Assessing the import of morphology in gender assignment: A computational study of French

Nazanin Shafiabadi Nina Nusbaumer Olivier Bonami

Université Paris Cité, LLF, CNRS

LLcD 2024

Gender Assignment

- Far from arbitrary (Corbett, 1991).

Gender Assignment

- Far from arbitrary (Corbett, 1991). Some predictors:
  - **semantic**: *femme*$_{\mathrm{F}}$ (woman), *homme*$_{\mathrm{M}}$ (man)

Gender Assignment

- Far from arbitrary (Corbett, 1991). Some predictors:
  - **semantic**: *femme*$_{\mathrm{F}}$ (woman), *homme*$_{\mathrm{M}}$ (man)
  - **phonological**:[paj]$_{\mathrm{F}}$ (straw), [aksɑ̃]$_{\mathrm{M}}$ (accent)

Gender Assignment

- Far from arbitrary (Corbett, 1991). Some predictors:
  - **semantic**: *femme*$_{\mathrm{F}}$ (woman), *homme*$_{\mathrm{M}}$ (man)
  - **phonological**:[paj]$_{\mathrm{F}}$ (straw), [aksɑ̃]$_{\mathrm{M}}$ (accent)
  - **morphological**: *chanteuse*$_{\mathrm{F}}$ (singerF), *chanteur*$_{\mathrm{M}}$ (singerM)

Gender Assignment

- Far from arbitrary (Corbett, 1991). Some predictors:
  - **semantic**: *femme*$_{\mathrm{F}}$ (woman), *homme*$_{\mathrm{M}}$ (man)
  - **phonological**:[paj]$_{\mathrm{F}}$ (straw), [aksɑ̃]$_{\mathrm{M}}$ (accent)
  - **morphological**: *chanteuse*$_{\mathrm{F}}$ (singerF), *chanteur*$_{\mathrm{M}}$ (singerM)

- **Entanglement** of these predictors:
  Is the gender of the noun *chanteuse*$_{\mathrm{F}}$ predictable because of its semantic properties or because of the suffix *-euse*?

Gender Assignment

- Far from arbitrary (Corbett, 1991). Some predictors:
  - **semantic**: *femme*$_{\mathrm{F}}$ (woman), *homme*$_{\mathrm{M}}$ (man)
  - **phonological**:[paj]$_{\mathrm{F}}$ (straw), [aksɑ̃]$_{\mathrm{M}}$ (accent)
  - **morphological**: *chanteuse*$_{\mathrm{F}}$ (singerF), *chanteur*$_{\mathrm{M}}$ (singerM)

- **Entanglement** of these predictors:
  Is the gender of the noun *chanteuse*$_{\mathrm{F}}$ predictable because of its semantic properties or because of the suffix *-euse*?

- Predicting gender in French:
  - 2-gender system: FEM, MASC
  - 1/3 of nouns with a unique gender end in a derivational suffix (Bonami et al., 2019)
  - The relevant suffixes are for the most part compatible with only one gender (Bonami & Boyé, 2019)

Gender Assignment

- Far from arbitrary (Corbett, 1991). Some predictors:
  - **semantic**: *femme*$_{\mathrm{F}}$ (woman), *homme*$_{\mathrm{M}}$ (man)
  - **phonological**:[paj]$_{\mathrm{F}}$ (straw), [aksɑ̃]$_{\mathrm{M}}$ (accent)
  - **morphological**: *chanteuse*$_{\mathrm{F}}$ (singerF), *chanteur*$_{\mathrm{M}}$ (singerM)

- **Entanglement** of these predictors:
  Is the gender of the noun *chanteuse*$_{\mathrm{F}}$ predictable because of its semantic properties or because of the suffix *-euse*?

- Predicting gender in French:
  - 2-gender system: FEM, MASC
  - 1/3 of nouns with a unique gender end in a derivational suffix (Bonami et al., 2019)
  - The relevant suffixes are for the most part compatible with only one gender (Bonami & Boyé, 2019)
- **Is gender prediction based on morphology or phonology?**

Previous Attempts

- Since the 2000s, various attempts to use machine learning techniques to study gender assignment.
- General idea:
  - Set up a (computational) classifier predicting gender from a set of other features.
  - Train the classifier on part of the data.
  - The accuracy of the classifier on unseen data reflects how hard it is to predict gender from that set of features.
- Some examples:

Study	Language	Genders	Datapoints	Method
Sokolik & Smith (1992)	French	2	600	Simple NN
Eddington (2002)	Spanish	2	2,416	Skousen's AML
Matthews (2005)	French	2	1,590	Perceptron / AML
Plaster et al. (2013)	Tsez	4	3,500	Decision Trees
Bonami et al. (2019)	French	2	3,683	Multilayer perceptron
Fedden et al. (in press)	German	3	30,576	Boosting Trees

Our Study

- Focus on predicting the gender of French nouns from their orthographic or phonemic form.
- Investigation of how the presence of suffixes influences prediction in the absence of explicit morphological boundaries.
- Primary data source: ***Flexique*** (Bonami et al., 2014):
  - French inflectional lexicon
  - Using only nouns (total: 29,411)

Gender	**Distribution**	
MASC	56%	17,192
FEM	39%	12,219
Both	5%	1,593

- For a subset of the nouns, morphological annotation from ***Echantinom*** (Bonami & Tribout, 2021) is available, and will be used for evaluation.

Data Preparation

- **Dataset**: French nouns with gender labels
  - **Training set**: 80% (random split) of ***Flexique*** (F), excluding all nouns that appear in ***Echantinom*** (E) and nouns with both genders (b)
  - **Validation set**: The remaining 20% of ***Flexique***
  - **Testing set**: ***Echantinom***, excluding all nouns with both genders

Set	Noun Count	Composition
Training	19,904	$(F - E - b) \times 80\%$
Validation	4,976	$(F - E - b) \times 20\%$
Test	4,532	$(E - b)$

- **Preprocessing**: Word forms reversed and split into characters

Model Selection

- **Model**: GenderLSTM (LSTM architecture)
- **Reason**:
  - Effective for handling sequential data
  - Generates a prediction after each character is processed
  - Relatively simple architecture

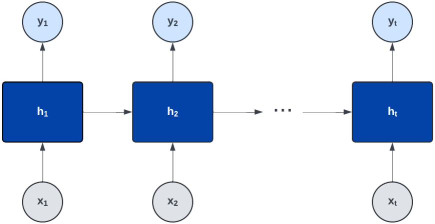


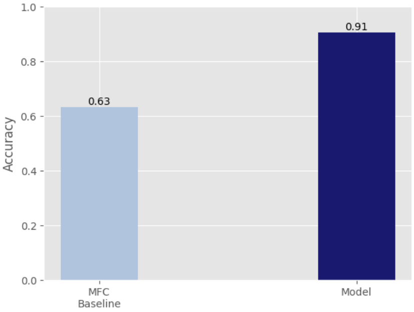
Figure: Unidirectional LSTM Architecture

Model Architecture

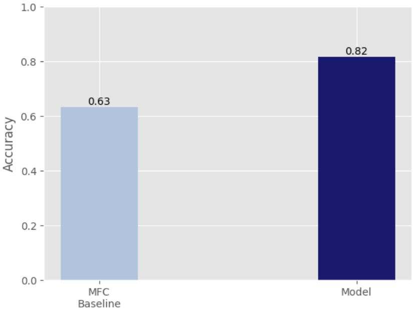
- **Embedding Layer**: Converts characters into vectors
- **LSTM Layer**: Processes character sequences
- **Linear Layer**: Transforms LSTM outputs into logits for class predictions
- **Output Layer**: Outputs probability distributions over the different genders for each character and the final hidden state for loss computation

Comparison with MFC Baseline

- The chosen architecture is successful in capturing the gender patterns
- Gender is more predictable from the orthographic form of the nouns than from the phonemic form



(a) Orthographic Model



(b) Phonemic Model

Comparing the Models' Performances

- Analyzed the predictions for **4,532** lemmas (Test set)

Comparing the Models' Performances

- Analyzed the predictions for **4,532** lemmas (Test set)

	FEM	MASC
Correct	87% (1438)	93% (2664)
False	13% (220)	7% (209)

Orthographic model accuracy

Comparing the Models' Performances

- Analyzed the predictions for **4,532** lemmas (Test set)

	FEM	MASC
Correct	87% (1438)	93% (2664)
False	13% (220)	7% (209)

Orthographic model accuracy

	FEM	MASC
Correct	70% (1167)	89% (2549)
False	30% (492)	11% (323)

Phonemic model accuracy

Comparing the Models' Performances

- Analyzed the predictions for **4,532** lemmas (Test set)

	FEM	MASC
Correct	87% (1438)	93% (2664)
False	13% (220)	7% (209)

Orthographic model accuracy

	FEM	MASC
Correct	70% (1167)	89% (2549)
False	30% (492)	11% (323)

Phonemic model accuracy

- They agree for **84%** of the lemmas.

What are frequent errors common to both models?

- Errors common to the orthographic and phonemic models.
- Testing using ***Echantinom***'s morphological annotations.
- Hypothesis:
  - The model learns from statistical associations between endings and gender.
  - These associations are mostly due to the presence of gender assigning suffixes.
  - Errors are driven by words that go against these associations.

What are frequent errors common to both models?

- Errors common to the orthographic and phonemic models.
- Testing using ***Echantinom***'s morphological annotations.
- Hypothesis:
  - The model learns from statistical associations between endings and gender.
  - These associations are mostly due to the presence of gender assigning suffixes.
  - Errors are driven by words that go against these associations.

Process	**Distribution** in the dataset	**Proportion** among errors	**Examples**
Simplex	43% (1,943)	66% (159)	*gruyère, crypte*
Suffix	37% (1,701)	11% (27)	*capitaine, fureur, minceur*
Nonconcat	2% (96)	7% (17)	*info, mayo, berlingue*
Conversion	10% (473)	7% (17)	*velouté, uniforme, hongre?*
Polylexical	5% (240)	6% (15)	*serre-tête, bain-marie, croquemitaine*
Prefix	2% (81)	3% (7)	*mi-juillet, antigène, demi-frère*

Some Common Simplicia Endings

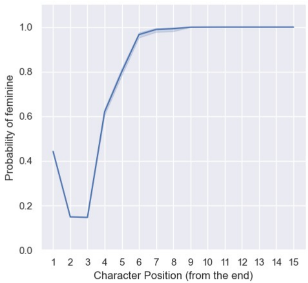
Ending	**Distribution** in the dataset		**Distribution** among errors		**Examples**
-age		(895)		(3)	*cage, rage, nage*
-ère		(469)		(3)	*frère, gruyère, cimetière*
-ion		(1791)		(5)	*scion, bastion, centurion*
-eur		(1552)		(4)	*fleur, rumeur, couleur*
-ule		(155)		(3)	M: *bidule*; F: *capsule, mandibule*
-ane		(113)		(4)	M: *crane, arcane*, F: *chicane*
-che		(255)		(3)	*prêche, guinche, panache*
-sse		(284)		(3)	*causse, bidasse, carrosse*
-oge		(9)		(2)	*toge, loge*

FEM MASC

Observations I

- Words with the same suffix have similar curves

(a) Orthographic Model



(b) Phonemic Model

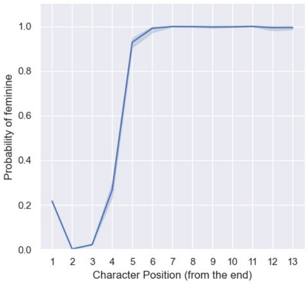
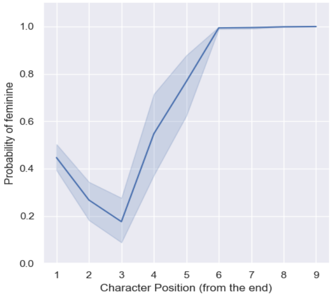


Figure: Average probability distribution of words with suffix ”ion” (162 samples)

Observations II

- There is greater variation within runs for the same word than runs for different words sharing the same suffix

(a) ”gradation”



(b) ”sensation”

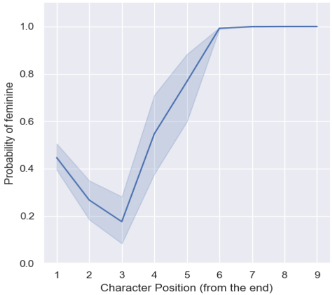
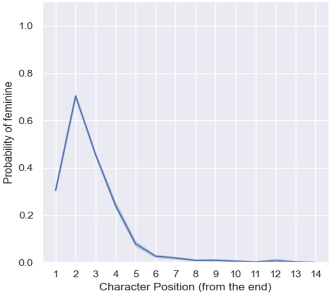


Figure: Comparison of the probability distribution of two words ending in ”ion” over 10 runs.

Interesting Cases I

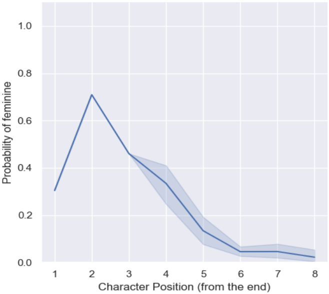
- Two suffixes with the same form but different genders and meanings have strikingly similar shapes.

(a) ”eurM” (193 samples)



e.g. *chauffeur* (driver)

(b) ”eurF” (15 samples)

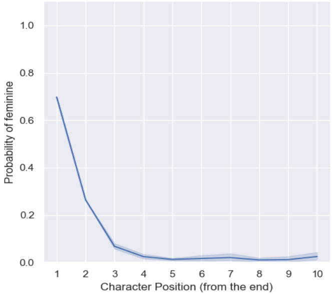


e.g. *fureur* (fury)

Figure: Probability distribution of masculine vs. feminine words ending in ”eur”.

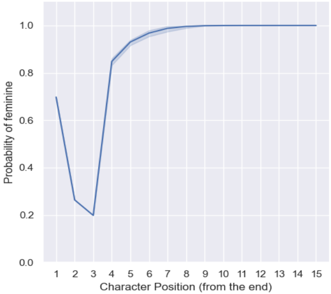
Interesting Cases II

- With similar but not identical suffixes, the shapes diverge.



(a) ”on” (72 samples)

e.g. *carton* (cardboard)



(b) ”ion” (162 samples)

e.g. *sensation* (sensation)

Figure: Probability distribution of words ending in ”on” vs. ”ion”.

Conclusions

- An RNN assigns gender correctly with high accuracy based only on the form of the noun (confirming previous work).
- Error patterns are driven by the presence of endings with strong statistical associations to one gender due to derivational morphology
- Analysis of the time course of predictions suggests that the model assigns gender based on derivational suffixes
  - The curves are the same for words with the same suffix
- Relation to Word and Paradigm?
  - Other talks in this workshop focus on prediction of one word from another word.
  - Another important question is prediction of the content of a word from its form (Beniamine 2018).
  - Here we have looked at how one morphosyntactic feature, gender, can be predicted from the form of nouns.
  - The same methods could be deployed more generally in the context of paradigms.

Thank You!

References I

Bonami, Olivier & Gilles Boyé (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, & Andrew Hippisley. Edinburgh: Edinburgh University Press, 171–192 (cit. on pp. 2–8).

Bonami, Olivier, Gauthier Caron, & Clément Plancq (2014). "Construction d'un lexique flexionnel phonétisé libre du français." In: *Actes du quatrième Congrès Mondial de Linguistique Française*. Ed. by Franck Neveu et al., pp. 2583–2596 (cit. on p. 10).

Bonami, Olivier, Matías Guzmán Naranjo, & Delphine Tribout (2019). "The role of morphology in gender assignment in French." In: *ISMO 2019*. Paris, France (cit. on pp. 2–9).

Bonami, Olivier & Delphine Tribout (2021). *Echantinom: Morphological Annotations for French*. `https://osf.io/rdxqk/`. Accessed: 2024-07-16 (cit. on p. 10).

Corbett, Greville G. (1991). *Gender*. Cambridge: Cambridge University Press (cit. on pp. 2–8).

Eddington, David (2002). "Spanish gender assignment in an analogical framework." In: *Journal of Quantitative Linguistics* 9, pp. 49–75 (cit. on p. 9).

References II

Fedden, Sebastian, Matías Guzmán Naranjo, & Greville G. Corbett (in press). "Typology meets statistical modeling: the German gender system." In: *Language* (cit. on p. 9).

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 (cit. on p. 9).

Plaster, Keith, Maria Polinsky, & Boris Harizanov (2013). "Noun classes grow on trees: Noun classification in the North-East Caucasus." In: *Language Typology and Historical Contingency. In honor of Johanna Nichols*. Ed. by Balthasar Bickel et al. Amsterdam: John Benjamins, pp. 153–169 (cit. on p. 9).

Sokolik, M. E. & Michael E. Smith (1992). "Assignment of gender to French nouns in primary and secondary language: a connectionist model." In: *Second Language Research* 8.1, pp. 39–58 (cit. on p. 9).