Phonolette, un phonologiseur automatique du français

Basilio Calderone, Nabil Hathout

CLLE, CNRS & Université Toulouse Jean Jaurès

Thématiques actuelles de la recherche en TAL
6 novembre 2023

La phonologisation automatique

Données

Phonolette

Résultats et évaluation

Conclusion

La phonologisation automatique

La phonologisation automatique vise à produire une séquence de /phonèmes/ qui transcrit la prononciation d'une séquence de <graphèmes> (*grapheme-to-phoneme*, G2P).

Par exemple en français :

- un mot comme <**ami**> se prononce $\longrightarrow$ /**ami**/
- un mot comme <**pari**> se prononce $\longrightarrow$ /**pari**/
- un mot comme <**pâté**> se prononce $\longrightarrow$ /**pate**/

Les chevrons <> délimitent une transcription orthographique

Les barres obliques // délimitent une transcription phonologique

Usages des transcriptions phonologiques

Nous avons besoin de transcriptions phonologiques fiables pour :

- la création de lexiques et de ressources lexicales
  - Phonolette doit servir à transcrire les formes contenues dans la table des lexèmes de la base de données morphologique Demonette-2 (Hathout & Namer, 2014; Namer et al., 2019; Fiammetta et al., 2023)
- les protocoles et les expériences psycholinguistiques
- les études phonologiques (études en synchronie)
- les études sur l'acquisition de la compétence phonologique chez l'enfant
- les tâches de phonologie computationnelles fondées sur le *machine learning*
- etc.

Transparence orthographique

Un facteur qui détermine la complexité de la phonologisation automatique est la **transparence orthographique** (TO).
La **TO** désigne, pour une langue donnée, le degré de correspondance entre l'orthographe et la phonologie

En français :

- **<solitaire>** se prononce $\longrightarrow$ **/sɔlitɛʁ/**
  - la séquence <ai> $\mapsto$ ɛ
- **<phonologie>** se prononce $\longrightarrow$ **/fɔnɔlɔʒi/**
  - la séquence <ph> $\mapsto$ f
- **<psychologie>** se prononce $\longrightarrow$ **/psikɔlɔʒi/**
- **<achat>** se prononce $\longrightarrow$ **/aʃa/**
- **<schéma>** se prononce $\longrightarrow$ **/ʃema/**
  - la séquence <ch> $\mapsto$ /k/ ou /ʃ/
    la séquence /ʃ/ $\mapsto$ <sch>

Transparence orthographique

Une langue est parfaitement orthographiquement transparente si la correspondance entre graphèmes et phonèmes est consistante : à un phonème correspond un et un seul graphème.

Le finnois présente une transparence ‘pure’ entre l’orthographe et la phonologie

- en finnois, le mot **<kinkku>** ‘jambon’ se prononce $\longrightarrow$ **/kinkku/**

L’italien a également une bonne transparence orthographique

- en italien, le mot **<vecchio>** ‘vieux’ se prononce $\longrightarrow$ **/vekkjo/** (<cch> $\mapsto$ /kk/)

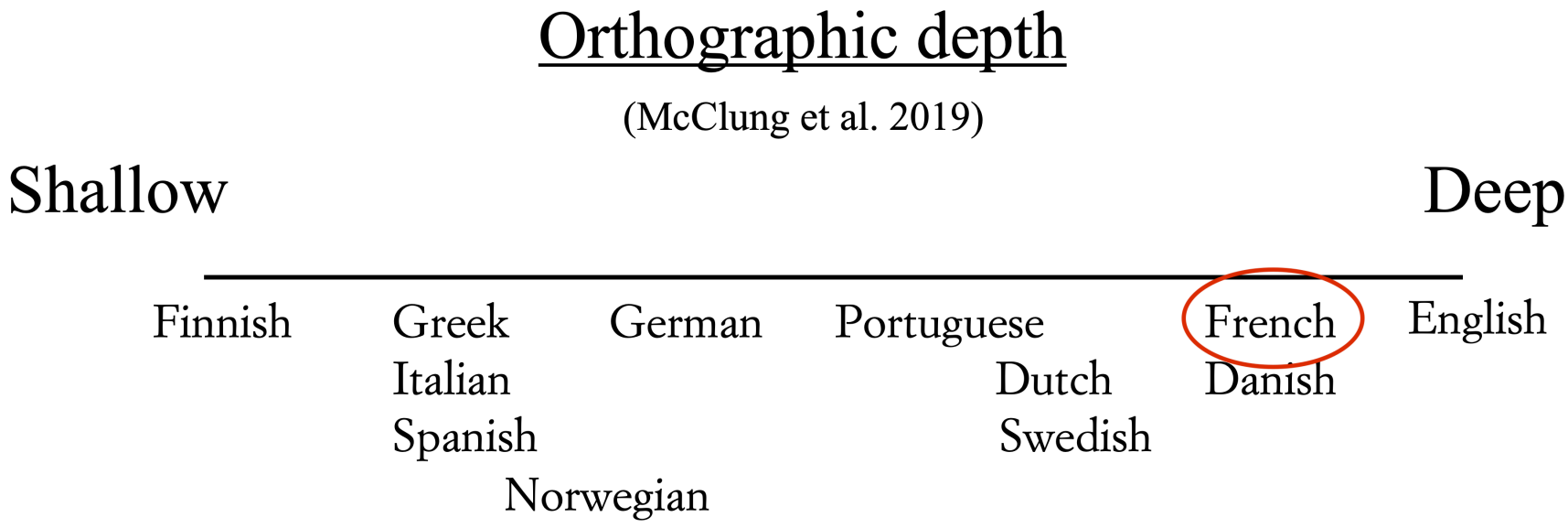
Transparence orthographique

À l'inverse, l'orthographe de l'anglais est opaque (on parle aussi d'orthographe profonde (*orthographic depth*)).

- le mot <**tolerate**> 'tolérer' se prononce ⟶ /**tɒlərɹeɪt**ʔ/
- le son /f/ peut être correspondre à <**f**ast>, <a**ff**air>, <**ph**ase>, <tou**gh**>, <ha**lf**>

Transparence orthographique

Les langues n'ont pas toutes la même transparence orthographique :



A gauche les langues les plus transparentes, correspondance 1 : 1, à droite les langues les plus opaques (correspondance N : 1, avec $N \gg 1$).

Graphotactique

Les emprunts et le mots étrangers qui ont une orthographe non standard sont une autre source de difficultés pour la phonologisation automatique.

Le mot <edelweiss>, d'origine allemande, perturbe les systèmes de phonologisation automatique car il ne respecte pas l'ordre habituel des caractères en français (<l> n'est normalement pas suivi par <w>).

Le domaine qui étudie les combinaisons de caractères valides dans une langue donnée est appelé **graphotactique**.

Graphotactique

Score de conformité graphotactique

Les **probabilités de transition** entre un caractère et le suivant permet de calculer un score de conformité graphotactique.

- Le **score de conformité graphotactique** est la moyenne géométrique des probabilités conditionnelles des n-grammes de caractères qui apparaissent dans un mot graphémique.
- La **probabilité conditionnelle** d'un n-grammes est la probabilité que le dernier caractère du n-gramme apparaissent après la séquence des $n-1$ caractères précédents.

Graphotactique

- La probabilité d'apparition du caractère <r> après la séquence <#pa> en début de mot est élevée car de nombreux mots commençant par <#par> en français : <parc>, <parasol>, <parent>, <pari>, <parfait>, <parti>, etc.
  $\Rightarrow$ <#pa><r> a une probabilité de transition élevée
- La séquence <#ig><l> a une probabilité de transition très faible elle n'apparaît que dans <igloo>

Le **score de conformité graphotactique** (SCG) du mot <parc> est calculé au moyen de la formule (*3*-grammes):

$$SCG_{3g}(<\#parc\$>) = \sqrt[4]{P(a|\#p) * P(r|pa) * P(c|ar) * P(\$|rc)}$$

$SCG_{3g}(<\#edelweiss\$>) = 0.01$ $SCG_{3g}(<\#parc\$>) = 0.15$

$\Rightarrow$ Le score de conformité graphotactique permet de
- détecter les mots qui ne sont pas conformes à l'orthographe du français.
- exclure les mots qui dégradent la qualité des prédictions du système.

Phonolette

Phonolette est un système *grapheme-to-phoneme* du français qui sera utilisé pour produire les transcriptions phonologiques des formes fléchies de la table des lexèmes de Démonette.

Phonolette est un réseau de neurones seq2seq entraîné sur un jeu de données qui combine les formes graphémiques de GLàFF (Sajous et al., 2013) et les transcripitons phonologiques de Flexique (Bonami et al., 2014).

Exactitude

Phonolette prédit correctement 97.82% des transcriptions phonologiques du jeu de données.

La prédiction est peu sensible à la **graphotactique**.
L'exactitude augmente à **98.11%** sur un jeu de données réduit aux mots les plus conformes à la graphotactique du français.

GLàFF

Phonolette a été développé pour transcrire les formes fléchies des entrées de la table des lexèmes de **Démonette-2** (Fiammetta et al., 2023)

Le lexique utilisé pour constituer la table des lexèmes de Demonette-2 est le lexique flexionnel **GLàFF** (Sajous et al., 2013).

- GLàFF contient 186 082 lemmes, **plus de 1.4 millions d'entrées**.
- GLàFF a été créé à partir de la nomenclature de GLAWI Sajous & Hathout (2015).
- GLàFF contient des mots lexicaux des catégories Nom (N) ; Verbe (V) ; Adjectif (A) ; Adverbe (R)
- GLàFF fourni des transcriptions phonologiques pour **88.85%** des entrées.

GLàFF

Qualité des annotations

✔ Les graphies des lemmes et des formes fléchies sont très fiables.

✗ La qualité des transcriptions phonologiques est **insuffisante** pour être incluses dans une ressource de référence comme Démonette-2

Les transcriptions phonologiques présentent des problèmes de cohérence et de standardisation.
Il s'agit d'une ressource créée par de simples locuteurs et non par des spécialistes.

Forme	**Traits morphosynt.**	**Lemme**	**Transcription phonologique**
chanlatté	Vmps-sm	chanlatter	chɑ̃late ; ʃɑ̃late
sauvageons	Ncmp	sauvageon	sauvageɔ̃ ; sovaʒɔ̃̃
brûleries	Ncfp	brûlerie	bʁylʁi ; bʁyləʁi
autiste	Ncfs	autiste	otist ; ɔtist
autocollant	Afpms	autocollant	otokɔlɑ̃
ainée	Ncfs	ainée	ene
ainées	Ncfp	ainée	ene ; ɛne

La variation phonologique

Il peut exister plusieurs prononciations équivalentes pour un même mot en fonction de la norme phonologique de référence et des variables diatopiques (français standard *vs* français régional parlé).

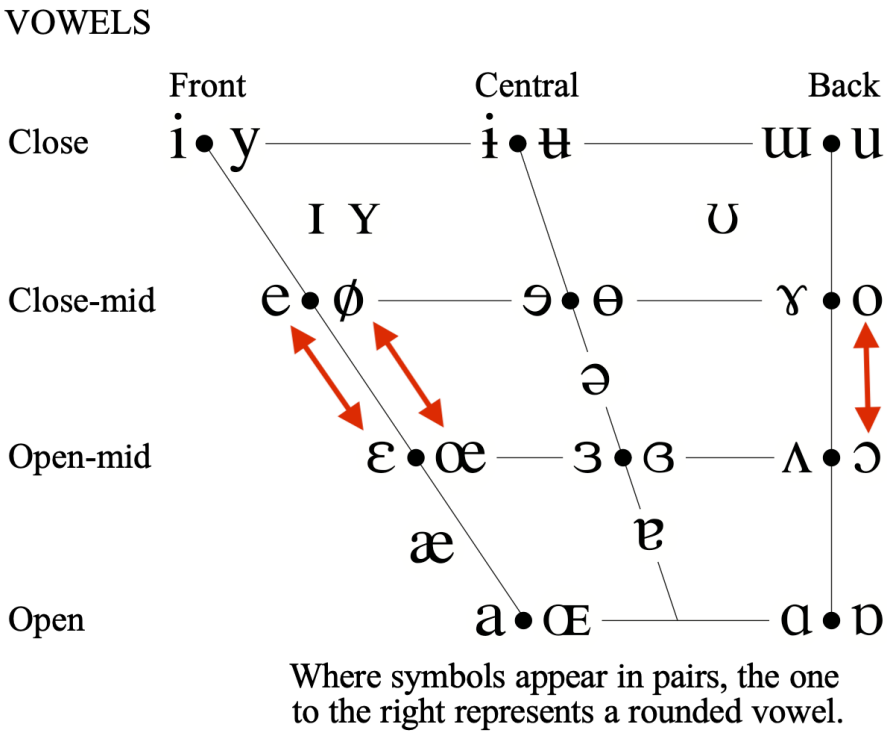
En français, cette variation concerne principalement trois groupes de voyelles sur la base du trait **voyelle fermée** vs. **voyelle ouverte**.

- La voyelle e vs. la voyelle ɛ
  <essayais> ↦ /esejɛ/ vs. /esɛjɛ/
- La voyelle ø vs. la voyelle œ
  <affeuré> ↦ /aføʁe/ vs. /afœʁe/
- La voyelle o vs. la voyelle ɔ
  <aumônières> ↦ /omonjɛʁ/ vs. /ɔmɔnjɛʁ/

Neutraliser la variation phonologique

L'**Alphabet Phonétique International (API)** est un standard pour la transcription phonétique de toutes les langues du monde. Il est utilisé pour la transcription phonologique des sons du langage parlé.

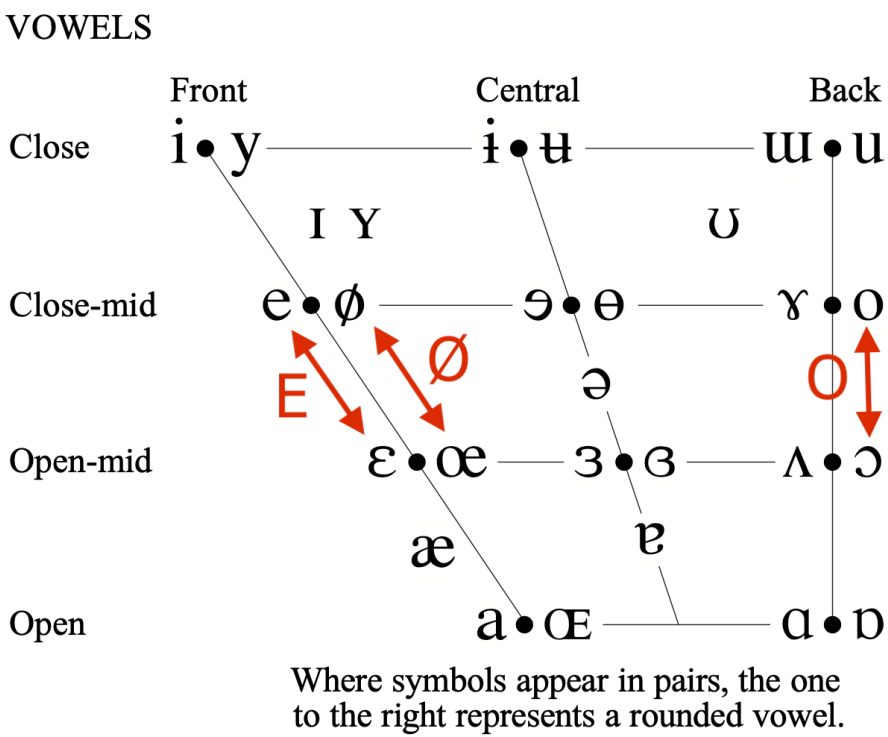
Table des voyelles en API :



Neutraliser la variation phonologique

Une solution à ce problème de variation phonologique est d'annuler (i.e. neutraliser) la différence entre ces couples de voyelles.

Table dans laquelle les variations vocaliques sont neutralisées :



<essayais> ↦ /esEjɛ/
<affeuré> ↦ /afØʁe/
<aumônières> ↦ /OmOnjɛʁ/

Nota bene

La neutralisation n'a lieu que dans contextes phonologiques très spécifiques. Il ne s'agit pas d'une substitution de phonèmes sytématique.
<litote> ↦ /litɔt/
<litière> ↦ /litjɛʁ/

Neutraliser la variation phonologique

La neutralisation des voyelles permet de représenter de façon globale et systématique les multiples réalisations phonologiques des formes du lexique.
Les neutralisations **ne font pas partie de l'API**.

Flexique et ses transcriptions phonologiques

Cette neutralisation des voyelles est utilisée dans le lexique **Flexique**.

Flexique (Bonami et al., 2014) est un lexique phonologique flexionnel du français.

- **Flexique** contient 47 242 lemmes et 363 293 formes fléchies munies de leur transcriptions phonologiques.
- Les entrées appartiennent aux catégories majeures :
  Nom (N) ; Verbe (V) ; Adjectif (Adj)

Neutraliser la variation phonologique

- ✔ Ses annotations sont très fiables. Elles ont fait l'objet d'un examen manuel rigoureux.
- ✗ **Flexique** ne fournit les transcriptions orthographiques des formes fléchies

Combiner GLÀFF et Flexique

- **GLÀFF** dispose d'une couverture large et de transcriptions orthographiques fiables.
- **Flexique** dispose de transcriptions phonologiques fiables.

Nous avons construit un jeu de données en croisant les transcriptions orthographiques de **GLÀFF** et les transcriptions phonologiques de **Flexique**.

La jointure entre les deux lexiques est réalisée sur les lemmes et les étiquettes morphosyntaxiques des formes fléchies.
Les étiquettes morphosyntaxiques de Flexique ont été recodées dans le jeu d'étiquette de GLàFF.

Jeu de donnée complet

Dataset complet

La jointure produit un jeu de données de 362 260 entrées.
Chaque forme fléchie est munie

- d'une transcription orthographique (qui provient de GLÀFF),
- une transcription phonologique (qui provient de Flexique),
- une catégorie grammaticale (POS).

transc. ortho (GLàFF)	**transc. phono (Flexique)**	**POS**
linguistique	lɛ̃gɥistik	N
rectangulaire	ʁɛktɑ̃gylɛʁ	A
régénériez	ʁeʒeneʁje	V

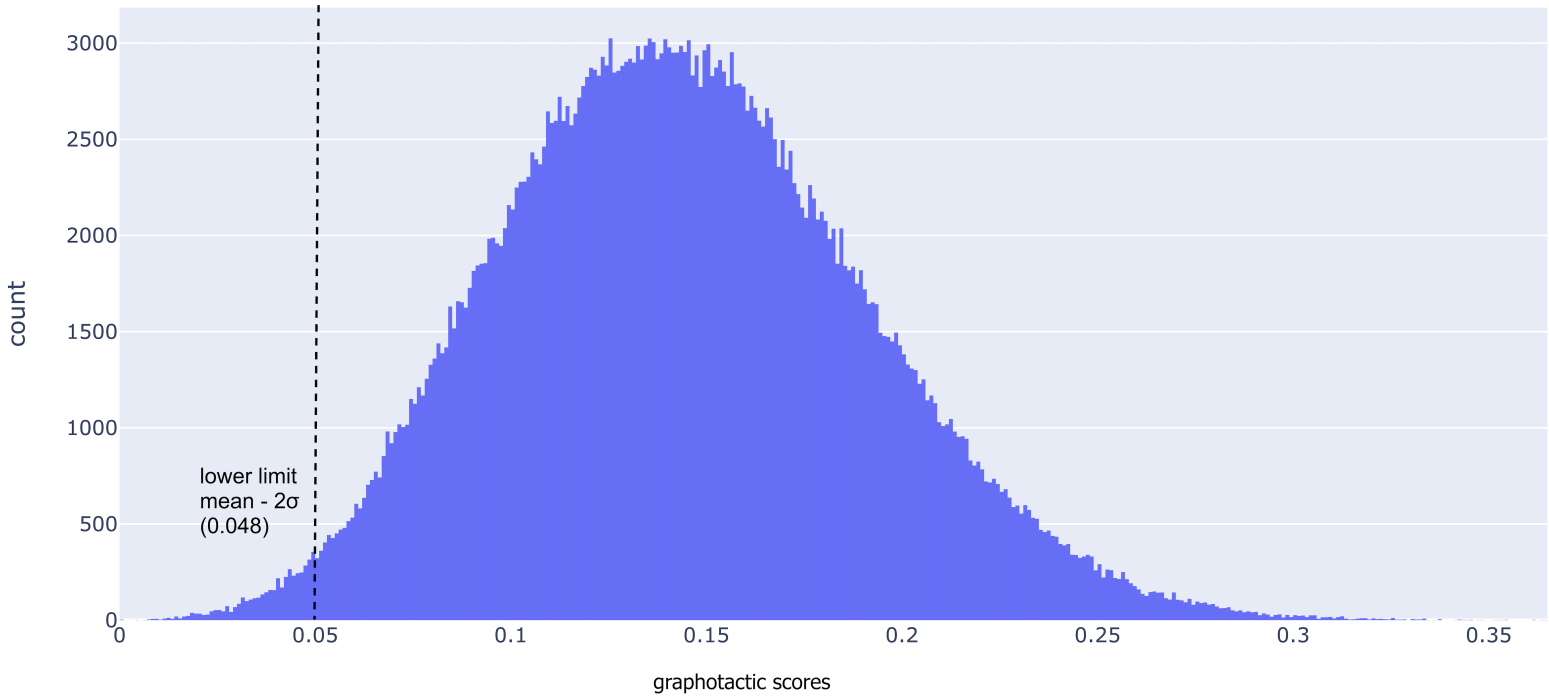
Jeu de données graphotactique

Nous avons également créé un deuxième jeu de données en sélectionnant les formes écrites les plus conformes à la **graphotactique** française.

Nous avons calculé le **score de conformité graphotactique** de chacune des 362 260 entrées du dataset complet (sur la base de la transcription orthographique).

Un score élevé indique que le mot est composé de séquences bien attestées dans la langue ;
un score faible indique la présence de séquences plus rares.

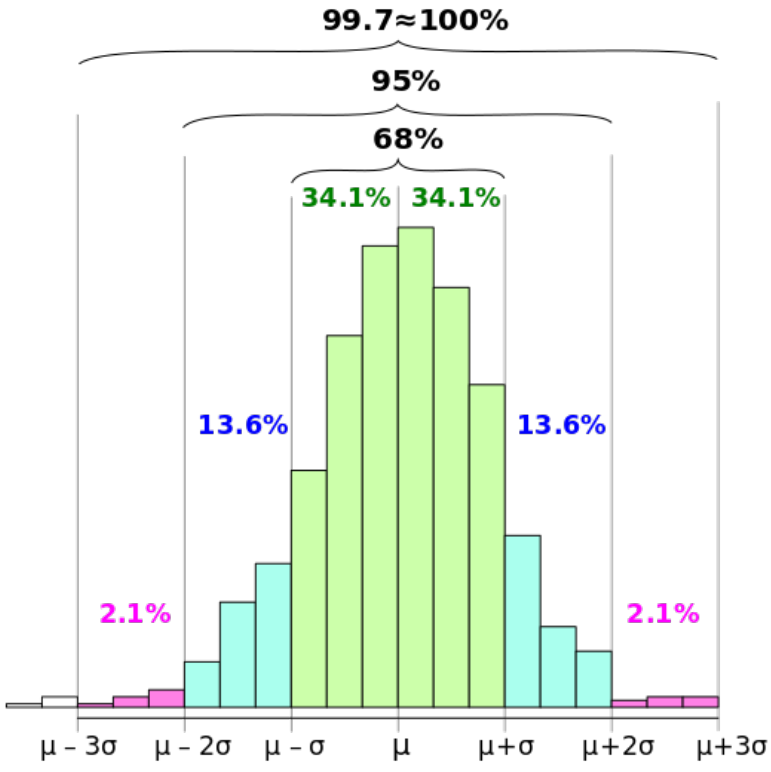
Jeu de données graphotactique



Les scores ont une distribution (quasi-)normale.

Dans une distribution normale presque toutes les valeurs se situent dans un intervalle centré autour de la moyenne et délimité par trois écarts-types de part et d'autre.
C'est la règle dite des « trois sigmas » ou « règle empirique ».

La règle empirique



La règle empirique stipule que si une variable est normalement distribuée,

- environ 68% de la distribution se situe à l'intérieur d'un écart-type de la moyenne
- 95% de la distribution est à moins de deux écarts-types de la moyenne
- 99,7% de la distribution est à moins de trois écarts types de la moyenne.

Jeu de données graphotactique

La distribution quasi-normale permet d'exclure les formes les moins conformes à la graphotactique française en fixant un seuil au bord gauche
La valeur du seuil est fixée à 0.048, soit la moyenne moins deux écarts-types.
Cette sélection élimine produit 5 810 mots.
Elle produit un jeu de données graphotactique qui contient donc 356 450 entrées.

Exemples de mots exclus

edelweiss, gecko, tsé-tsé, guppy, tsar, jenny, sexy, geisha, kevlar, tsunami, breitschwanz, one-step, watt

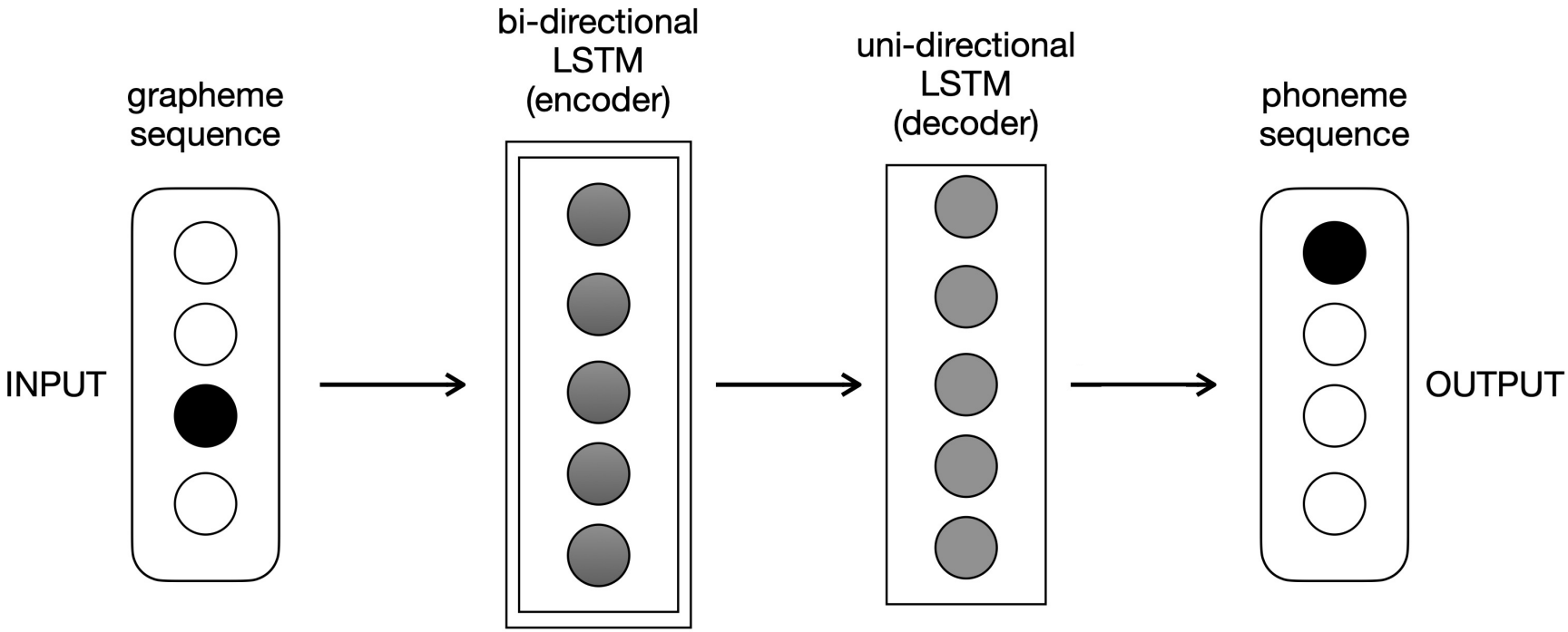
Phonolette

Phonolette est basée sur une architecture de réseau à mémoire à long terme (LSTM, Hochreiter & Schmidhuber (1997)). Les LSTM sont des réseaux conçus pour le traitement des séquences (*seq2seq*).
Ils peuvent être entraînés à prédire une séquence de phonèmes (**output**) à partir d'une séquence de lettres (**input**).

Architecture

- Un encodeur LSTM bidirectionnel et un décodeur LSTM unidirectionnel de 100 neurones chacun
- L'encodeur lit la séquence d'entrée dans les deux sens (de gauche à droite et de droite à gauche) et produit une matrice d'activation de dimension 200 (100 dimensions pour chaque direction).
- La matrice est transmise au décodeur qui produit une transcription en phonèmes de l'entrée

Phonolette


grapheme
sequence
bi-directional
LSTM
(encoder)
uni-directional
LSTM
(decoder)
phoneme
sequence
INPUT
OUTPUT

Données d'entraînement

Les entrées sont les formes orthographiques de **GLÀFF**.

Les sorties transcriptions phonologiques de **Flexique** correspondantes.

La qualité des prédictions est améliorée significativement lorsque les catégories grammaticales (POS) sont fournies en entrée.

Finale <ent>

Dans un nom, la séquence finale <ent> correspond généralement au phonème /ɑ̃/ (<dent> $\mapsto$ /dɑ̃/).
Dans un verbe, elle correspond à la séquence vide (<glacent> $\mapsto$ /glas/).

POS + transc. ortho (input)	**transc. phono (output)**
#N#linguistique$	#lɛ̃gɥistik$
#A#rectangulaire$	#ʁEktɑ̃gylEʁ$
#V#régénériez$	#ʁEʒEnEʁjE$

Implémentation

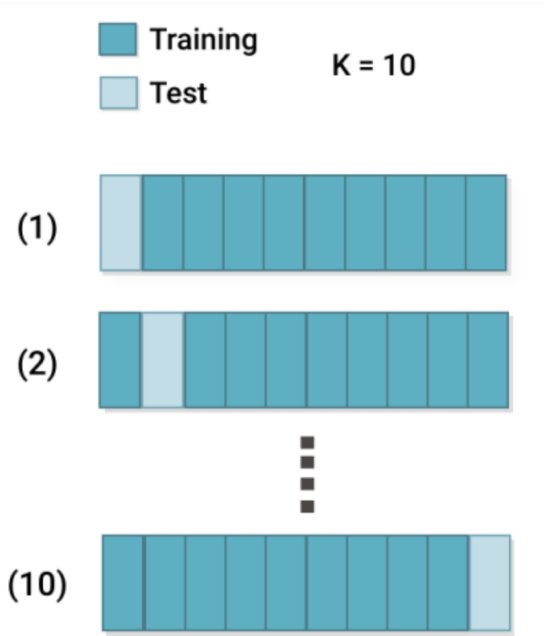
Codage one-hot. Les 44 lettres (graphèmes), les 3 caractères qui codent le POS et les 45 phonèmes sont tous représentés par des vecteurs **one-hot**. Codage d'une variable à n états sur n bits dont un seul prend la valeur 1 et où tous les autres valent 0.

Fonction de perte = entropie croisée catégorielle. (*Categorical Cross Entropy*).

Batch = 32. Lors de l'apprentissage, 32 échantillons sont traités avant la mise à jour des paramètres du modèle.

Implémentation

10-fold cross validation. L'apprentissage a été effectuée en utilisant la validation croisée à 10 "plis". Le jeu de données est divisé en 10 sous-ensembles appelés "plis". Le système utilise successivement l'un des plis comme données d'évaluation (*test set*) et les 9 autres comme données d'entraînement (*training set*).



WER et PER

Deux mesures permettent d' évaluer les performances d'un système G2P (Gorman et al., 2020) :

- le **WER** (*word error rate*) est le pourcentage de mots dont la transcription est incorrecte.
- le **PER**(*phone error rate*) est la somme des distances d'édition de Levenshtein entre les transcriptions prédites et les transcriptions de référence divisée par la somme des longueurs des transcriptions de référence :

$$PER = 100 \times \frac{\sum_{i=1}^{n} d(s_i, r_i)}{\sum_{i=1}^{n} |r_i|}$$

PER est exprimé en pourcentage.

WER et PER

	WER	PER
Dataset complet	2.18%	0.55%
Dataset graphotactique	1.89%	0.48%

Les performances de Phonolette sont légèrement meilleures sur le jeu de données graphotactique (en WER et en PER).
$\Rightarrow$ Les mots dont la graphotactique n'est pas française sont plus difficile à prédire.

Les valeurs de PER sont faibles.
Les erreurs concernent globalement 1 phonème (1 prédiction) sur 200.

Qualité de la prédiction par POS

Dataset	POS	correct	%	incorrect	%	total
complet	A	33 163	96.19	1310	3.81	34 473
	N	48 124	90.31	5159	9.69	53 283
	V	273 091	99.48	1413	0.52	274 504
	total	354 378	97.82	7882	2.18	362 260
graphotactique	A	32 289	96.20	1275	3.80	33 564
	N	45 060	91.63	4115	8.37	49 175
	V	272 386	99.51	1325	0.49	273 711
	total	349 735	98.11	6715	1.89	356 450

- La transcription phonémique des **verbes** est presque parfaite pour les deux jeu de données.
- L'exactitude de la prédiction pour les **adjectifs** est presque identique pour les deux jeux de données
- La transcription des **noms** est la plus difficile.
  Les formes dont la graphotactique n'est pas française (les emprunts) sont essentiellement des noms.

Qualité de la prédiction par POS

La distribution des POS explique les différences dans la qualité des prédictions.

les **verbes** (270 000 formes) sont la classe la plus importante et la plus redondante en termes d'observations auxquelles Phonolette est exposées. Chaque verbe a 51 formes fléchies.

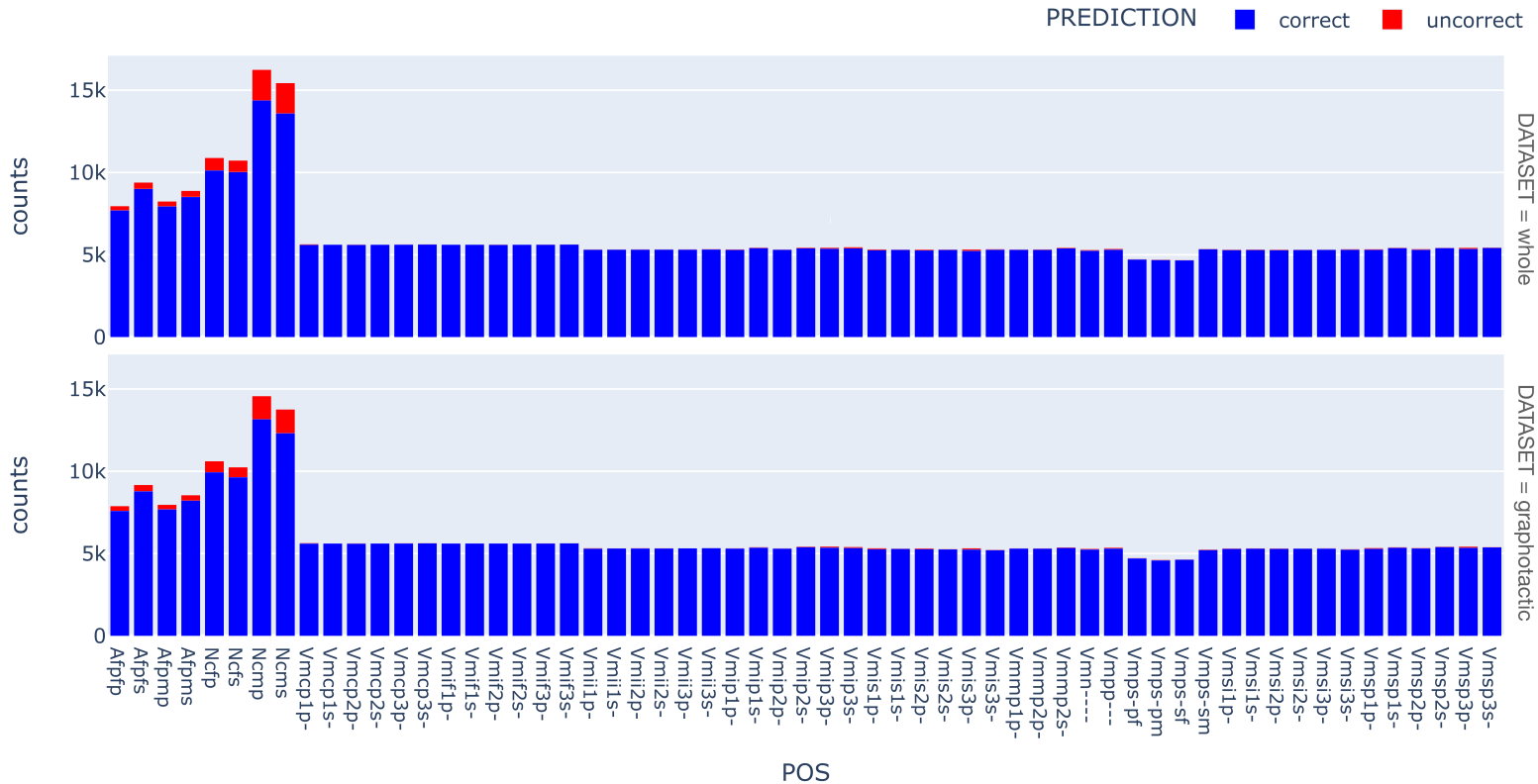
Les **noms** et les **adjectifs** ne représentent ensemble que 25% des données d'apprentissage.

Les adjectifs ont 4 formes, et les noms seulement 2 formes.

L'exactitude semble dépendre du nombre de lemmes et de la taille du paradigme flexionnel.

Qualité de la prédiction par POS

Prédictions correctes et incorrectes pour chaque cellule du paradigme pour les trois catégories V, N et A.



Erreurs dans les transcriptions des noms

Une partie des erreurs pour les noms (et les adjectifs) sont en réalité des variantes :

- insertion de consonnes finales : <persil> ↦ /pɛʁsi/ (/pɛʁsil/)
- voyelles incorrectes : <emmenthal> ↦ /Emɛ̃tal/ (/Emɑ̃tal/)
- voyelles incorrectes : <hautin> ↦ /Otɛ̃/ (/otɛ̃/)

D'autres sont de vraies erreurs :

- suppression de consonnes finales : <audit> ↦ /Odi/ (/Odit/)
- mots composés : <contre-emploi> ↦ /kɔ̃tʁəplE/ (/kɔ̃tʁɑ̃plwa/)
- mots étrangers : <shinto> ↦ /ʃĩto/ (/ʃinto/)

Erreurs dans la transcription des verbes

La finale <ient> se prononce /jɛ̃/ à la 3^{e} personne du singulier <appartient>, <convient>, <obtient>, etc.

Le modèle n'a pas accès à la totalité de l'étiquette morphosyntaxique **Vmip3s-**, mais seulement à la catégorie. Au pluriel, <ent$> correspond à la séquence vide.

Les séquences prédites sont : apaʁti, kɔ̃vi, ɔpti

Erreurs dans la transcription

Les erreurs les plus fréquentes de Phonolette sur le dataset complet.

Oper.	Neutr.	Phon.	%	∑%	Prédiction	Target	Forme
sub.		O/o	3.12	3.12	EgzOsfɛʁ	Egzosfɛʁ	exosphère
sub.	oui	E/ɛ	3.09	6.21	sEgmɑt̪al	sɛgmɑt̪al	segmentale
ins		s	2.98	9.19	tʁɑšibEʁjɛ̃	tʁɑs̃sibEʁjɛ̃	transsibérien
sub.	oui	ɛ/E	2.39	11.58	ɛd̪iʒɛstjɔ̃	ɛd̪iʒEstjɔ̃	indigestions
sub.	oui	O/ɔ	1.99	13.57	Optik	ɔptik	optiques
ins.		t	1.61	15.18	kasjɔ̃	katjɔ̃	cations
sub.		ɑ7ɛn	1.6	16.78	kamEʁamɑ̃	kamEʁamɛn	cameramen
suppr.		t	1.26	18.05	ʁEpit	ʁEpi	répits
sub.		z/s	1.24	19.29	ʁEzEksjɔ̃	ʁEsEksjɔ̃	résections
sub.		o/O	1.11	20.4	povʁət̪e	pOvʁət̪e	pauvreté

- La plupart des erreurs concernent la neutralisation des voyelles.
- La prononciation ou la non-prononciation de l'occlusive /t/ à la fin d'un mot.
- Le remplacement de la fricative voisée /z/ par la fricative devoisée /s/.

PER

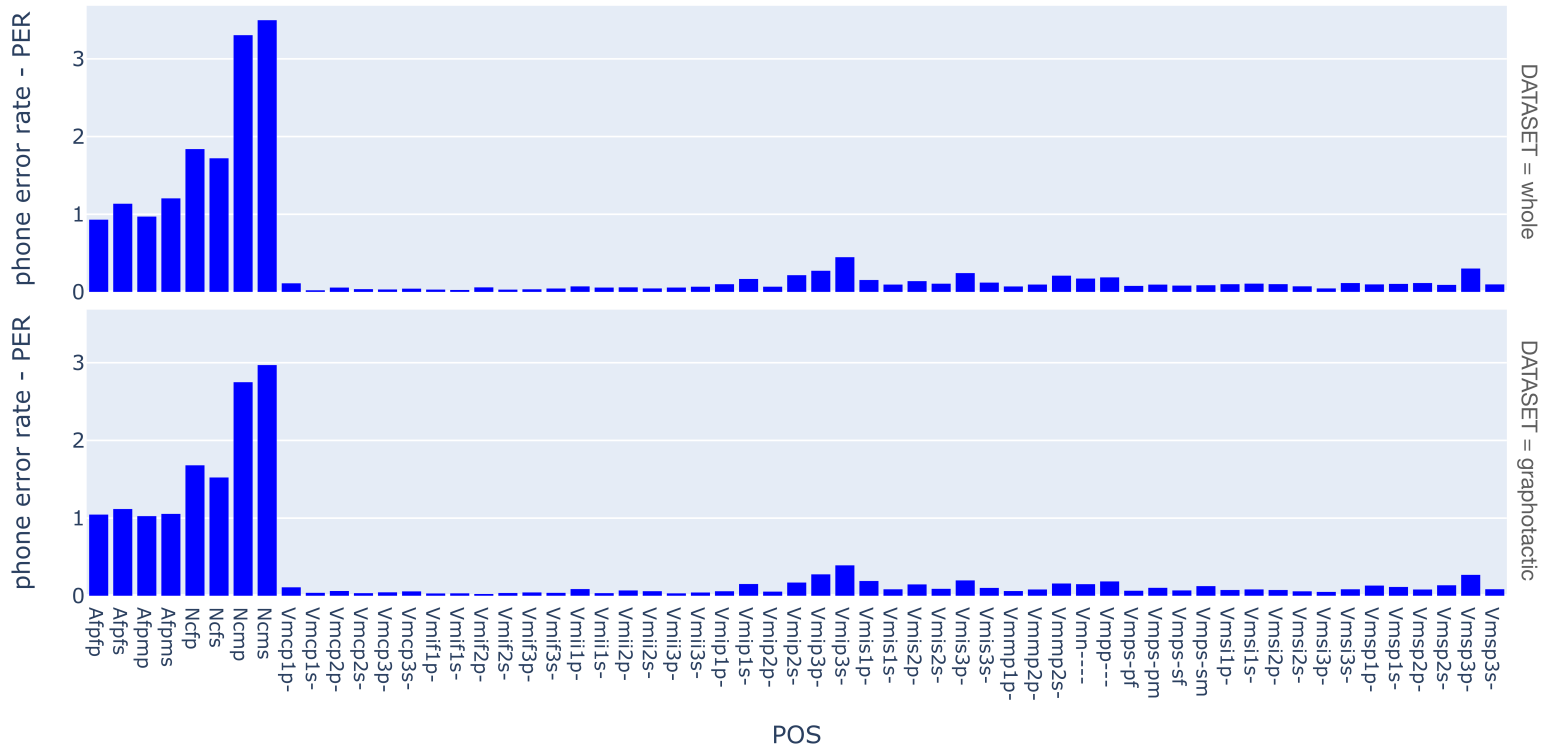
Dataset	POS	PER
complet	A	1.06%
	N	2.70%
	V	0.10%
graphotactique	A	1.06%
	N	2.29%
	V	0.09%

Les différences dans la difficulté des prédictions apparaît aussi dans le PER.

- Les **noms** représentent la catégorie la plus difficile à phonologiser.
- Pas de différence entre les deux jeux de données pour les **adjectifs** et les **verbes**.

PER

PER pour chaque cellule des paradigmes des V, N et A.



Ouvrir la boîte

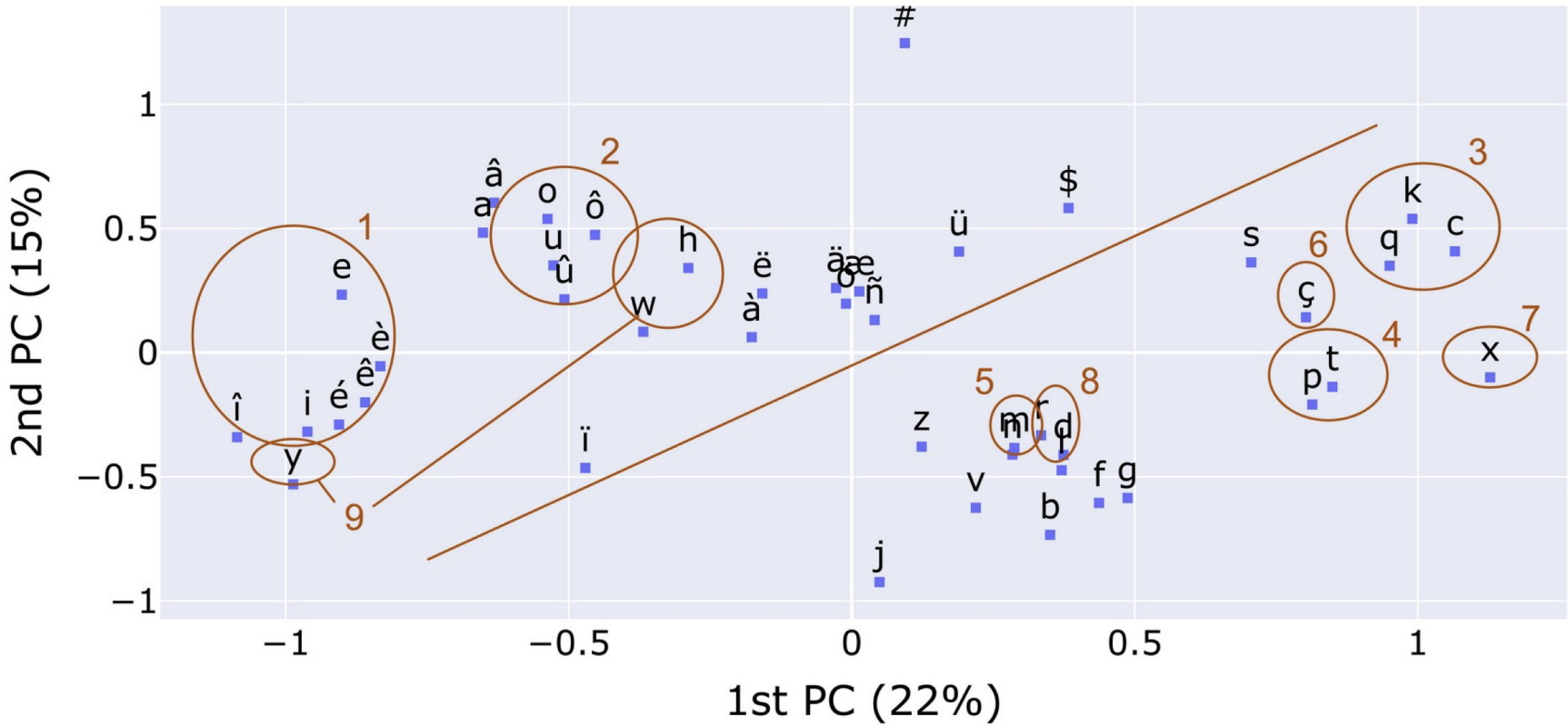
Nous avons vu que Phonolette transcrit des séquences de graphèmes en séquences de phonèmes en utilisant une architecture LSTM encodeur-décodeur/

Plus précisément, à la fin de la phase d'apprentissage, l'encodeur produit une représentation temporelle de la séquence de graphèmes d'entrée. Cette représentation est ensuite transmise au décodeur, qui l'utilise pour produire une séquence phonologique en sortie

Les états d'activation peuvent être récupérés pour toutes les formes écrites du lexique, mais aussi pour les 44 caractères individuels (c'est-à-dire les unigrammes) qui composent les formes

Les états d'activation des 44 caractères forment une **matrice** de 44 lignes et 200 colonnes (correspondant aux 100 dimensions du codeur pour chacune des deux directions).

Ouvrir la boîte



- distinction voyelles et consonnes
- voyelles antérieures (groupe 1) et voyelles postérieures (groupe 2)
- groupe occlusifs (3 et 4) et groupe nasal (5)
- groupe liquides <r> et <l> (8)

Conclusion

Nous avons présenté Phonolette, un modèle pour la phonologisation de la langue française à partir de son orthographe

Bonnes performances :

- $97,82$ **%** de prédictions correctes pour le jeu de données complet
- 98.11 **%** pour le jeu de données graphotactique

Directions futures possibles :

- codage des données phonologiquement motivé (non one-hot)
- introduction de la syllabe
- utilisation de Phonolette pour d'autres langues

Références

Bonami, Olivier., Caron Gauthier & Clément Plancq. 2014. Construction d'un lexique flexionnel phonétisé libre du français. *4e Congrès Mondial de Linguistique Française* 2583–2596. doi:https://doi.org/10.1051/shsconf/20140801223.

Fiammetta, Namer, Nabil Hathout, Dany Amiot, Lucie Barque, Olivier Bonami, Gilles Boyé, Basilio Calderone, Julie Cattini, Georgette Dal, Alexander Delaporte, Guillaume Duboisdindien, Achille Falaise, Natalia Grabar, Pauline Haas, Frédérique Henry, Mathilde Huguin, Nyoman Juniarta, Loïc Liégeois, Stéphanie Lignon, Lucie Macchi, Grigoriy Manucharian, Caroline Masson, Fabio Montermini, Nadejda Okinina, Franck Sajous, Daniele Sanacore, Mai Thi Tran, Juliette Thuilier, Yannick Toussaint & Delphine Tribout. 2023. Démonette-2, a derivational database for french with broad lexical coverage and fine-grained morphological descriptions. *Lexique* 33. à paraître.

Références

Gorman, Kyle, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu & Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th sigmorphon workshop on computational research in phonetics, phonology, and morphology*, 40–50. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.sigmorphon-1.2. https://aclanthology.org/2020.sigmorphon-1.2.

Hathout, Nabil & Fiammetta Namer. 2014. Démonette, a french derivational morpho-semantic network. *Linguistic Issues in Language Technology* 11(5). 125–168.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8). 1735–1780.

Namer, Fiammetta, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout & Delphine Tribout. 2019. Demonette2 – une base de données dérivationnelles du français à grande échelle : premiers résultats. In *Actes de la 26e conférence annuelle sur le traitement automatique des langues naturelles (taln-2019)*, 233–243. Toulouse.

Références

Sajous, Franck & Nabil Hathout. 2015. GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*, 405–426. Herstmonceux, England.

Sajous, Franck, Nabil Hathout & Basilio Calderone. 2013. GLÁFF, un Gros Lexique á tout Faire du Français. In *Actes de la 20e conférence sur le traitement automatique des langues naturelles (taln'2013)*, 285–298. Les Sables d'Olonne, France.