$$

The NIR serves as our baseline.

# Reporting results

- In each case we evaluate accuracy using 10-fold cross-validation, and report No Information Rate, average accuracy, and 95% confidence intervals around the average.
- For instance, here are the results for prediction of gender from phonology for the whole dataset ($N = 3683$)

| | M | F |
|---|---|---|
| M | 1125 | 352 |
| F | 384 | 1822 |
| Acc: | 0.80 | |
| 95% CI: | (0.79, 0.81) | |
| NIR: | 0.59 | |



# Results and discussion

# Predictors considered

**Phonology**

- Number of phonemes
- Number of syllables
- Identity of the last three segments

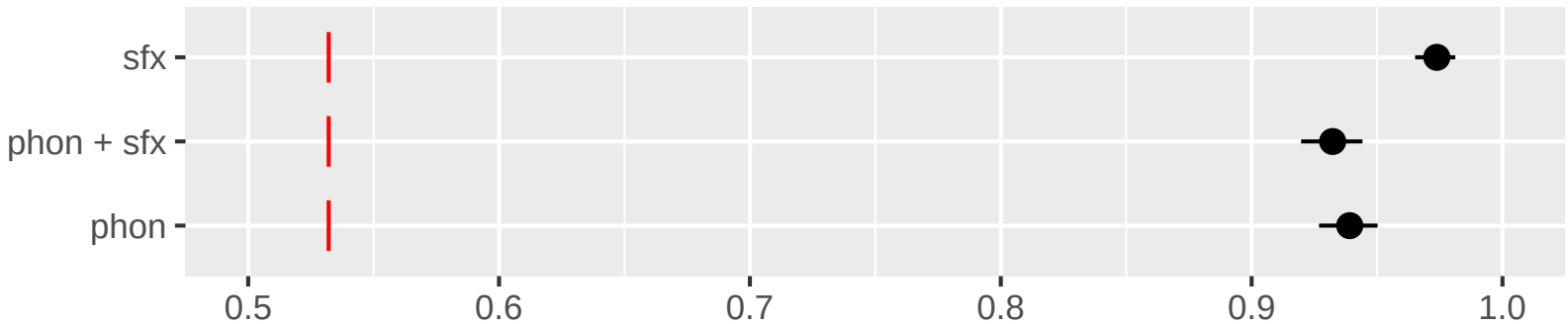
**Morphology**

- Type of outermost process
- Identity of the suffix, if any
- Identity of the prefix, if any

**Semantics** We tried to include animacy in our models, but this turned out to never improve accuracy. We do not report further on this.

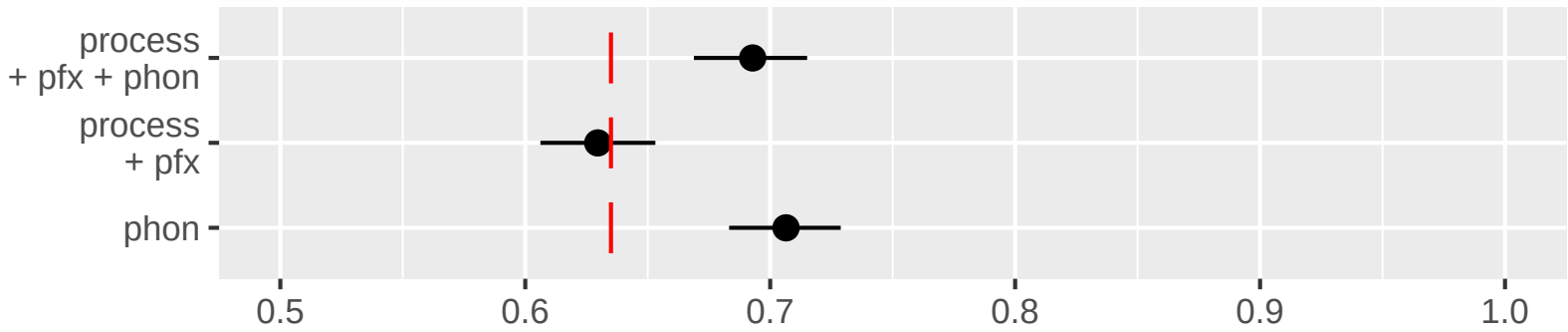
# Results on suffixed nouns ($N = 1558$)

- Prediction from morphology is nearly perfectly accurate, as expected.
- Accuracy of prediction from just phonology is lower but still very high.
- Adding phonology to morphology does not lead to any improvement, but there is no room for improvement anyway.



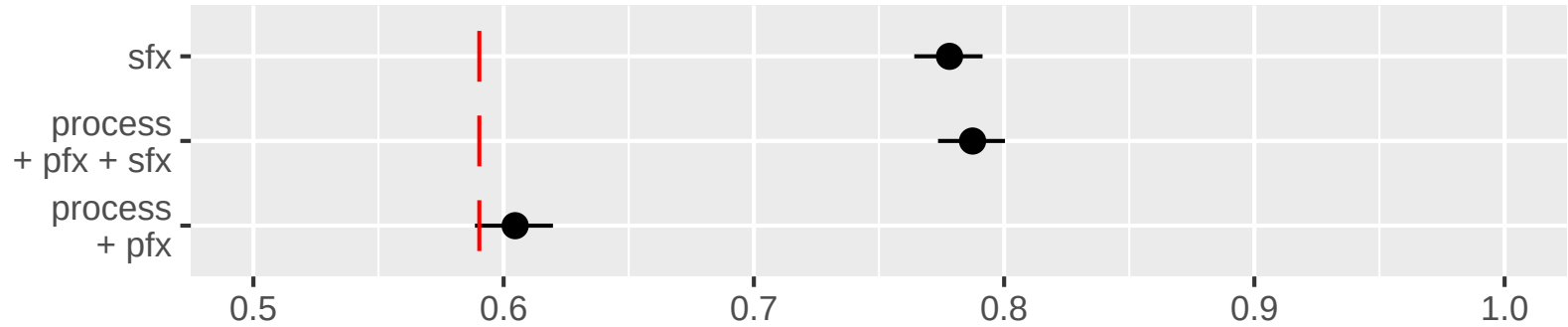
# Results on non-suffixed nouns ($N = 2125$)

- Prediction from morphology is at chance level.
- Prediction from phonology is significant but much lower than what we observed for suffixed nouns.
- Adding morphology to phonology does not lead to any improvement.

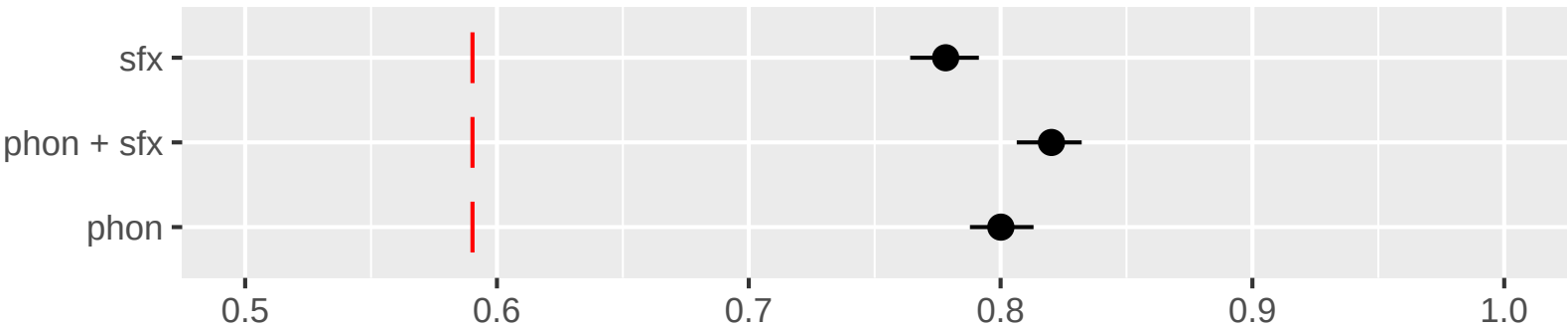


# Results on the whole dataset ($N = 3683$)

- Among morphological predictors, suffixation drives prediction: no evidence that adding prefix and type of outermost process to suffix leads to better accuracy.




- Prediction from phonology only is more accurate than prediction from morphology only.



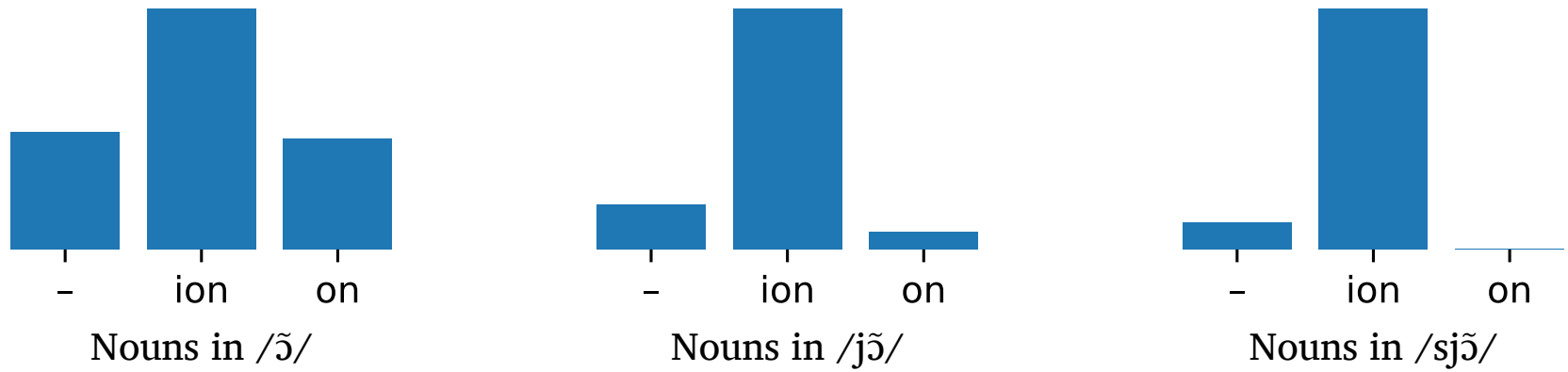

- No evidence that adding morphological predictors to a model already taking into account phonology leads to better accuracy.

# Discussion I

- Conclusions:
  1. Suffixes are excellent predictors of gender. The few cases where one suffix is compatible with two genders (e.g. *-oire*) are not enough to get in the way.
  2. Phonology is an imperfect but quite good proxy for the information provided by suffixes. Homophonous suffixes in both genders (e.g. *-eur*$_M$ vs. *-eur*$_F$) are enough to make a measurable difference, but barely.
  3. On the other hand, phonology does have a sizeable effect on gender predictability, even where suffixes can't help.

# Discussion II

- Why is phonology such a good proxy for morphology?
  - The prevalence of suffixed nouns (42% of our data) entails that morphology has a strong influence on the phonological makeup of the lexicon (Lindsay and Aronoff, 2013).




  - As a consequence, the identity of the last three segments predicts quite well the presence of a specific suffix, so that adding morphological knowledge to phonological knowledge can only lead to small improvements in gender prediction.

# Discussion III

- Important lessons:
  1. Fine predictability:
     - The literature takes note of some fine correlations between gender and some phonological or morphological variables.
     - By assessing overall predictability in a random sample, we establish that these are not statistically powerful enough to have an effect on the accuracy of gender prediction.
     - This highlights the importance of sampling decisions when studying gender predictability.
  2. Semantic predictability:
     - The work of adding good semantic predictors to our models still needs to be done.
     - However we predict little additional effect of semantic prediction, given the very high accuracy we already reach without it.
  3. Processing:
     - The ultimate cause of gender predictability is morphology and the way it shapes the phonological makeup of the lexicon.
     - However, it does not follow that a speaker needs to attend to morphology to attain accurate prediction of gender.
     - Next step: Assess through behavioral experiments which predictors speakers actually rely on.

# References I


Aikhenvald, Alexandra Y. (2016). *How gender shapes the world*. Oxford: Oxford University Press (cit. on p. 5).

Bonami, Olivier and Gilles Boyé (2005). "Construire le paradigme d'un adjectif." In: *Recherches Linguistiques de Vincennes* 34, pp. 77–98 (cit. on p. 31).

— (2019). "Paradigm uniformity and the French gender system." In: *Perspectives on morphology: Papers in honour of Greville G. Corbett*. Ed. by Matthew Baerman, Oliver Bond, and Andrew Hippisley. Edinburgh: Edinburgh University Press, pp. 171–192 (cit. on pp. 4, 11).

Boroditsky, Lera, Lauren A. Schmidt, and Webb Phillips (2003). "Sex, syntax, and semantics." In: *Language in Mind: Advances in the Study of Language and Thought*. Ed. by Dedre Gentner and Susan Goldin-Meadow. Cambridge: MIT Press, pp. 61–79 (cit. on p. 5).

Burnett, Heather and Olivier Bonami (2019a). "A Conceptual Spaces Model of Socially Motivated Language Change." In: *Proceedings of the 2nd Meeting of the Society for Computation in Linguistics* (cit. on p. 4).

— (2019b). "Linguistic Prescription, Ideological Structure and the Actuation of Linguistic Changes: Grammatical Gender in French Parliamentary Debates." In: *Language in Society* 49.1, pp. 65–93 (cit. on p. 4).

Corbett (1991). *Gender*. Cambridge: Cambridge University Press (cit. on p. 2).

Damourette, Jacques and Édouard Pichon (1930). *Des mots à la pensée. Essai de grammaire de la langue française*. Vol. 1. Paris: d'Artrey (cit. on p. 5).

Dye, Melody et al. (2017). "A functional theory of gender paradigms." In: *Perspectives on Morphological Organization: Data and Analyses* 10, p. 212 (cit. on p. 5).

# References II


Guzman Naranjo, Matías (2019). *Analogical classification in formal grammar*. Berlin: Language Science Press (cit. on p. 13).

Hathout, Nabil, Franck Sajous, and Basilio Calderone (2014). "GLÀFF, a Large Versatile French Lexicon." In: *Proceedings of LREC 2014* (cit. on p. 9).

Holmes, Virginia M. and Juan Segui (2006). "Assigning Grammatical Gender During Word Production." In: *Journal of Psycholinguistic Research* 35.1, pp. 5–30 (cit. on p. 3).

Koehl, Aurore (2012). "La construction morphologique des noms désadjectivaux suffixés en français." PhD thesis. Université de Lorraine (cit. on p. 3).

Landor, Roland (2014). "Grammatical categories and cognition across five languages: the case of grammatical gender and its potential effects on the conceptualization of objects." PhD thesis. Griffith University (cit. on p. 5).

Lindsay, Mark and Mark Aronoff (2013). "Natural Selection in Self-Organizing Morphological Systems." In: *Morphology in Toulouse: Selected Proceedings of Décembrettes 7*. Ed. by Nabil Hathout, Fabio Montermini, and Jesse Tseng. Munich: Lincom Europa, pp. 133–153 (cit. on p. 23).

Matthews, Clive A. (2005). "French Gender Attribution on the Basis of Similarity: A Comparison Between AM and Connectionist Models." In: *Journal of Quantitative Linguistics* 12, pp. 262–296. (Visited on 07/31/2016) (cit. on pp. 3, 13).

Mickan, Anne, Maren Schiefke, and Anatol Stefanowitsch (2014). "Key is a llave is a Schlüssel: A failure to replicate an experiment from Boroditsky et al. 2003." In: *Yearbook of the German cognitive linguistics association* 2.1, pp. 39–50 (cit. on p. 5).

# References III


New, Boris et al. (2007). "The use of film subtitles to estimate word frequencies." In: *Applied Psycholinguistics* 28, pp. 661–677 (cit. on p. 8).

Surridge, Marie (1985). "Le genre grammatical des composés en français." In: *Canadian Journal of Linguistics* 30.3, pp. 247–271 (cit. on p. 2).

— (1986). "Genre grammatical et dérivation lexicale en français." In: *Canadian Journal of Linguistics* 31.3, pp. 267–283 (cit. on p. 2).

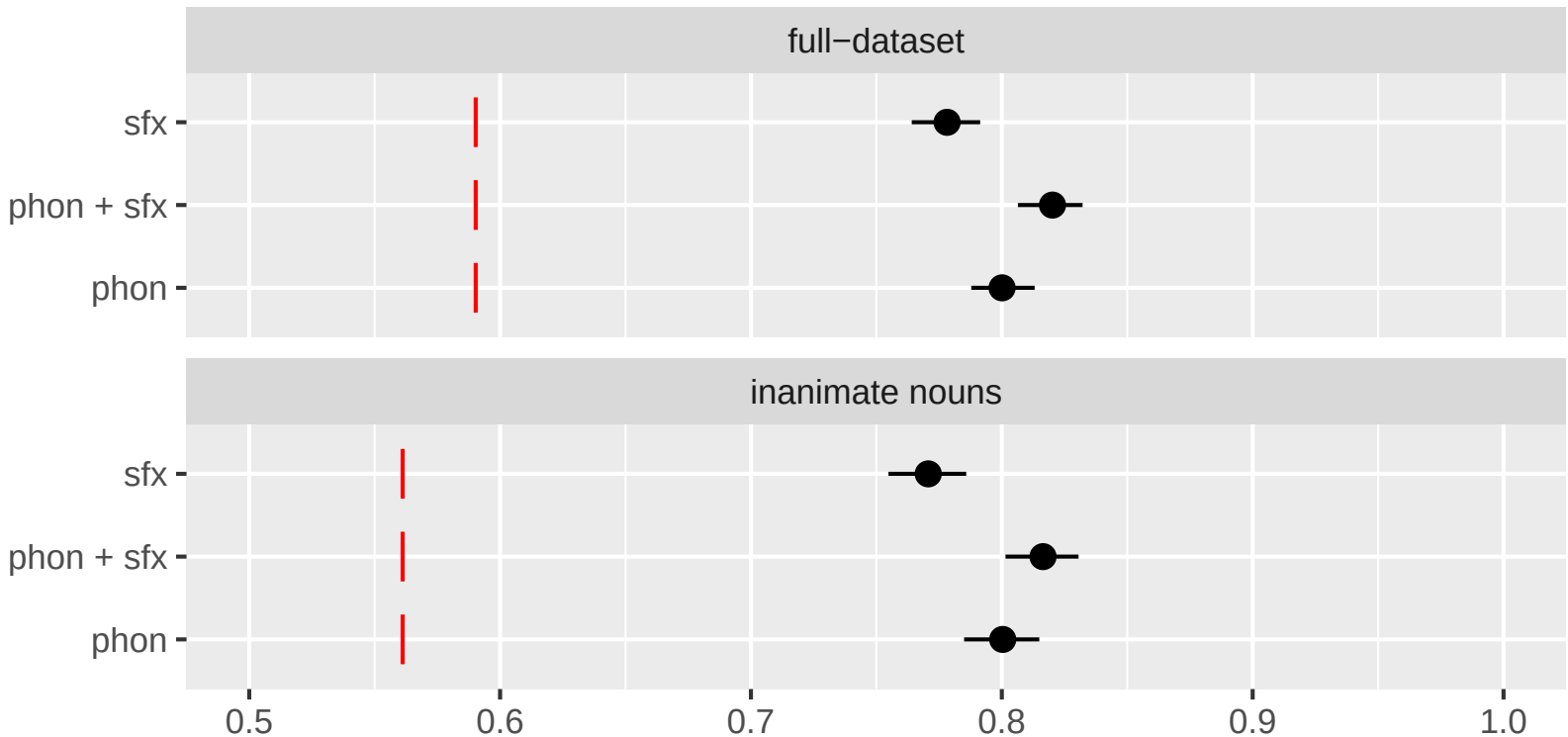
— (1989). "Le facteur sémantique dans l'attibution du genre en français." In: *Canadian Journal of Linguistics* 34.1, pp. 19–44 (cit. on p. 2).

Tucker, G Richard, Wallace E Lambert, and André Rigault (1977). *The French speaker's skill with grammatical gender: An example of rule-governed behavior*. Berlin: Mouton De Gruyter (cit. on pp. 2, 3).

Williams, Adina et al. (accepted). "Adjective Selection and the Grammatical Gender of Inanimate Nouns: An Information-Theoretic Investigation of Neo-Whorfian Claims." In: *Transactions of the Association for Computational Linguistics* (cit. on p. 5).

# A role for animacy?

- Because we were uncertain about the quality of our data for human nouns, we re-ran all models on the subset of inanimate nouns.
- Although the numbers are obviously not exactly the same, we observe no qualitative difference. Our conclusions still hold.

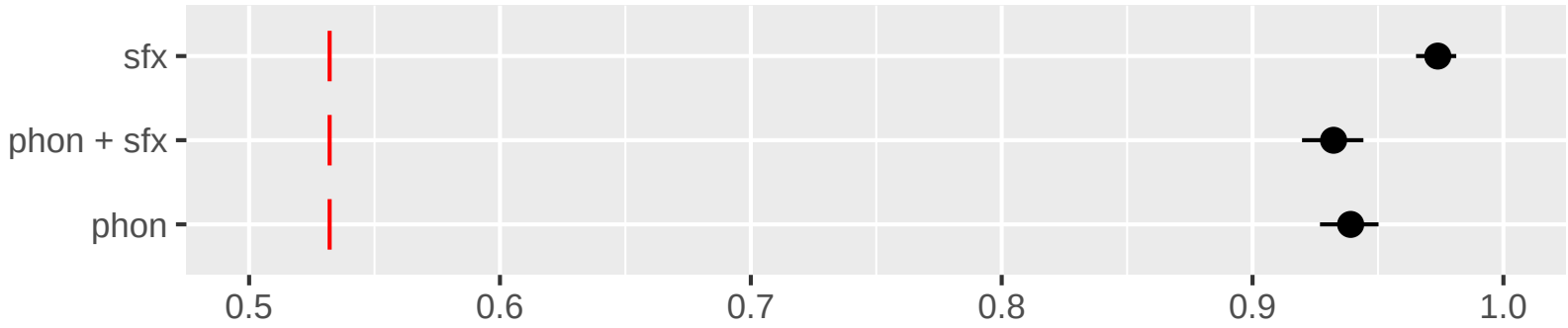


# One or two suffixes? I

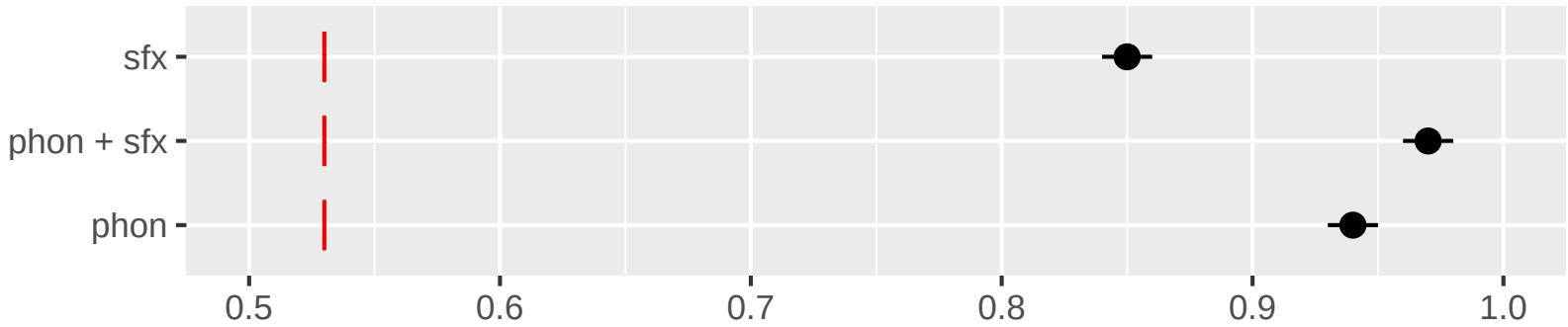
- We took the decision to treat pairs such as *-eur* vs. *-euse*, *-ier* vs. *-ière* as separate suffixes.
- An alternative view would hold that such pairs should count as variants of the same suffix, and hence that the contrast between e.g. *-eur* and *-euse* should not participate in prediction of gender from morphology.

# One or two suffixes? II

- Taking such an approach obviously reduces the predictive power of morphological information.
  - Results on suffixed nouns, without pairing of gendered suffixes




  - Results on suffixed nouns, with pairing of gendered suffixes



# One or two suffixes? III

- Rationale for our decision:
  1. Groupings are a lot less obvious than one would think: often there is actually homophony between one member of a gendered pair and a gender-specific suffix.
     - *froussard* 'coward' (< *frousse* 'fear') and *veinarde* 'lucky person' (< *veine* 'luck') are obviously similar → gender contrasting pair of suffixes
     - *bombarde* 'bombard' (< *bombe* 'bomb') has no masculine analogue → separate suffix?
  2. If we start grouping suffixes with (mostly) the same meaning, why stop at such gendered pairs? Why not e.g. group *-age* and *-ment*?
  3. Sequences such as *-euse*, *-rice*, are morphologically unsegmentable units (Bonami and Boyé, 2005). But it does not make sense to say that *-eur* and *-euse* are allomorphs, since they need not convey the same meaning (see animates). Hence paired suffixes have a dubious ontological status.
  4. It does not make sense to compare whole word phonology to abstract morphology. A fair comparison would look at pairs of suffixes vs. stem phonology. But that is a completely different endeavor.