Received: March 2023 / Accepted: June 2023
Published on line: December 2023

Phonolette: a grapheme-to-phoneme converter for French

Phonolette : un convertisseur graphème-phonème du français

Basilio Calderone
CNRS, CLLE, Université de Toulouse
basilio.calderone@univ-tlse2.fr

Nabil Hathout
CNRS, CLLE, Université de Toulouse
nabil.hathout@univ-tlse2.fr

Olivier Bonami
Université de Paris Cité, LLF, CNRS
olivier.bonami@u-paris.fr

Résumé

L'article présente Phonolette, un phonologiseur du français, capable de prédire une transcription phonologique d'un mot à partir de sa représentation orthographique. Phonolette est basé sur une architecture LSTM bidirectionnelle. Le protocole d'apprentissage de Phonolette combine les formes orthographiques du lexique GLÀFF et les transcriptions phonologiques de Flexique. Deux jeux de données ont été utilisés : l'intersection des entrées de GLÀFF et de Flexique ; la même intersection, mais en conservant seulement les formes qui ont une graphotactique française. Les résultats obtenus sont prometteurs. La précision est de 97,82 % sur le jeu complet et de 98,11 % sur le jeu réduit.

Mots-clés : phonologiseur, transcription phonémique, LSTM, ressources lexicales, français

Abstract

This paper presents Phonolette, a phonologizer for the French language, capable of predicting a phonological transcription of a word from its orthographic representation. Phonolette is based on a bidirectional LSTM architecture. The training protocol of Phonolette combines orthographic forms from the GLÀFF lexicon and phonological transcriptions from Flexique. Two datasets were used: the intersection of GLÀFF and Flexique entries; and the same

intersection, but keeping only those forms with French graphotactics. The results are promising. Accuracy is 97.82% for the full dataset and 98.11% for the reduced dataset.

Keywords: phonologizer, phonemic transcription, LSTM, lexical resources, French

Funding: The Demonext project has been funded by the French National Research Agency (ANR), under the reference ANR-17-CE23-0005. The description of the project, its results and publications are available at https://www.demonext.xyz/

1. Introduction

Automatic phonologization aims to produce a sequence of phonemes that transcribe the pronunciation of a sequence of graphemes. This is a classical speech processing task for which several benchmarks exist (van Esch et al., 2016; Yolchuyeva et al., 2019; Lee et al., 2020). Work and systems on grapheme-phoneme (G2P) conversion have mostly focused on English. More recently, systems have been developed for other languages, especially poorly endowed languages (Gorman et al., 2020). The availability of lexicons containing reliable phonological transcriptions of written forms is essential for linguistic research, especially in quantitative linguistics and experimental psycholinguistics.

In this paper, we present the French G2P system Phonolette. This system predicts phonological transcriptions from written forms. Phonolette is based on a bidirectional LSTM network processing sequences. Model training relies on two French inflectional lexicons, Flexique (Bonami et al., 2014), a lexicon whose phonemic transcriptions are normalized and carefully checked, and GLÀFF (Sajous et al., 2013), a very large lexicon extracted from the GLAWI electronic dictionary (Sajous and Hathout, 2015; Sajous et al., 2020), which is derived from the entries documenting French words in the French version of Wiktionary, so-called Wiktionnaire.

More specifically, Phonolette is trained on a dataset that combines the graphemic forms of GLÀFF and the phonemic transcriptions provided by Flexique. A second dataset was created by eliminating the forms that do not fully comply with French graphotactics (these are usually loans). The results of Phonolette are promising, reaching an accuracy of 97.82% for the full dataset and 98.11% for the second dataset.

The paper presents the architecture of the model and its performance according to different metrics for nouns, verbs and adjectives.

2. Datasets

Phonolette's reference lexicon is Flexique (Bonami et al., 2014). Flexique provides reliable phonological transcriptions that have undergone rigorous manual review.[1] It uses specific coding for certain vowels. E is a representation that neutralizes the differences between the vowels /e/ and /ɛ/, O between the vowels /o/ and /ɔ/ and Ø between the vowels /ø/ and /œ/.

This coding allows to represent the multiple phonological realizations of some forms without having to arbitrarily choose one of them. Each entry has a unique transcription as close as possible to its surface form from which the different possible realizations can be deduced.

Furthermore, the vowel /ə/ is systematically included in the transcriptions even in words where its actual realization is rare, except at the end of words, where it is systematically omitted because realization is predictable without lexical knowledge (Dell, 1995). Flexique has three tables, one for each of the categories noun, verb and adjective. It contains 47,242 lemmas and the transcriptions of 363,293 inflected forms, but does not provide their written forms. It therefore provides the output data of Phonolette, but not the input data. We extracted this information, i.e. the written forms of the inflected forms, from another French lexicon, the GLÀFF lexicon (Sajous et al., 2013), which has a very large coverage with 186 082 lexemes and over 1.4 million inflected forms. GLÀFF provides the lemma of each inflected form, a morphosyntactic label and, for 90% of them, one or more phonemic transcriptions. While GLÀFF's written forms are reliable, its phonemic transcriptions are not standardized and often inconsistent. In order to exploit the respective strengths of GLÀFF and Flexique, we built a dataset with written transcriptions from the former and phonemic transcriptions from the latter. The two resources are joined on the lemmas and morphosyntactic labels of the inflected forms. This dataset contains 362,260 entries.

We also created a second dataset by selecting only the written forms that conform to French graphotactics (or orthotactics). The first dataset will be referred to as the 'full dataset' and the second as the 'graphotactic dataset'.

The term graphotactics refers to the permissible sequences of characters in a given language, just as phonotactics refers to the permissible combinations of phonemes. It can be observed in graphical representations by using the transition probability of the characters they contain. In our case, this probability is the same as a conditional probability. For example, the probability of having a character <r> after a sequence <#pa> at the beginning of a word is high in French because the lexicon contains many words beginning with <#par> (*parc* 'park', *parasol* 'umbrella', *parent* 'parent', *pari* 'bet', *parfait* 'perfect', *parti* 'gone', *parcourir* 'to

[1] Flexique does not contain proper names and toponyms, and this inevitably represents a limitation in this study.

search', *pardonner* 'to forgive', *parfumer* 'to perfume', etc.). As a result, <#par> has a high transition probability. Conversely, a sequence like <#igl> has a very low transition probability, because the probability of having an <l> after the <#ig> sequence at the beginning of the word is very low. It is only found in French in the word *igloo* 'igloo', a loan word from Inuktitut.

Graphotactics is clearly a factor in sequence learning. Sequences with high transition probabilities are learned better (and lead to more correct predictions) than those with low transition probabilities, because of the systematic redundancy of these sequences in the lexicon. However, we do not know how significant this is.

The graphotactic acceptability of a word can be estimated by the geometric mean of the transition probabilities of the trigrams it contains. This measure, which we call the *graphotactic score*, is available in a lexicon such as PsychoGLÀFF, a resource dedicated to psycholinguistic studies (Calderone et al., 2014). For example, the score for the word <parc> is given by the following formula:

$$score_{3gram}(parc) = \sqrt[4]{P(a|STARTp) * P(r|pa) * P(c|ar) * P(END|rc)}$$

where START and END represent the beginning and the end of the word.
A high graphotactic score indicates that a word is composed of sequences well attested in the lexicon; a low score indicates the presence of rarer sequences.

Figure 1 shows the distribution of graphotactic scores in the full dataset. We can see that the scores have a quasi-normal distribution, in which almost all values lie in an interval centered around the mean and bounded by three standard deviations on either side (the so-called three-sigma rule or empirical rule).

We can therefore set a threshold at the left edge of the curve to exclude words that do not conform to French graphotactics. The value is 0.048, i.e. the mean minus two standard deviations. This minimum threshold for the graphotactic score excludes 5,810 words from the full dataset. Our second dataset therefore contains 356 450 entries.

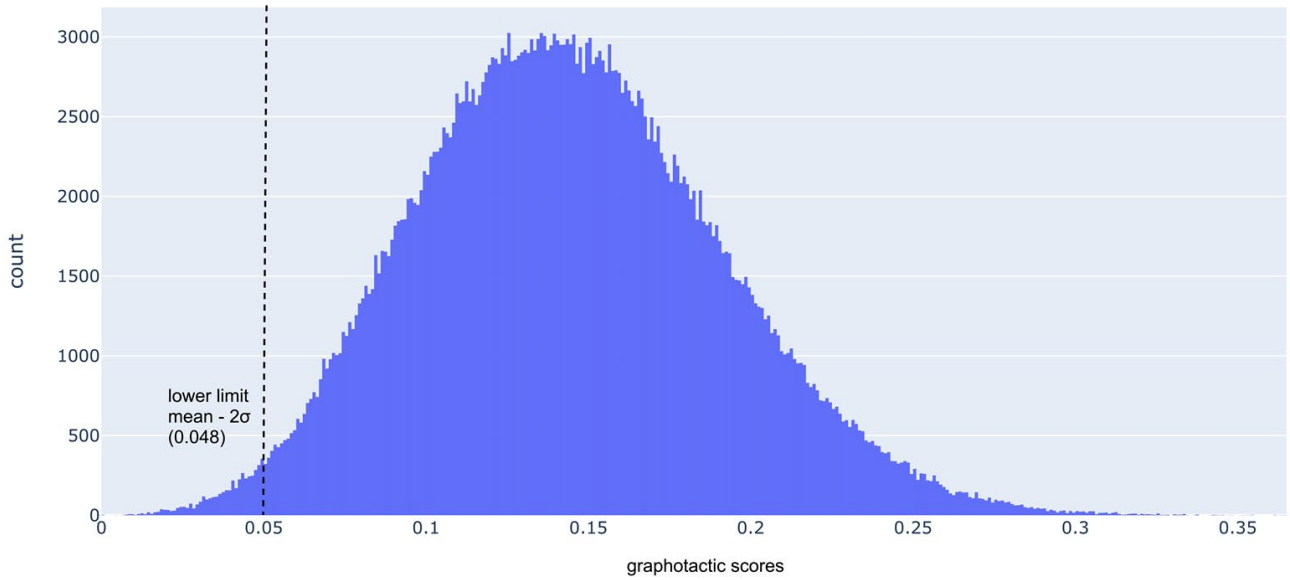

Figure 1. Distribution of graphotactic scores in the full dataset

3. Model

Phonolette is based on a Long Short-Term Memory (LSTM) network architecture (Hochreiter and Schmidhuber, 1997). LSTMs are networks designed for sequence processing. In particular, they are capable of predicting a phoneme sequence from a sequence of letters.

3.1 Architecture

Phonolette consists of a bidirectional LSTM encoder and a unidirectional LSTM decoder. The encoder reads the input sequence in both directions (left to right and right to left) and constructs activation states which are then passed to the decoder which outputs a phoneme transcription of the input. The encoder and decoder are composed of 100 neurons. The encoder produces an activation matrix of dimension 200 from the input sequence (100 dimensions for the activation produced during the left-to-right scan and 100 dimensions during the right-to-left scan). To improve the prediction, we used the teacher forcing protocol. It consists in providing the decoder at time ***t*** with the phoneme to be predicted at time ***t-1*** instead of the phoneme (actually) predicted at ***t-1***. Figure 2 shows the architecture of Phonolette.

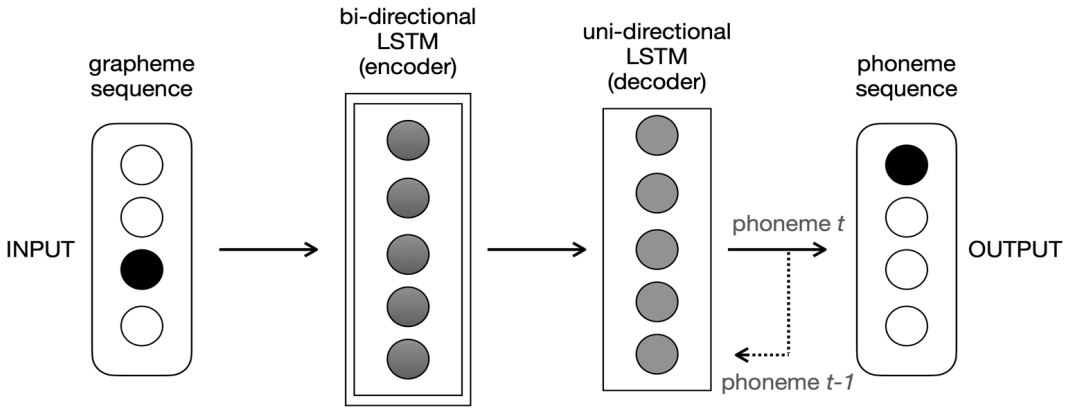

Figure 2: The Phonolette architecture consists of two LSTM networks. The teacher forcing protocol is represented by the dotted line.

3.2 Data description

When developing the model, we noticed that the part of speech played a crucial role in the predictions. For instance, a final <ent> sequence in a noun normally corresponds to the phoneme /ɑ̃/, while when it occurs at the end of a verb, it corresponds to the empty sequence. Therefore, we added the part of speech (POS) at the beginning of the written forms provided to the model as inputs[2]. Table 1 shows some examples of the inputs and outputs of the model.

Written form	**Tag**	**POS + Form (input)**	**Phono (output)**
linguistique	Ncfs	#Nlinguistique$	#lɛ̃gɥistik$
rectangulaire	Afpms	#Arectangulaire$	#ʁEktɑ̃gylɛʁ$
régénériez	Vmsp2p-	#Vrégénériez$	#ʁEʒEnEʁje$

Table 1. Examples of inputs (POS + written form) and targets (phonological transcripts). Tags are labels describing the morphosyntactic features of the forms. The symbols # and $ represent the beginning and end of the input and output sequences respectively.

The 44 lowercase letters (i.e. graphemes) that appear in the input forms, the 3 uppercase letters that encode the POSs, and the 45 phonemes that appear in the transcripts are all encoded as one-hot vectors. The training was done using 10-fold cross-validation. The dataset is divided into 10 subsets called “folds”. The system successively uses one of the subsets as

[2] When compared with inputs consisting of orthographic sequences only, using POS information improves Phonolette accuracy by 2.7%. The position of POS information, at the beginning or the end of the sequence, does not make any difference in the results because Phonolette uses a bidirectional LSTM.

evaluation data and the other 9 as training data. Phonolette uses categorical cross-entropy as a loss function and a batch size of 32.

4. Results and evaluation

Gorman et al. (2020) propose two metrics to assess the performance of G2P systems: (1) WER, word error rate, that is the percentage of words that are transcribed incorrectly. (2) PER, phone error rate, that is the sum of the Levenshtein distances between the predicted and reference (i.e. observed) transcripts divided by the sum of the lengths of the reference transcripts, where n is the number of entries, s_i is the predicted transcript and r_i is the reference transcript.

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

Like WER, PER is expressed as a percentage; the lower the value, the better the performance of the system. Table 2 shows that Phonolette performs relatively well on both datasets, with an overall accuracy of 97.82% for the full dataset and 98.11% for the graphotactic dataset. The PER is low for both datasets.

	WER	**PER**
whole dataset	2.18%	0.55%
graphotactic dataset	1.89%	0.48%

Table 2. WER and PER for the two datasets. Predictions obtained by 10-fold cross-validation.

The improvement in Phonolette's performance is small but considerable on the graphotactic dataset. This result shows that it is more difficult to process words with non-French graphotactics. The PER values are low for both datasets, with a slightly better performance for the second dataset, 0.48%, compared to 0.55% for the whole dataset.

A detailed analysis (Table 3) shows that prediction quality varies according to the POS of the forms. Phonolette predicts the phonemic transcription of verbs almost perfectly in both datasets. For adjectives, the accuracy is almost identical in both datasets (96.19% for the full dataset and 96.20% for the graphotactic dataset). On the other hand, nouns are the most difficult to transcribe.

Dataset	POS	correct	%	uncorrect	%	total
full	**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
graphotactic	**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

Table 3. Accuracy of the predicted transcripts by category on both datasets (10-fold cross-validation).

This behavior results from the basic characteristic of the POSs and their distribution. With more than 270,000 forms, verbs are the largest class and therefore the most redundant in terms of the observations to which Phonolette is exposed. Each verb lexeme yields 51 inflected forms. In comparison, nouns and adjectives together account for only 25% of the learning data. Nouns have only 2 forms and adjectives only 4 forms. This imbalance explains the differences in Phonolette's performance. Figure 3 shows, for both datasets, the number of correct and incorrect predictions for each paradigm cell of the three categories A, N and V.

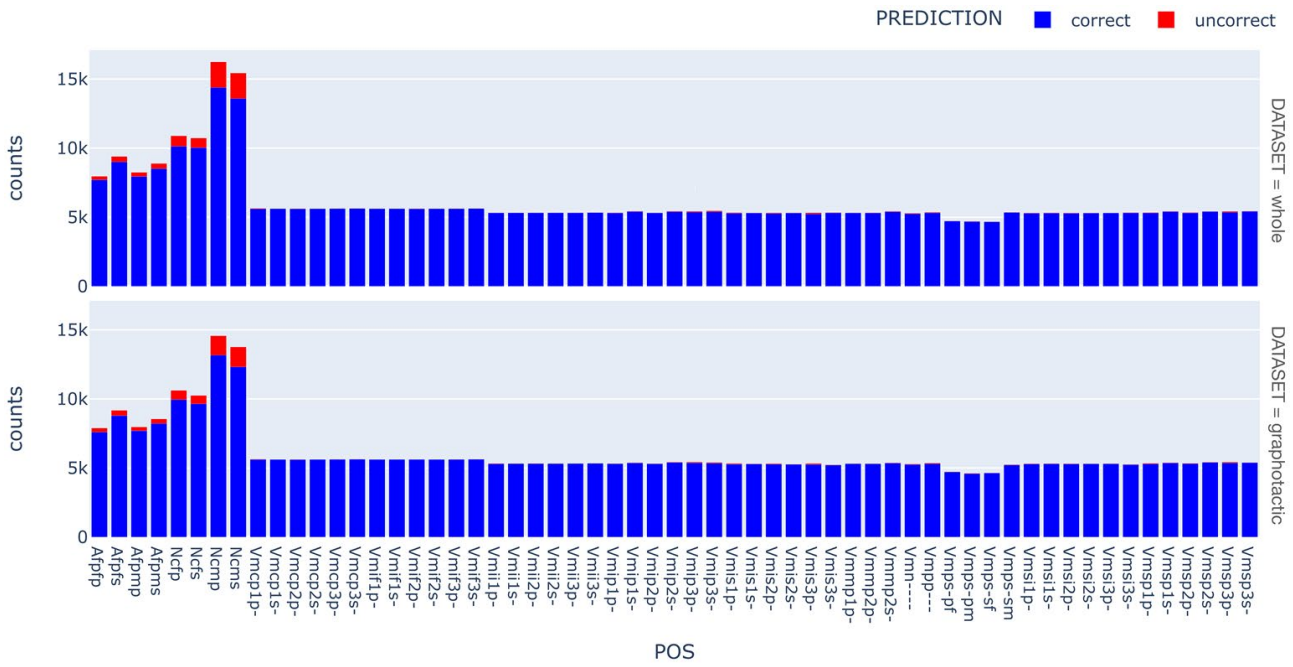

Figure 3. Number of correct and incorrect predictions for each paradigm cell for the three categories A, N and V, and for both datasets

The same is true for the PER ratio, as we can see in Table 4.

Dataset	POS	PER
full	**A**	1.06%
	N	2.70%
	V	0.10%
graphotactic	**A**	1.06%
	N	2.29%
	V	0.09%

Table 4: PER, phone error rate, by POS for both datasets (10-fold cross-validation).

Figure 4 shows the PER for each paradigm cell of the three POSs.

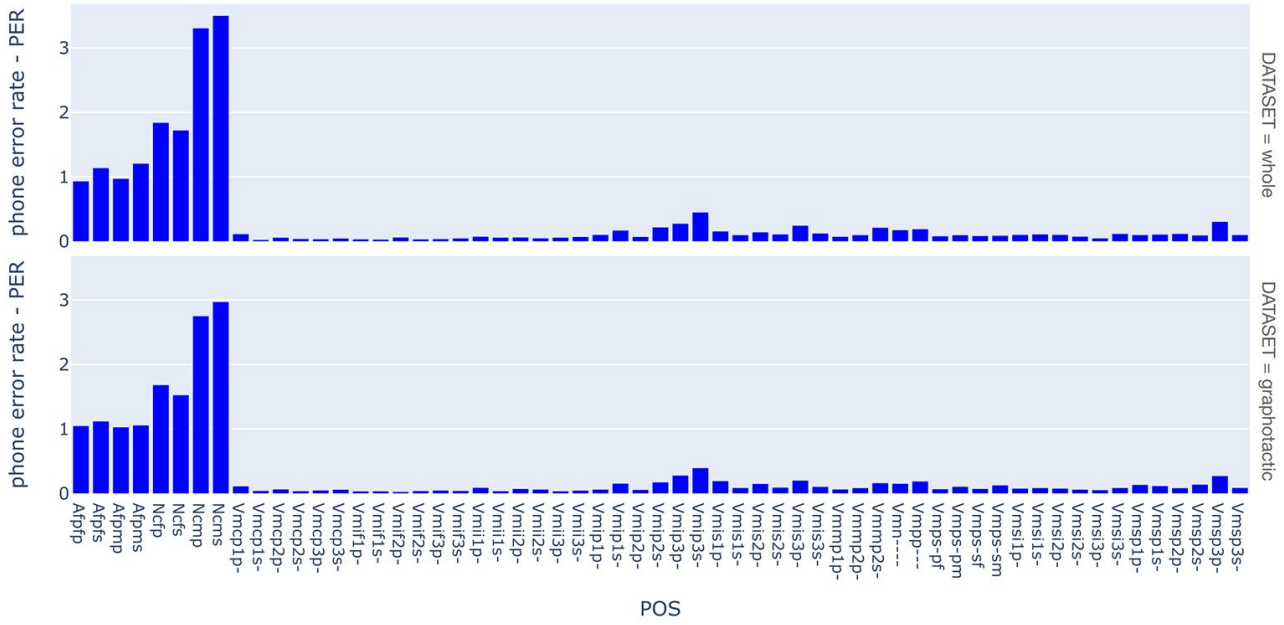

Figure 4. PER for each paradigm cell of the three POSs A, N, and V, and for the two datasets.

5. Error analysis

In this section we will discuss some of the phonologization errors made by Phonolette on the two datasets. Tables 6 and 7 show the 20% of all errors and their percentages for the full and the graphotactic datasets respectively.

Oper.	**Neutr.**	**Phon.**	**%**	**Prediction**	**Target**	**Example**
sub.		O/o	3.12	EgzOsfɛʁ	Egzosfɛʁ	exosphère
sub.	yes	E/ɛ	3.09	sEgmɑ̃tal	sɛgmɑ̃tal	segmentale
ins		s	2.98	tʁɑ̃sibEʁjɛ̃	tʁɑ̃ssibEʁjɛ̃	transsibérien
sub.	yes	ɛ/E	2.39	ɛ̃diʒɛstjɔ̃	ɛ̃diʒEstjɔ̃	indigestions
sub.	yes	O/ɔ	1.99	Optik	ɔptik	optiques
ins.		t	1.61	kasjɔ̃	katjɔ̃	cations
sub.		ɑ̃/ɛn	1.6	kamEʁamɑ̃	kamEʁamɛn	cameramen
del.		t	1.26	ʁEpit	ʁEpi	répits
sub.		z/s	1.24	ʁEzEksjɔ̃	ʁEsEksjɔ̃	résections
sub.		o/O	1.11	povʁəte	pOvʁəte	pauvreté

Table 5. The 10 most frequent errors of Phonolette on the full dataset (10-fold cross-validation). These errors represent 20.40% of the total number of errors on this dataset. The Flexique's neutralizations are also reported. Coding of operations: sub. = substitution; ins. = insertion; del. = deletion.

Oper.	Neutr.	Phon.	%	Prediction	Target	Example
sub.	yes	E/ɛ	3.67	pElәʁinaʒ	pɛlәʁinaʒ	pèlerinages
sub.	yes	ɛ/E	2.94	ɛ̃tɛʁaksjɔ̃	ɛ̃tEʁaksjɔ̃	interactions
sub.		O/o	2.62	tɛʁmOdinamik	tɛʁmodinamik	thermodynamique
del.		ɑ̃	2.22	kɔ̃fly	kɔ̃flyɑ̃	confluent
sub.	yes	O/ɔ	1.87	Optik	ɔptik	optiques
sub.	yes	ɔ/O	1.57	lɔtәʁi	lOtәʁi	loterie
sub.		ɑ̃/ɛn	1.45	kamEʁamɑ̃	kamEʁamɛn	cameramen
ins.		t	1.13	ʁEpit	ʁEpi	répits
sub.		E/ə	0.77	pOtɑ̃sjalizEʁE	pOtɑ̃sjalizəʁE	potentialiserai
sub.		o/O	0.74	ɑ̃tʁopomɔʁfik	ɑ̃tʁOpOmɔʁfik	anthropomorphique
del.		t	0.72	dEfisi	dEfisit	déficit
sub.		ə/E	0.65	sɔ̃bʁәʁo	sɔ̃bʁEʁo	sombrero

Table 6. The 12 most frequent errors of Phonolette on the graphotactic dataset (10-fold cross-validation). These errors represent 20.35% of the total number of errors on this dataset. The Flexique's neutralizations are also reported. Coding of operations: sub. = substitution; ins. = insertion; del. = deletion.

Most of the errors concern the coding of E and O, which occur in Flexique transcripts. Recall that this encoding neutralizes the phonological difference between the vowel pairs /e/ and /ɛ/ and /o/ and /ɔ/, respectively. The choice between neutralizing the vowels or specifying them proves difficult. Another difficult task for Phonolette in both datasets is the pronunciation or non-pronunciation of the occlusive /t/ at the end of a word. Phonolette sometimes inserts or deletes it incorrectly, as in /ʁEpit/ instead of /ʁEpi/ for <répits> or /dEfisi/ instead of /dEfisit/ for <déficits>. Conversely, the replacement of the voiced fricative /z/ with the voiceless fricative /s/, as in the case of /ʁEzEksjɔ̃ /, <résections>, instead of /ʁEsEksjɔ̃/, is a frequent error only in the full dataset. Other errors are more frequent in the graphotactic dataset, such as the omission of certain nasal vowels like /ɑ̃/ or, conversely, the use of a nasal vowel instead of the consonant group V + nasal, as in the case of <cameramen>, phonologized as /kamEʁamɑ̃/ when the target transcription is /kamEʁamɛn/.

6. Internal representation of graphemes

We have seen that Phonolette transcribes grapheme sequences into phoneme sequences using an LSTM encoder-decoder architecture. More specifically, at the end of the learning phase, the encoder produces a representation of the input grapheme sequence. This representation is then passed to the decoder, which uses it to produce a phonological sequence as output. The vector representations produced by the encoder are independent of the phonological target of the decoder. They reflect only the distribution and frequency of the characters composing the written forms and correspond to the activation state of the encoder. The activation states can be retrieved for all the written forms of the lexicon, but also for the 44 individual characters (i.e., unigrams) that compose the forms. The activation states of the 44 characters form a matrix of 44 rows and 200 columns (corresponding to the 100 dimensions of the encoder for each of the two directions). Figure 5 shows the first two components of a Principal Component Analysis (PCA) analysis of the encoder activation matrix of the 44 graphemes.

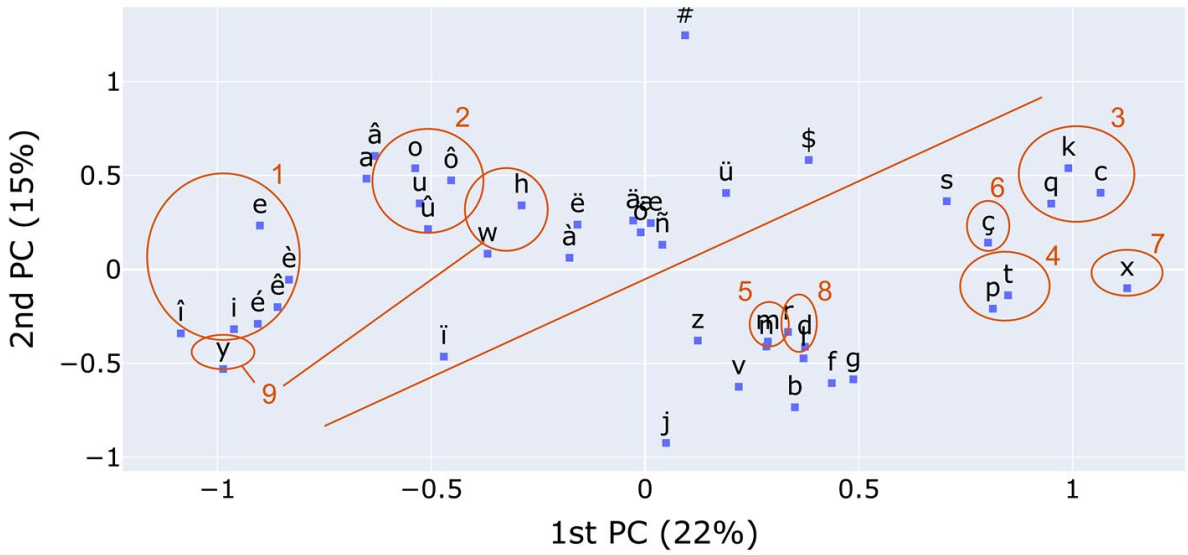

Figure 5. PCA analysis of the activation state of the Phonolette encoder for the 44 graphemes that occur in the written forms of the inflected forms.

We can see in this figure that the representations of graphemes exhibiting similar phonological behavior are close. A first clear division can be observed between consonants (bottom right) and vowels (top left). Within the vowel group, front vowels (group 1) can be distinguished from back vowels (group 2). Within the consonant group, several natural classes can be identified, such as group 3, which realizes the phoneme /k/, or group 4, which is formed by the voiceless stops /p/ and /t/. The labial nasal /m/ and the dental nasal /n/ form group 5. The position of the cedilla, ç (group 6), near the occlusive /k/ and the fricative /s/ is also

interesting. Particularly interesting is the case of the character <x>, group 7, which phonologically can realize the diphone /ks/ or /gz/. In the figure, it is rather isolated but rightly close to the area of the occlusives (group 3 and 4). The liquid consonants <r> and <l> are also grouped and form group 8 in the center of the graph. Group 9 identifies three characters <y>, <w> et <h> that appear to have similar distributions given their phonological behavior. With some exceptions (especially at word start as in <wagon> /vagɔ̃/ the three characters are often accompanied by vowel in diphthong position (<kiwi> /kiwi/).

7. Conclusion

In this article we have presented Phonolette, a phonologizer capable of transcribing French written inflected forms. Phonolette was trained and evaluated on a dataset created from two French lexicons, GLÀFF and Flexique. A second dataset was created by excluding the words that do not conform to French graphotactics. These words are mainly loanwords. Overall, the results of Phonolette are satisfactory, with an accuracy of 97.82% for the first dataset and 98.11% for the second. We have seen that the accuracy depends on the POS: the transcription of nouns contains more errors, while verbs are transcribed almost perfectly (99.5%) due to their high redundancy. Phonolette could also be used to detect transcription errors in GLÀFF, but also in Flexique, by comparing the predictions with the transcriptions of the two resources. Finally, Phonolette should be compared with recent Transformer models such as LeBenchmark (Solène et al., 2021) to explore the differences between LSTM-based and Transformer-based phonologizers.

References

Bonami, O., Caron, G., & Plancq, C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. In F. Neveu et al. (Eds), *Actes du quatrième Congrès Mondial de Linguistique Française* (pp. 2583–2596). Berlin, Germany. https://doi.org/10.1051/shsconf/20140801223

Calderone, B., Hathout, N., & Sajous, F. (2014). From GLÀFF to PsychoGLÀFF: a large psycholinguistics-oriented French lexical resource. In *Proceedings of the 16th EURALEX Conference* (pp. 431–446), Bolzano, Italy. http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202014/euralex_2014_032_p_431.pdf

Calderone, B., Hathout, N., & Bonami, O. (2021). Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 274–282). Online. http://dx.doi.org/10.18653/v1/2021.sigmorphon-1.28

Dell, F. (1995). Consonant clusters and phonological syllables in French, *Lingua*, *95*, 5–26. https://doi.org/10.1016/0024-3841(95)90099-3

van Esch, D., Chua, M., & Rao, K. (2016). Predicting pronunciations with syllabification and stress with recurrent neural networks. In *INTERSPEECH 2016: 17th Annual Conference of the International Speech Communication Association* (pp. 2841–2845), San Francisco, USA. http://dx.doi.org/10.21437/Interspeech.2016-1419

Gorman, K., Ashby, L. F. E., Goyzueta, A., McCarthy, A. D., Wu, S., & You, D. (2020). The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 40–50). Online. http://dx.doi.org/10.18653/v1/2020.sigmorphon-1.2

Lee, J. L., Ashby, L. F. E., Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A. D., & Gorman, K. (2020). Massively multilingual pronunciation mining with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 4216–4221). Marseille, France. https://aclanthology.org/2020.lrec-1.521

Sajous, F., Calderone, B., & Hathout, N. (2020). Extraire et encoder l'information lexicale de Wiktionary : quel boulot pour étrangler le goulot ! *Lexique*, *27*, 121-144. http://www.peren-revues.fr/lexique/569

Sajous, F., Hathout, N., & Calderone, B. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de la 20e conférence Traitement Automatique des Langues Naturelles*. (pp. 285-298). Les Sables-d'Olonne, France. https://aclanthology.org/F13-1021.pdf

Sajous, F., & Hathout, N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the 4th Fourth biennial conference on electronic lexicography* (pp. 405–426). Herstmonceux, United Kingdom. https://shs.hal.science/halshs-01191012v1/file/Sajous_Hathout_ELEX2015_GLAWI.pdf

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. (2019). Transformer based grapheme-to-phoneme conversion. In *INTERSPEECH 2019: 20th Annual Conference of the International Speech Communication Association* (pp. 2095–2099), Graz, Austria. https://doi.org/10.21437/Interspeech.2019-1954

Solène, E., Nguyen, M. H., Le, H., Zanon Boito, M., Mdhaffar, S. et al. (2021). Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark. In *35th Conference on Neural Information Processing Systems*. Online, USA. https://hal.science/hal-03407172v1/file/FLOWBERT_NEURIPS2021.pdf