Comparing NLP solutions for the disambiguation of French heterophonic homographs for end-to-end TTS systems

Maria-Loulou Hajj, Martin Lenglet, Olivier Perrotin, Gérard Bailly

Comparing NLP solutions for the disambiguation of French heterophonic homographs for end-to-end TTS systems*

Maria-Loulou Hajj[0000−0002−2405−1713], Martin Lenglet[0000−0002−3359−6846], Olivier Perrotin[0000−0002−9909−6078], and Gérard Bailly[0000−0003−4598−9206]

Grenoble-Alps Univ, GIPSA-Lab
11, rue des Mathématiques, St Martin d'Hères, France
firstname.lastname@gipsa-lab.fr

Abstract. This paper presents a study on different NLP solutions for French homographs disambiguation for text-to-speech systems. Solutions are compared using a home-made corpus of 8137 sentences extracted from the Web, comprising roughly one hundred instances of each of 34 pairs of prototypical words. A disambiguation system based on per-case Linear Discriminant Analysis (LDA) classifiers using contextual word embeddings as input features achieves state-of-the-art F-scores superior to 0.96.

Keywords: End-to-End Text-to-Speech · Letter-to-Sound · Heterophonic Homographs.

1 Introduction

English and French are considered to have the most opaque orthographies among languages with alphabetic (as opposed to logographic or syllabary) writing systems: fluent reading of French requires a visual attention span (VAS = the number of distinct visual elements that can be processed simultaneously at a glance) [3] of up-to 5 to 6 characters (see Fig. 1). Note that this VAS is highly structured: the whole span is not screened and processed for all characters. On the other hand, some words may require a larger span (and likely several saccades and fixations through the text) to get properly pronounced, such as homographs. An heterophonic homograph is 'one of two or more words spelled alike but different in meaning or pronunciation' (such as 'mon fils' my son vs. 'des fils' which is the plural of 'fil', a thread or wire). The correct classification of a homograph has always been an issue for natural language processing when it comes to text-to-speech (TTS) systems and the analysis of texts. And so, using a word that has homographs, can change the meaning and context of a text if read or analyzed

* Supported by the ANR 19-P3IA-0003 MIAI. This work was performed using HPC/AI resources from GENCI- IDRIS (Grant AD011011542)

by a machine. Moreover, French has a significant amount of homographs[1] which are usually not well analyzed [4].

In this study, we will try to get the best results possible generating the pronunciation of homographs, using different approaches and methods. We will first show that popular end-to-end neural TTS systems with text input such as Tacotron2 can be trained to perform a rather performative letter-to-sound (LTS) alignment and mapping, using both aligned and non-aligned acoustic corpora as well a pronunciation dictionary. We also tested the performance of a part-of-speech (POS) tagging transformer to bias this LTS mapping. We finally compared these models with homograph classification models also built on FlauBERT; a French version of BERT. Classification is performed by Linear Discriminant Analysis (LDA) trained on a corpus of 8137 homographs observed in context.

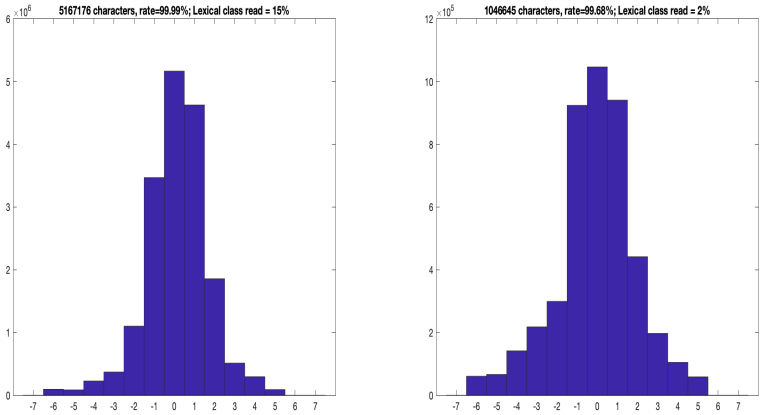

Fig. 1. Histograms of the distance of context characters left and right from the current character a minimum-length decision tree [2] has to question to pronounce it correctly. Minimum length training of these isolated words is very efficient. Left: French (using entries from the Robert dictionary); Right: English (using CMUDICT). Word entries are augmented with POS tags. On these datasets, French uses much more this information than English. Weighted means are 0.15 vs. -0.07: French seems to use a bit more look-ahead context. Note however that these data depend on the grapheme-to-phoneme alignments. Both writings need an attention span of more than 10 letters.

2 State of the art

Early solutions for letter-to-sound conversion consisted in storing orthography/pronunciation pairs of a finite list of words in a lexicon enriched with lexical, syntactic or semantic information that condition the retrieval of the right pronunciation given

[1] 789 according to `https://fr.wiktionary.org/wiki/Catégorie:Homographes_non_homophones_en_français`

orthography and information provided by POS taggers or homograph disambiguators [5]. Statistical LTS models have then been introduced to generalize LTS mapping to out-of-vocabulary words: neural sequence-to-sequence models are the current state of the art [1, 16].

The first generation of end-to-end (E2E) neural TTS such as Tacotron2 [13] or Deep Voice [11] proposed to generalise from character input to acoustic output from fairly large sets of parallel text and speech audio data, implicitly learning LTS mappings. Taylor and Richmond [15] showed that this implicit LTS models underperformed explicit LTS. Reported LTS errors (close to 10%) were quite alarming. Latest generation of E2E models now opt for a phonetic input [12]. Note that phonological variations (ways of words are pronounced) depend on linguistic context – hopefully captured by the text encoder – but also on speaking style and speaker components that usually bias embeddings computed by the phonetic encoder of current E2E neural TTS. An external component should thus restore the covariation between these segmental and suprasegemental components. With implicit LTS models, speaker and expressivity components can implicitely modulate the phonological variations at all levels in a more ecological framework. Experiments comparing explicit and implicit LTS have not yet been confirmed on French. our experience with implicit LTS on French is rather more positive, given appropriate supervision (see below).

Concerning the processing of specific LTS mappings by E2E TTS, Taylor et al [15] compared explicit vs. implicit LTS results focusing on French liaisons. They show that Tacotron2 over-inserts liaison sounds, leading to a significant preference for an explicit LTS control.

Nicolis et al [10] describe an explicit heterophonic homograph disambiguation system for English based on per-case classifiers using contextual word embeddings as input features. They report an accuracy of 0.991 with as little as 100 sentences of training material.

The current paper builds on these experiments. Our main contributions are:

- Enhanced implicit LTS for E2E TTS using grapheme-to-phoneme alignments gathered during pre-training the TTS system with both orthographic and phonetic input. Implicit LTS for general text input (more than 100 hours of read speech) achieves an accuracy of 0.989 for all input characters and 0.999 when considering only word characters.
- Homograph-specific LDA classifiers using contextual word embeddings using a similar approach as proposed by [10] achieves good performance
- Performance and generalization can be improved by clustering homographs into groups

3 Dataset and models

We collected 8137 sentences comprising at least one heterophonic homograph. The samples were collected from various sources: articles from various journals, google searches, etc. Most sentences are kept in their original phrasing, in-between punctuation marks. We nevertheless cleaned (removing lists of proper

Table 1. Multispeaker audio data used to train the Tacotron2.

Speaker	Sex	Type	#Utts		#Duration	
			All	Aligned	All	Aligned
NEB[a]	F	Audiobooks	83021	45099	71:16	34:30
DG	M	Audiobooks	20179	7749	17:16	6:31
RO	F	Read sentences	9371	0	0:00	9:57
IZ	F	Scripted dialogs	11073	386	9:28	0:17
AD	F	Read sentences	6476	2853	5:05	2:14
Total			130105	56102	112:59	43:35

[a] Part of this data is available at `https://zenodo.org/record/4580406`.

nouns, dates, etc) and shortened part of them (removing unnecessary clauses, inserts, etc). Since some homographs occur frequently (e.g. we have 1373 occurrences of the auxiliary "est" in the homograph dataset), we end-up with 9997 homographs.

3.1 Our baseline: end-to-end TTS augmented with phone prediction

Using Multidimensional Scaling, we analyzed the latent space computed by the output of the text encoder of a Tacotron2 [13] trained on 113 hours of a multispeaker French data (see Table 1). Speaker embeddings are simply added to the output of the text encoder. This baseline model was trained for 250 Epochs, a batchsize of 128 and a learning rate of $4e^{-5}$. We used the HPC facilities provided by the Jean-Zay supercomputer in Paris.

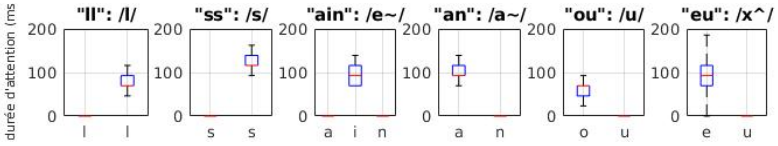

Fig. 2. Distributions of durations of activation (ms) of 6 character sequences: when one phoneme is encoded by two letters, the second character gets mostly activated in double consonants, while the first is activated for vowels.

The text encoder was trained using both text and phonetic input when hand-checked (38% of the utterances): while this representation mixing has been shown to improve spectrogram estimation [6], it also provides a letter-to-sound alignment [1] as a by-product of the Tacotron2 attention map: using activation patterns (see sample statistics in Fig. 2) and joint projection of input phones and characters embeddings into a common latent space (see Fig. 3) using Multidimensional Scaling (MDS), we obtain a lawful correspondence between phones and characters [9], including pauses and spaces/punctuations.

We thus enriched the original Tacotron2 with an additional task: phone prediction from the text encoder's output. This prediction is simply performed

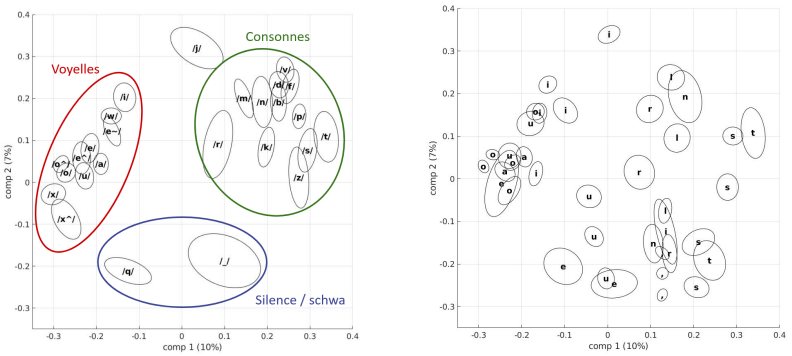

Fig. 3. Projection of the embeddings of the 25 most frequent phones (left) and the 15 most frequent characters (right) at the output of the text encoder. The projection is performed on the first factorial plane of Multidimensional Scaling (MDS) of the embeddings of all input symbols of 10% of the training data. A Gaussian mixture model clusters scattered distributions of each character. As an example, "s" has four clusters, one overlapping the phone /s/, one on /z/ and two on the silence /_/, reflecting the pronunciation of "sot", "asie" vs. "tu es bien"; "o" has five clusters, overlapping the phones /u/, /o/, /o͡/ (/ɔ/), /x̂/ (/ʃ/), and /õ/ (/ɔ̃/), reflecting the pronunciation of "loup", "dos", "cor", "coeur" and "long".

by a full-connected layer with softmax. This model is named TC2 in the following. The set of target phonemes comprises the input phoneme inventory augmented with a "silent" symbol and several diphones such as /k&s/, /i&j/, /d&z͡/ (/dʒ/) ... paired with single characters such as "x" (in "six"), "y" (in "appuyer") or "j" (as in "jazz"). We also have symbols for hiatus, syntactic vs. breath pauses, often paired with punctuations and sometimes with spaces.

For aligned utterances, two input/output patterns are thus provided: the output pattern for orthographic input has "silent" symbols and diphones, while that for phonetic input has one-to-one correspondence except for spaces and punctuations that may be associated with "silent" symbols or pauses. Of course, both utterances are associated with the same spectrogram. Please find an excerpt of our corpus below (note that the name of the audiofile and the timestamps have been removed):

```
sans l'impérieuse exigence,|s a~ _ _ _ l _ e~ _ p e r j x _ z _ _ e^ g&z i z^ a~ _ s _ __
{s a~} {l e~ p e r j x z} {e^ g z i z^ a~ s},|s a~ _ l e~ p e r j x z _ e^ g z i z^ a~ s __
```

Note that TC2 was only exposed to the homographs used in the audiobooks (plus one exemplar per homograph from the Robert lexicon, see below). The column #obs/Audio of Table 2 gives the number of occurrences of each homograph in the 113 hours corpus: This distribution is very uneven: the auxiliary verb "est" occurs more than 10000 times whereas the homographs "adoptions", "détections", "négligent", "somnolent" and "pressent" occur less than once.

The resulting letter-to-sound mapping is quite accurate (see Fig. 4). The accuracy of phonetic predictions of 4771684 word characters is superior to 99.9% and close to 98.9% when considering punctuations and other special characters.

Note that this mapping opens up the possibility to improve pronunciation accuracy for words not present in the audiobooks (modern terms[2], loan or rare words, etc). We thus performed the letter-to-sound alignment of the pronunciation input from the Robert dictionary (1995 version) and also include some conjugated forms. We thus added one exemplar of each homograph given in minimal context (adding sufficient grammatical words for disambiguisation, e.g. "Ils convient." vs "Il convient"). When training TC2 speaker embeddings, these additional "normative" out-of-context 104332 entries are set for all speakers with no dialectal nor style variation.

The final parametrization of TC2 was trained in two steps: 10 epochs for training the phone prediction layer from the frozen baseline, then 40 epochs for fine-tuning the whole model.

While not having observed so many homographs (see column "audio" in Table 2), the pronunciation accuracy of homographs is quite good (see F-scores in Table 2). Some scores ("convient", "minerai", etc) are not so high for several reasons: large asymmetry of the empirical distributions (frequency of appearance of each homograph is highly imbalanced: "est" as auxiliary is twenty times more frequent than "est" as noun, the use of "minerai" or "violent" as conjugated verbs in our audiobooks are never encountered), limited amount of training material and detection capabilities of the text encoder, etc.

3.2 Part-of-speech (POS) Tagging

Most homographs can be solved using POS Tagging. The few of them with identical POS require computation of grammatical variables (such as "convient" or "os") or semantic analysis (such as "fils"). POS Tagging has been shown to improve pronunciation accuracy, phrasing and prosody [14]. In this paper, we used the *Hugging Face* French POS tagger trained on the free French-treebank dataset [3].

The pronunciation accuracy of homographs given this POS prediction (see F-scores in Table 2 is rather disappointing: probably because of the asymmetrical distributions of the homographs in the French-treebank dataset: this largely explains the poor tagging of "content", "couvent" and "parent" as verbs. It also could be due to the tag inventory that could not be appropriate to homograph disambiguation: as an example, infinitives are either classified as verbs or infinitives, adjectives may be confused with nouns or past participes.

[2] e.g. the root "techniqu" is only used 5 times in our audiobook database: with no additional patterns from a pronunciation lexicon, "ch" will likely be mispronounced with the post-alveolar fricative ʃ.

[3] https://huggingface.co/gilf/french-postag-model

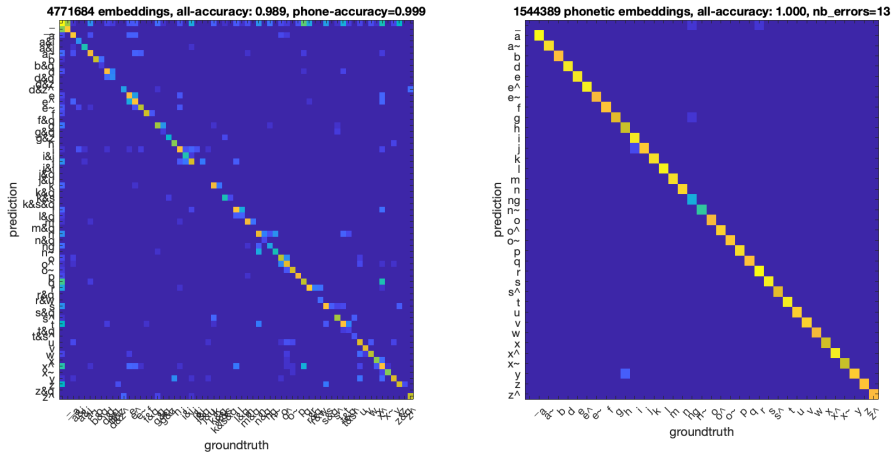

Fig. 4. Confusion matrix of phonetic prediction for orthographic (left) vs. phonetic (right) input. On average, we have three characters per phone. Trained with phonetic alignments, the output of Tacotron2 text encoder actually embeds quite precise phonetic representations.

3.3 Linear discriminant analysis of BERT embeddings

Self-supervised NLP models such as Bidirectional Encoder Representations from Transformers (BERT) use "auxiliary" or "pre-training" task – such as predicting masked words, next words in an utterance, sentence order, etc. – to learn latent representations that are further used for downstream supervised tasks [7]. Word embeddings computed by the last transformer before the final softmax layer are often used has representation vectors for the supervised tasks.

Since our task consists in separating two classes (no words with more than two pronunciations in French), we combined feature selection and linear discriminant analysis (LDA) on the BERT embeddings for each pair of homographs (see Fig. 5).

We used FlauBERT, which stands for "French Language Understanding via Bidirectional Encoder Representations from Transformers", one of the few pre-trained French version of BERT [8]. From the original 1024 embedding dimensions of the Flaubert large-cased model, we iteratively select the 110 dimensions that are the most relevant for the LDA classification.

We explored two different ways of grouping homographs:

Embeddings of word pairs (B-wrd) builds a model for each homograph. An example of this method is found in Appendix 7.1.

Embeddings of class pairs (B-grp) builds a model specific to each class of homograph. Some classes are quite large such as homographs ending with "ions" or "ent"; some are word-specfific and equivalent to B-wrd such as "fils" or "plus". This method allows us to build more general models, trained on larger groups, that can be further applied on words not seen or trained before

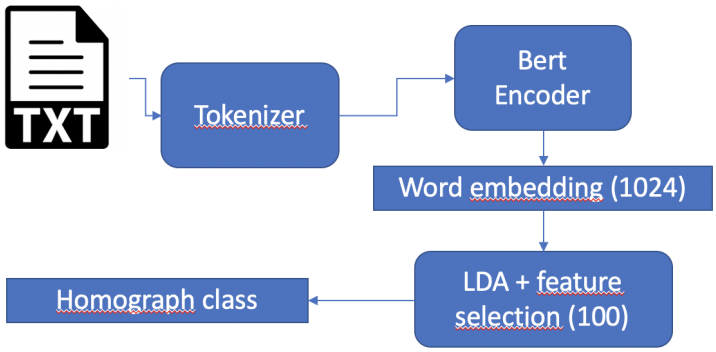

Fig. 5. Combining BERT word embeddings (using a tokenizer with a vocabulary of 68729 words and sub-word units) with feature selection and Linear Discriminant Analysis.

(or homographs that are rather difficult to observe in context). Example of this method on a poetry paragraph from a Facebook post complaining opacity of French orthography is found in Appendix 7.2.

4 Results

The goal of this research was to evaluate performance and accuracy of each of four NLP methods, which were applied for the disambiguation of French homographs: TC2, POS tagging, B-wrd and B-grp. The results of F1-scores for each homograph after each solution are found in Table 2.

Amongst these solutions, the POS tagging performed the worst, with a mean F1-score of .67, when POS tags are sufficient for disambiguation. This low score is likely to be explained by the poor and unbalanced representation of homographs in the French French-treebank. It also shows that the use of BERT embeddings is highly dependent of the targeted task: the POS ambiguities of the rare homographs are outliers when considering the empiral distribution of word tokens.

TC2, with a mean F1-score of .8, has rather good results distributed across all homographs compared to the POS tagging method. Especially for the homographs where POS tagging performed poorly, TC2 had much higher scores. Rare homographs requiring calculation of grammatical variables such as number get low scores, like "convient", "os". More detailed results about the phonetics generated for some homographs by TC2 are given in Fig. 6.

Both methods of LDA from FlauBERT embeddings, B-wrd and B-grp, got the highest and best scores distributed across all homographs but also compared to POS tagging and Tacotron2. Both performed very well, but with minor differences.

For B-wrd, the F1-score ranges from 0.843 to 0.999, making it the second-best approach applied in our research.

Table 2. Results of F1-Score for each homograph after the different methods used. Number of observations in the 113 hours of read speech and in the 8137 sentences are given in colmuns Audio (word count) and Homo (for each homograph). Abbreviations for POS tags: Ns/Np=Noun singular/plural; Vs/Vp/Inf=Verb singular/plural/Infinitive; A/Pps=Adjective/Past Participle; Pas=Negation. No errors are scored .999. F-scores below .25 and .50 are enlighten in red and blue.

Homograph	POS	Sounds	#obs		F1-Score			
			Audio	Homo	POS	TC2	B-wrd	B-grp
Actions	Np/Vp	[aksjɔ̃]/[aktjɔ̃]	38	44/37	.22	.95	.991	.995
Adoptions	Np/Vp	[adɔpsjɔ̃]/[adɔptjɔ̃]	0	149/149	.92	.97	.993	.993
Affections	Np/Vp	[afɛksjɔ̃]/[afɛktjɔ̃]	10	148/106	.99	.98	.999	.999
Collections	Np/Vp	[kɔlɛksjɔ̃]/[kɔlɛktjɔ̃]	11	145/139	.04	.97	.978	.989
Détections	Np/Vp	[detɛksjɔ̃]/[detɛktjɔ̃]	0	58/44	.99	.97	.967	.989
Intentions	Np/Vp	[ɛ̃tɑ̃sjɔ̃]/[ɛ̃tɑ̃tjɔ̃]	13	101/40	.27	.91	.950	.987
Options	Np/Vp	[ɔpsjɔ̃]/[ɔptjɔ̃]	6	80/38	.80	.69	.843	.999
Portions	Np/Vp	[pɔʁsjɔ̃]/[pɔʁtjɔ̃]	12	93/52	.80	.73	.972	.999
Affluent	Ns/Vp	[aflyɑ̃]/[afly]	4	143/148	.98	.89	.999	.999
Couvent	Ns/Vp	[kuvɑ̃]/[kuv]	13	149/142	.04	.83	.976	.989
Ferment	Ns/Vp	[fɛʁmɑ̃]/[fɛʁm]	7	102/104	.68	.84	.995	.995
Parent	Ns/Vp	[paʁɑ̃]/[paʁ]	9	140/153	.08	.94	.987	.997
Résident[a]	Ns/Vp	[rezidɑ̃]/[rezid]	0	101/99	.95	.62	-	.990
As	N/Vs	[as]/[a]	342	157/141	.95	.94	.995	.999
Bus	N/Vs	[bys]/[by]	4	163/149	.15	.92	.993	.993
But	Ns/Vs	[byt]/[by]	88	125/149	.50	.91	.993	.993
Sens	N/Vs	[sɑ̃s]/[sɑ̃]	213	208/155	.99	.95	.997	.997
Vis	N/Vs	[vis]/[vi]	128	120/99	.98	.94	.990	.999
Content	A/Vp	[kɔ̃tɑ̃]/[kɔ̃t]	56	167/190	.04	.94	.997	.999
Excellent	A/Vp	[ɛksɛlɑ̃]/[ɛksɛl]	60	213/96	.56	.80	.979	.999
Négligent	A/Vp	[negliʒɑ̃]/[negliʒ]	1	106/108	.92	.79	.978	.981
Somnolent	A/Vp	[sɔmnɔlɑ̃]/[sɔmnɔl]	1	77/56	.97	.56	.978	.999
Urgent	A/Vp	[yʁʒɑ̃]/[yʁʒ]	16	166/114	.98	.94	.975	.978
Violent	A/Vp	[vjɔlɑ̃]/[vjɔl]	48	248/110	.86	.75	.991	.999
Convient	Vs/Vp	[kɔ̃vjɛ̃]/[kɔ̃vi]	40	108/71	-	.41	.962	.965
Pressent	Vs/Vp	[prɛsɑ̃]/[prɛs]	1	64/84	-	.73	.976	.994
Est	Ns/Vs	[ɛst]/[ɛ]	10624	114/116	.94	.90	.957	.965
Minerai	Ns/Vs	[minərɛ]/[minəre]	4	36/17	.97	.48	.962	.970
Cacher	A/Inf	[kaʃɛʁ]/[kaʃe]	56	69/112	.83	.77	.987	.991
Fier	A/Inf	[fjɛʁ]/[fje]	35	217/148	.11	.94	.992	.997
Fils	Ns/Np	[fis]/ [fil]	270	158/105	-	.85	.995	
Os	Ns/Np	[ɔs]/[o]	38	91/72	-	.32	.966	
Plus	A/Pas	[plys]/[ply]	5307	219/404	-	.71	.967	
Reporter	Ns/Inf	[ʁəpɔʁtɛʁ]/[ʁəpɔʁte]	107	59/84	.75	.58	.994	.995
Supporter	Ns/Inf	[sypɔʁtɛʁ]/[sypɔʁte]	42	89/105	.91	.70	.982	.986

[a] As an example, this homograph was not included in the training dataset. The B-wrd model nevertheless scores .999.

Finally, for B-grp, the F1-score ranges from 0.957 to 0.997. Its results are always higher than the minimum of B-wrd's results.

5 Comments

We tried to populate the corpus of different French homographs with a balanced set of utterances in their respective contexts, in order to unbias homograph recognition. Several solutions were applied for the disambiguation of French homographs. We worked on POS, Tacotron2, LDA/BERT models and trained the model to get more accurate results using four different methods:

- **POS** is very accurate for most homographs but its performance heavily depends on empirical distributions of the underlying BERT model and the corpus with TAG labels used for supervised training. Hence, the extremely low scores especially below .25 (highlighted in red) are probably due to underrepresented homophones (e.g. "collections", "intentions", "couvent" or "parent" as verbs. It is our low anchor.
- **Tacotron2** performed quite well on average. Compared to POS tagging, it has full coverage of homographs and its text encoder seems to be able to perform some semantic calculations to solve complex cases such as "fils" or "plus". Poor performance for "minerai", "reporter", "convient" and "os" are largely explained by the empirical distribution of the examplars in the 113speech data: all "minerai" and "reporter" are nouns, the 40 "convient" are all from "convenir" (none from "convier") and only 1 "os" is singular.
- **B-wrd** works on the embeddings of words pairs of the homograph extracted from balanced corpora. Unlike both previously mentioned solutions, this one is very accurate, with almost perfect scores.
- **B-grp** groups homographs according to the proximity of POS tags and grammatical variables involved in the disambiguation: the prediction of the pronunciation and meaning of the homograph would depend on the group it belongs. The two advantages of this model are: (a) its generalisation capabilities (see performance on the unseen homograph "résident" in table 2), (b) its robustness, since LDA works on bigger samples.

6 Conclusions and perspectives

We extend the work performed by Nicolis et al [10] on English. We collected a significant database of heterophonic homographs for French. We show that the grouping of homographs into grammatical cases offers generalization and robustness. Some groupings are not so successful: "est" and "minerai" should certainly be treated separately. More generally, grouping should be automatized, in particular when considering other rare homographs (e.g. "bis", "hélas", "sus"). We did not consider here the problem of phonological variation: some homographs could be pronounced with optional liaisons, mute-es, schwas, depending on numerous factors such as speed, context or speaker's style. One possibility is to use

representation mixing and only overwrite parts of the words that are ambiguous: "fils" when used as "thread/wire" could be rewriten as `"{f i l}s`, "est" when used as the auxiliary "is" as `{e^}t`, etc. The idea is to combine the B-grp precision for solving rare heterophonic homographs with the TC2 flexibility for handling implicit LTS. First results are encouraging but the interaction with other components should be analyzed.

We also show that text encoders of current end-to-end TTS are capable of performing quite impressive LTS mapping, given proper LTS alignment and mixed input training. Augmenting training material with homographs in context – and not only entries in isolation provided by the text/phonetic alignment data – will certainly improve performance of LTS mapping while keeping the flexibility of orthographic input for phonological variation.

We are currently exploring the impact of multi-speaker training on LTS mapping, in particular phonological variations and phrasing. Building TTS with orthographic input is a prerequisite for shaping latent spaces that can capture segmental and suprasegmental variations in the same embeddings.

References

1. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech communication **50**(5), 434–451 (2008)
2. Black, A.W., Lenzo, K., Pagel, V.: Issues in building general letter to sound rules. In: The third ESCA/COCOSDA workshop (ETRW) on speech synthesis. Jenolan Caves House, Blue Mountains, Australia (1998)
3. Bosse, M.L., Tainturier, M.J., Valdois, S.: Developmental dyslexia: The visual attention span deficit hypothesis. Cognition **104**(2), 198–230 (2007)
4. Goldman, J.P., Laenzlinger, C., Wehrli, E.: La phonétisation de" plus"," tous" et de certains nombres: une analyse phono-syntaxique. Actes de TALN99, Cargese, Corse pp. 165–174 (1999)
5. Gorman, K., Mazovetskiy, G., Nikolaev, V.: Improving homograph disambiguation with supervised machine learning. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
6. Kastner, K., Santos, J.F., Bengio, Y., Courville, A.: Representation mixing for tts synthesis. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5906–5910. IEEE (2019)
7. Kumar, A.: Nlp pre-trained models explained with examples (Sep 2021)
8. Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D.: Flaubert: Unsupervised language model pre-training for french (2019). https://doi.org/10.48550/ARXIV.1912.05372, https://arxiv.org/abs/1912.05372
9. Lenglet, M., Perrotin, O., Bailly, G.: Modélisation de la parole avec tacotron2: Analyse acoustique et phonétique des plongements de caractère. In: 34^{e} Journées d'Études sur la Parole (JEP). pp. 845–854. Noirmoutier, France (2022)
10. Nicolis, M., Klimkov, V.: Homograph disambiguation with contextual word embeddings for TTS systems. In: ISCA Speech Synthesis Workshop (SSW). pp. 222–226 (2021). https://doi.org/10.21437/SSW.2021-39
11. Ping, W., Peng, K., Gibiansky, A., Arik, S.O., Kannan, A., Narang, S., Raiman, J., Miller, J.: Deep voice 3: Scaling text-to-speech with convolutional sequence learning. arXiv preprint arXiv:1710.07654 (2017)

12. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y.: Fastspeech: Fast, robust and controllable text to speech. Advances in Neural Information Processing Systems **32** (2019)
13. Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., et al.: Natural tts synthesis by conditioning wavenet on mel spectrogram predictions (Feb 2018), https://arxiv.org/abs/1712.05884
14. Sun, M., Bellegarda, J.R.: Improved pos tagging for text-to-speech synthesis. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5384–5387. IEEE (2011)
15. Taylor, J., Richmond, K.: Analysis of pronunciation learning in end-to-end speech synthesis. In: INTERSPEECH. pp. 2070–2074 (2019)
16. Yao, K., Zweig, G.: Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. arXiv preprint arXiv:1506.00196 (2015)

7 appendices

7.1 Example of embeddings of word pairs (B-wrd)

For example, processing two sentences using the word "as" by the LDA from FlauBERT embeddings of "as":

```
% do_traite_homographes_heterophones.py
>>Les as.
['Les</w>', 'as</w>', '.</w>']
Les {a s}.
>>Tu les as mangés.
['Tu</w>', 'les</w>', 'as</w>', 'mangés</w>', '.</w>']
Tu les {a} mangés.
```

7.2 Example of embeddings of class pairs (B-grp)

Phonetization of heterophone homographes of a FaceBook post of French poetry with no errors:

```
Nous {p o^ r t j o~} les {p o^ r s j o~}.
Les poules du {k u v a~} {k u v}.
Mes {f i s} ont cassé mes {f i l}.
Il {e^} à l'{e^ s t}.
Je {v i} ces {v i s}.
Cet homme {e^} {f j e^ r}. Peut-on s'y {f j e}?
Avant, nous {e d i t j o~} de belles {e d i s j o~}.
Je suis {k o~ t a~} qu'ils {k o~ t} ces histoires.
Il { k o~ v j e~} qu'ils {k o~ v i} leurs amis.
Ils ont un caractère {v j o^ l a~}: ils {v j o^ l} leurs promesses.
Nos {e~ t a~ s j o~} sont que nous {e~ t a~ t j o~} ce procès.
Ils {n e g l i z^} leurs devoirs, je suis moins {n e g l i z^ a~} qu'eux.
Ils {r e z i d} à Paris chez le {r e z i d a~} d'une nation étrangère.
Les cuisiniers {e k s e^ l} à faire ce mets {e k s e l a~}.
Les poissons {a f l y} à un {a f l y a~}.
```

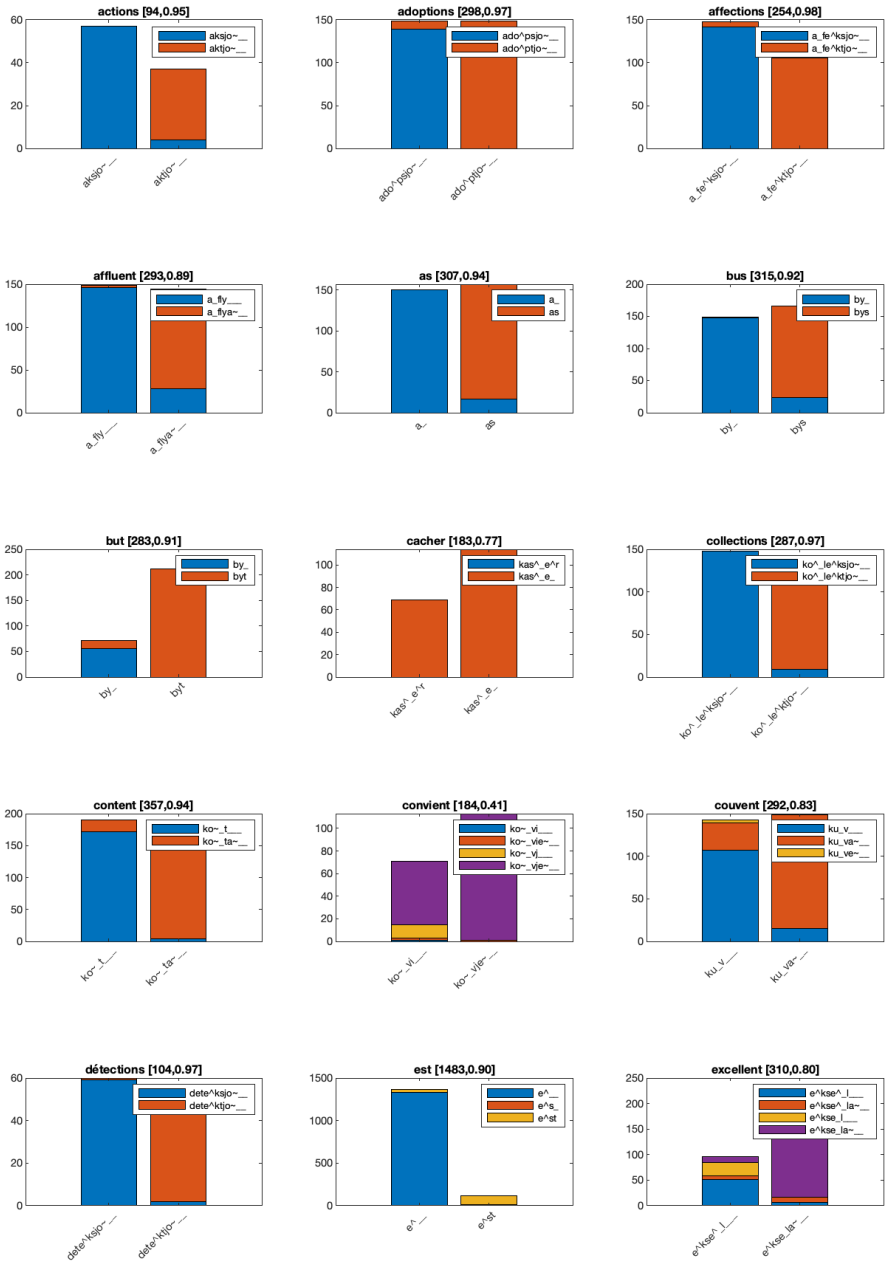

Fig. 6. Average F1-score of the first 15 homographs processed by our Tacotron2 end-to-end system augmented with phone prediction. Phonological variations of homographs are aggregated.