A toolbox for phonologizing French infant-directed speech corpora

Maria Julia Carbajal[1], Camillia Bouchon[2], Emmanuel Dupoux[1] & Sharon Peperkamp[1]

[1]LSCP, Ecole normale supérieure, PSL University, EHESS, CNRS, Paris, France
[2]Universitat Pompeu Fabra, Barcelona, Spain

December 19, 2018

Abstract

We developed an automatic French text phonologizer that transforms orthographic transcriptions of speech into approximate phonological transcriptions taking into account four phonological rules, i.e. liaison, liquid deletion, *enchaînement*, and "je"-devoicing. Using this tool, we compiled a phonologized corpus of speech input to infants under two years old, based on CHILDES transcriptions. We provide a description of the phonologizer as well as of the corpus.

Introduction

Many researchers studying language processing – whether in adults, children or infants – would like to have access to phonological transcriptions of spontaneous speech corpora. Unfortunately, this type of transcription is rarely available, and building it from scratch can be costly and time consuming. Alternatively, given an orthographically transcribed corpus – which is often easier to obtain –, a rough phonological approximation could be derived using a phonological dictionary. However, the surface form of a given word can differ from its canonical phonological form due to phonological processes that apply in continuous speech. While it is not possible to recreate the exact phonological form of an utterance without having access to its audio recording, it is possible to improve the first-order phonological dictionary approximation by applying a set of simple phonological rules. This solution, while not perfect, offers a good balance between accuracy and use of resources.

Here, we apply this method to obtain a phonologically transcribed corpus of French input to infants under two years old[1], which can be used by language development researchers. Orthographic transcriptions of interactions between parents and infants under the age of two years were obtained from the CHILDES database[2]. In the following sections we describe the CHILDES corpora that we used and the pipeline we implemented to obtain a corpus of phonologized French infant-directed speech. The phonologized corpus, as well as all the scripts needed to generate it, are publicly available at https://github.com/juliacarbajal/french_phonologizer. The scripts are easy to modify, such that phonological rules may be added, removed or modified, according to the user's needs and insights.

Corpora

To construct the phonologized corpus of infant-directed speech we compiled French corpora from the CHILDES database (http://childes.talkbank.org) that respect three conditions: (1) The corpus includes at least one transcribed session where the target child was between the ages of 0;0 and 2;0. (2) The transcription includes speech from adults. (3) The transcribed conversation is casual (i.e., no specific task or experiment involved, but may include interactions between the investigator and the parents). Based on these criteria, we included - partially or completely - the following corpora:

- Champaud (Champaud, 1994)
- Geneva (Hamann et al., 2003)
- Hunkeler (Hunkeler, 2005)
- Lyon (Demuth and Tremblay, 2008)
- Paris (Morgenstern and Parisse, 2007)
- Pauline (Bassano and Mendes-Maillochon, 1994)
- Yamaguchi (Yamaguchi, 2012)
- York (Plunkett, 2002)

[1]It should be noted that given the nature of the available corpora, we have no way to distinguish between direct and indirect input.
[2]http://childes.talkbank.org

Based on the supplementary information provided in each CHAT file, speakers in each transcription were tagged as adults or children. Only speech from adults was retained for processing. The final compiled corpus contains 98108 utterances produced by adults in the presence of children under the age of two years. (In the vast majority of cases the children were in their second year of life; the youngest age is 0;11.12).

Pipeline

The processing pipeline is divided in three steps, implemented in separate scripts:

1. Cleaning: `clean_corpus.py` takes speech transcriptions written in CHAT format[3] and returns an orthographic transcription, without annotations. During this step, utterances produced by children are filtered out.
2. Phonologizing: `phonologize.py` takes each clean orthographic transcription and transforms it into a phonological transcription (based on European French phonology). A base phonological form for each word is retrieved from the French lexical database *Lexique 3* (v3.80, New, Pallier, Ferrand, and Matos, 2001)[4]. Next, four French phonological rules are applied, in the following order: liaison, liquid deletion, *enchaînement*, and "je"-devoicing. The output is written using the same phonetic symbols as those used by Lexique.
3. Compiling: `compile.py` compiles a single corpus by concatenating the transcriptions obtained in previous steps. The script allows to define the age range, the type of transcription (e.g., clean orthographic transcriptions, or phonologized transcriptions with one or more of the rules applied) as well as other formatting options.

All scripts were written in Python 2. In the following sections we provide a detailed description of these processing steps.

Cleaning

Speech transcriptions written in CHAT format typically contain additional information regarding speakers and activities (e.g., a description of the present situation marked by `%sit` or `%act`), as well as numerous special codes and symbols indicating properties of the speech (e.g., the use of `[/]` marking the repetition of a word after a pause). Using TalkBank's CHAT transcription manual (see link in footnote 3) as a reference, `clean_corpus.py` cleans the transcriptions, removing or replacing annotations when appropriate. As a general rule, pauses are replaced by commas, and words that could not be transcribed (due to being inaudible or unrecognized by the transcriber) – usually coded as `xxx` or `www` – are replaced by a hashtag symbol #. The output file contains one utterance per line, each preceded by the name of the original CHAT file as well as the age of the infant at the moment of the recording (separated in years, months and days). Below is an example of two extracts of speech produced by the mother in a transcription from the Lyon corpus (Demuth and Tremblay, 2008) when the child was 1 year and 23 days old.

Raw CHAT format:

```
*MOT: <viens [/] viens avec moi> [=! chuchote]. 279942_283165
%act: Mot prend Ana et la haise et les font tourner en direction du salon.
*MOT: viens voir.
...
*MOT: non: xxx (...) pas à la bouche. 1158210_1165915
%sit: Ana met un cube à la bouche.
%act: Mot fait non de la tête et enlève le cube de la bouche de sa fille.
*MOT: mais!
```

After cleaning:

```
ana02a.cha 1 0 23  viens , viens avec moi .
ana02a.cha 1 0 23  viens voir .
...
ana02b.cha 1 0 23  non # , pas à la bouche.
ana02b.cha 1 0 23  mais!
```

This cleaning script takes the full set of CHAT files from a given corpus (which should be located in a folder called `corpora/corpus_name/raw/`) as input and returns an output file (`extract.txt`) with the compiled, cleaned-up corpus. If more than one corpus is available, it will create one `extract.txt` file per corpus, each located in `corpora/corpus_name/clean/`.

[3] `http://talkbank.org/manuals/CHAT.pdf`
[4] `http://www.lexique.org`

Phonologizing

As previously mentioned, the phonologizer first retrieves the canonical phonological form of each word in the corpus from a French dictionary.[5] Our dictionary is based on Lexique 3.80 (New et al., 2001), which we extended to include many words that appear in the corpus but that do not exist in the Lexique database. This includes colloquial words used by parents when talking to young children (e.g., *calinou*, *pinpin*), onomatopoeic words (e.g. *guili*, *oupla*, *wouf*), transcriptions of words with partial omissions (e.g., *(a)ttention*, *c(r)ocodile*), proper nouns (e.g., *Chloé*, *Mickey*, *Babar*) and misspelled words (e.g., *ca* instead of *ça*). Any remaining words, which were hence neither in Lexique nor added by hand to the dictionary, were transcribed with a hashtag symbol #. All hashtag occurrences (including the ones marking orthographically untranscribed words introduced during the cleaning step described above) were kept in the corpus, since the rules we implemented apply word-finally and are conditioned by the following word; thus, any rule should be blocked from applying if there is no phonological transcription of the following word.

The phonologizer script (`phonologize.py`) then proceeds to apply four French phonological rules that alter the surface form of words when embedded in utterances: liaison, liquid deletion, "je"-devoicing and *enchaînement*. To obtain only a first-order phonological approximation of the corpus, without any rules applied (i.e., only based on the phonological dictionary), an alternative script is provided, `recode.py`, which can be found in the same Github repository.

Below we describe the four phonological rules implemented in `phonologize.py`, with examples extracted from the compiled corpus. It should be noted that, while most of these phonological rules apply in specific morphosyntactic contexts, we do not generally count with this type of information. Indeed, obtaining a reliable morphosyntactic tagging of the corpus would be difficult due to the numerous irregularities occurring in the transcriptions. Instead, contexts for each rule are evaluated, when necessary and possible, based on local information such as sequences of words or grammatical classifications extracted from the Lexique dictionary.

Liaison

French liaison is a phonological process that concerns a limited set of words, mostly function words and prenominal adjectives. These words end in a latent consonant, most often /t/, /n/, /z/, or /ʁ/, which surfaces if the following word starts with a vowel or – with some exceptions – a semivowel and if the two-word sequence is prosodically cohesive (they are typically part of the same phonological phrase). For instance, the adjective *petit* 'small' is pronounced [pəti] in *petit chien* [pətiʃjɛ̃] 'small dog', but as [pətit] in *petit arbre* [pətitaʁbʁ] 'small tree'. In addition, the plural marker /z/ for nouns and adjectives (orthographically represented by 's' or 'x') is a liaison consonant; compare, for instance, *jolis dessins* [ʒolidesɛ̃] 'nice drawings' and *jolis arbres* [ʒolizaʁbʁ] 'nice trees'. Similarly, the third person marker /t/ for verbs (in both singular and plural forms) is a liaison consonant, but it surfaces only within clitic groups; compare, for instance, *elle en veut aussi* [ɛlɑ̃vøosi] 'she wants some too' and *en veut-elle aussi?* [ɑ̃vøtɛlosi] 'does she want some too?'.

Liaison does not not apply before vowel-initial proper names, interjections, and a few more words, such as *et* 'and' and *où* 'where'. These words were thus included in an exceptions list. Likewise, liaison does not apply before vowel-initial words containing an *h-aspiré* (that is, an orthographic *h* that is not pronounced yet behaves as a consonant for the purposes of phonology), as in *les hérissons* [leeʁisɔ̃] 'the hedgehogs'. We obtained a list of words beginning in *h-aspiré* from Wikipedia (`https://fr.wikipedia.org/wiki/H_aspiré`).

Liaison cases have traditionally been separated into mandatory and optional, with the latter showing a large amount of variation in the extent to which the rule actually applies: according to a spoken corpus analysis by Boula De Mareüil, Adda-Decker, and Gendner, 2003, production rates may vary depending on the specific morphosyntactic context, from 70% to 99%. In our pipeline, we implemented mandatory cases of liaison, which are thus likely or very likely to have been produced, but not optional ones. For instance, we implemented liaison in adjective + noun sequences, in which it is mandatory, but not in plural noun + adjective sequences, in which it is optional. In the absence of morphosyntactic tagging of the corpus, however, it is not always possible to accurately separate contexts in which a word undergoes obligatory liaison from those in which it does not. We separated contexts as best as we could in three ways. First, we defined a set of words that block liaison in the preceding word, because they typically occur at the beginning of a phonological phrase. For instance, we implemented that liaison never applies before the word *aussi* [osi] 'too'; compare for example *nous avons* [nuzavɔ̃] 'we have' and *nous aussi* [nuosi] 'we too'. Second, for certain words we defined conditions that have to be met for liaison to apply. For instance, *quand* 'when' undergoes liaison when followed by a pronoun (e.g., *quand on voit* [kɑ̃tɔ̃vwa] 'when one sees') but otherwise doesn't (e.g., *quand aller* [kɑ̃ale] 'when to go').

1. For the following words, we applied liaison throughout, modulo some exceptions (as specified in the script):
    - Articles *des, les, un*
    - Personal and indefinite pronouns *ils, elles, on, nous, vous, quelqu'un, autres*
    - Possessive pronouns *mon, ton, son, mes, tes, ses, nos, vos, leurs*
    - Prepositions *aux, en, sans, sous*
    - Adverbs *plus, tout, très*

[5] In the dictionary, a distinction is made for the phonemes /œ̃/ as in *brun* and /ɛ̃/ as in *brin*. As this distinction is absent in many varieties of French, particularly in France, we offer an option to collapse these two phonemes in the `compile.py` script.

- Numbers *six, dix*
- Indefinite, demonstrative, and interrogative adjectives *certains, certaines, plusieurs, aucun, quelques, ces, quels, quelles*

2. For the following words, we applied liaison only in specific contexts
   - 3rd person verbs in clitic groups (e.g., *faudrait-il, peut-on, appellent-elles*)
   - Masculine singular forms of prenominal adjectives ending in a liaison consonant, e.g. *bon, grand, petit, premier*
   - Plural form of all prenominal adjectives, e.g. *petit(e)s, grand(e)s*
   - *vas/allez/allons + y*
   - *pends/prenez/prenons + en*
   - *dans + un/une*
   - *quand + il/elle/on/ils/elles*
   - *quant + à/aux*

In some cases in which the word undergoing liaison originally ends with a nasal vowel, liaison induces denasalization of the vowel, e.g., bon [bɔ̃] turns into [bɔn] before a vowel-initial word, as in *bon ami* [bɔnami] 'good friend'. The full list of words undergoing denasalization can be found in the script `phonologize.py`.

Finally, a number of additional exceptions and corrections were added within the function that applies liaison. Thus, in some cases liaison would incorrectly be triggered due to the absence of a missing comma in the transcriptions. For instance, in *Ça y est il va partir* 'That's it he's about to leave', liaison does not apply between *est* and *il*, and one would normally put a comma after *est*. Other cases require a sequence of more than two words to infer whether liaison can apply or not. For instance, liaison typically does not apply in *était une* 'was a' (as in *elle était une gamine* [ɛletɛyngamin] 'she was a small girl'), but it does apply in *il était une fois* [iletɛ**t**ynfwa] 'once upon a time'; conversely, liaison applies within *vas-y* [va**z**i] 'go ahead', but not in *tu vas y* (as in, for instance, *tu vas y arriver* [tyvaiaʁive] 'you will get there').

Below is an example of a phonologized utterance, before and after applying liaison.

Clean orthographic transcription:

`ana04b.cha 1 01 27 ah c' est un éléphant !`

Base phonological form:

`ana04b.cha 1 01 27 a s' e 1 e-le-f@ !` (In IPA: [a s e œ̃ e.le.fɑ̃])

Phonological form with liaison:

`ana04b.cha 1 01 27 a s' et 1n e-le-f@ !` (In IPA: [a s et œ̃n e.le.fɑ̃])

Note that at this point in the pipeline, enchaînement has not yet taken place. That step, which would render the correct syllabification of this phrase (i.e., [a se.tœ̃.ne.le.fɑ̃]), will be explained further below.

Liquid deletion

In French, when a word finishing in an obstruent-liquid cluster (e.g., *table* [tabl] 'table') is followed by a consonant-initial word, the liquid will often be omitted, e.g., *table marron* 'brown table' will be pronounced as [tabmaʁɔ̃]. This phonological process has been found to occur frequently in child-directed speech (87% of all OL#C contexts in the audio files of the Yamaguchi corpus analyzed by Peperkamp and Hegde, *In preparation*). According to the traditional textbook view, liquid deletion may also apply before pauses (Dell, 1995). However, due to the low production rate of liquid deletion before pauses (31%) found by Peperkamp & Hegde (in prep.), we decided to exclude these cases. Below is an example of an utterance before and after liquid deletion:

Clean orthographic transcription:

`ana02a.cha 1 0 23 et voilà je suis passée de l' autre côté !`

Base phonological form:

`ana02a.cha 1 0 23 e vwa-la Z2 s8i pa-se d2 l otR ko-te !` (In IPA: [e vwa.la ʒə sɥi pa.se də l otʁ ko.te])

Phonological form after liquid deletion:

`ana02a.cha 1 0 23 e vwa-la Z2 s8i pa-se d2 l ot ko-te !` (In IPA: [e vwa.la ʒə sɥi pa.se də l ot ko.te])

While in our pipeline we have applied this rule after liaison, the order of implementation does not affect the results, as liaison and liquid deletion do not interact.

Enchaînement

The next step in the pipeline is *enchaînement* (resyllabification) of word-final consonants before vowel-initial words. This holds for all word-final consonants, whether they result from liaison (e.g. *un ami* [œ̃.**n**a.mi] 'a friend') or not (*une étoile* [y.ne.twal] 'a star'). In the case of a word finishing in an obstruent-liquid cluster, the whole cluster is resyllabified, as in *être assis* [ɛ.tʁa.si] 'to be seated'. Below is an example of *enchaînement* applied to the liaison case shown earlier.

Phonological form with liaison:

`ana04b.cha 1 01 27 a s' et 1n e-le-f@ !` (In IPA: [a s et œ̃n e.le.fɑ̃])

Phonological form with liaison and enchaînement:

`ana04b.cha 1 01 27 a se t1 ne-le-f@ !` (In IPA: [a se tœ̃ ne.le.fɑ̃])

je-Devoicing

French obstruents in word-final position may undergo voicing assimilation if the following word is part of the same phonological phrase and begins with an obstruent of the opposite voicing value. Thus, a voiced obstruent may become voiceless if followed by a voiceless obstruent, and a voiceless obstruent may become voiced if followed by a voiced obstruent. Voicing assimilation is optional; according to a corpus study of journalistic speech it applies in less than 25% of cases (Hallé and Adda-Decker, 2007). We therefore did not implement it, with one exception, though. Indeed, there is one very frequent case of voicing assimilation, concerning the pronoun *je* 'I'. In phrases such as *je t'écoute* 'I hear you' or *je pense* 'I think', where *je* is immediately followed by a voiceless obstruent, the schwa in *je* will often be omitted and the consonant devoiced, resulting in [ʃte.kut] or [ʃpɑ̃s], respectively. Additionally, if the following word begins with /s/, as in *je sais pas* 'I don't know', the devoiced *je* [ʃ] may be merged with the following [s], resulting in a single consonant, [ʃɛ.pa]. This occurs very frequently with the verbs *sais* 'know' and *suis* 'am'. In our pipeline, we implemented the devoicing of *je* before voiceless obstruents only when this word was transcribed either as `j(e)` or `j'` (thus indicating an ommitted schwa). Additionally, if the following word was either *sais* or *suis*, we implemented the merging of the two voiceless consonants. In practice, this phonological rule was implemented at the end of the *enchaînement* step.

Other phonological rules

Other phonological rules may be included in the pipeline to improve the phonological transcriptions. One such rule is schwa insertion, i.e., the insertion of /ə/ after a word-final consonant cluster if the following word begins with a consonant cluster, as in *faible pluie* [fɛbləplɥi] 'light rain' (cf. *faible* [fɛbl]). While this rule is not uncommon in French, its use is inconsistent and very few appropriate contexts are found in our corpus. Thus, we excluded it from our final pipeline but left it as an optional step in the script `phonologize.py`.

Below is an example of an utterance, before and after schwa insertion. In this example, neither liaison nor liquid deletion applied.

Clean orthographic transcription:

`leonard-10-2_06_07.cha 2 06 07 une grosse barbe blanche .`

Base phonological form:

`leonard-10-2_06_07.cha 2 06 07 yn gRos baRb bl@S .` (In IPA: [yn gʁos baʁb blɑ̃ʃ])

Phonological form after schwa insertion:

`leonard-10-2_06_07.cha 2 06 07 yn gRos baRb°  bl@S .` (In IPA: [yn gʁos baʁ.bə blɑ̃ʃ])

Output of the phonologizing step

The phonologizer script `phonologize.py` produces two types of output files:

- Phonologized transcriptions: After each step in the pipeline, a new phonological transcription is produced with the applied rule, as shown in the examples above. The transcription files are called `phonologized_<X>.txt`, where `<X>` indicates the rules applied so far, i.e., `L` = liaison, `D` = liquid deletion, `E` = enchaînement plus "je"-devoicing. Thus, the phonologized transcription with all four rules applied is called `phonologized_L_D_E.txt`.
- Lists of cases: From each step, a list of applied cases (and in the case of liaison, additionally a list of rejected cases) is printed for debugging. These lists provide the utterance number, the orthographic form, the phonological form with the applied rule, and a brief context of up to 5 words. Below is an example of two printed liaison cases:

```
3027 est une        et yn               ça c' est une carotte
3045 des oiseaux    dez wa-zo           c' est des oiseaux .
```

These output files are generated in a folder called `output/` containing a subfolder for each corpus.

Compiling

The final script `compile.py` gathers all previously processed corpora and produces one output file with a compiled corpus. This script allows to select the age range as well as the type of transcriptions to compile (e.g., the cleaned-up orthographic transcriptions, or the phonologized transcriptions with all or some of the rules applied). Additionally, it provides the option to remove geminates (e.g., in *Elle lit* 'she reads' [ɛl li] → [ɛ li]), as well as to merge the rounded front nasal vowel (/œ̃/) with the unrounded one (/ɛ̃/), as is characteristic of many varieties of French. The resulting compiled corpus will be located in a folder called `compiled_corpus/`.

Phonologized corpus of infant-directed speech

Using these scripts, we built a phonologized corpus of infant-directed speech based on the CHILDES corpora mentioned earlier, called `corpus_phono_L_D_E_0y0m_2y0m.txt`, which can be found on https://github.com/juliacarbajal/french_phonologizer under `compiled_corpus/`. In this corpus we applied liaison, liquid deletion, *enchaînement*, and "je"-devoicing, but not schwa insertion. Furthermore, we merged the nasal vowels mentioned before, but we did not remove geminates.

Acknowledgments

This research was funded by a doctoral fellowship from Ecole des Neurosciences de Paris to MJC, a Marie Curie fellowship to CB (H2020-MSCA-IF-2015; Project 707996), as well as by grants from the Agence Nationale de la Recherche (ANR-17-CE28-0007-01 and ANR-17-EURE-0017).

References

Bassano, D. & Mendes-Maillochon, I. (1994). Early grammatical and prosodic marking of utterance modality in French: a longitudinal case study. *Journal of Child Language*, *21*(3), 649–675.

Boula De Mareüil, P., Adda-Decker, M., & Gendner, V. (2003). Liaisons in French: a corpus-based study using morpho-syntactic information. In *Proc. of the 15th International Congress of Phonetic Sciences, Barcelona, Spain*.

Champaud, C. (1994). The development of verb forms in French children at around two years of age: some comparisons with Romance and non-Romance languages. In *Fisrt lisbon meeting on child language, lisbon, portugal* (pp. 14–17).

Dell, F. (1995). Consonant clusters and phonological syllables in French. *Lingua*, *95*(1-3), 5–26.

Demuth, K. & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language*, *35*(1), 99–127.

Hallé, P. & Adda-Decker, M. (2007). Voicing assimilation in journalistic speech. In *16th international congress of phonetic sciences* (pp. 493–496).

Hamann, C., Ohayon, S., Dubé, S., Frauenfelder, U. H., Rizzi, L., Starke, M., & Zesiger, P. (2003). Aspects of grammatical development in young French children with SLI. *Developmental Science*, *6*(2), 151–158.

Hunkeler, H. (2005). Aspects of the evolution of the early lexicon in the interactions mother-child: case study of two dizygotic twin children between 15 and 26 months. *University of Rouen*.

Morgenstern, A. & Parisse, C. (2007). Codage et interprétation du langage spontané d'enfants de 1 à 3 ans. *Corpus n6: "Interprétation, contextes, codage"*, 55–78.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: lexique[pleaseinsert"prerenderunicode–"intopreamble]//a lexical database for contemporary french: lexique[pleaseinsert"prerenderunicode–"intopreamble]. *L'année psychologique*, *101*(3), 447–462.

Peperkamp, S. & Hegde, M. (*In preparation*). Liquid deletion and liaison in child-directed speech.

Plunkett, B. (2002). Null Subjects in child French interrogatives: A view from the York Corpus. *Romance corpus linguistics: Corpora and spoken language*, 441–452.

Yamaguchi, N. (2012). *Parcours d'acquisition des sons du langage chez deux enfants francophones.* (Doctoral dissertation, Université de la Sorbonne nouvelle-Paris III).