# Automatic Extraction of Textual and Phonemic Complexity for French Cued Speech

**Magali Norré,[1,3] Brigitte Bigi,[1] Núria Gala,[1]**
**Ludivine Javourey-Drevet,[2] Thomas François[3]**
[1] Aix Marseille Univ, LPL, CNRS (UMR 7309), Aix-en-Provence, France
[2] Univ Lille, SCALab, CNRS (UMR 9193), Lille, France
[3] Université catholique de Louvain, CENTAL, ILC, Louvain-la-Neuve, Belgium
{magali.norre, nuria.gala}@univ-amu.fr, brigitte.bigi@cnrs.fr,
ludivine.javourey@univ-lille.fr, thomas.francois@uclouvain.be

**Abstract**

This article presents the results of an analysis of a written corpus with the view of automatically generating it in French Cued Speech (CS). CS is a communication system developed for people with hearing impairment to complement speech reading at the phonetic level using hands. This visual communication mode uses handshapes in different positions near the face in combination with the mouthshape (called 'cues' or 'keys') to make the phonemes of spoken language look different from each other. Despite many studies demonstrating its benefits, there are few resources available for learning and practicing it, especially in French. As part of a wider project aimed at creating an online learning platform with automatically generated videos using an augmented reality system displaying a virtual coding, we propose to identify, extract, and analyze 41 textual and phonemic features that might be more complex to (de)code in French CS. For the automatic extraction of complexity, several tools are used: FABRA for readability, SPPAS for phonetization and CS key generation. The results show some strong correlations between readability features, few between phonemic variables, and few between the two types. An initial model is proposed for selecting texts to be recorded for learning French CS.

**Keywords:** Cued speech, hearing loss, automatic features extraction, readability

## 1. Introduction

About 5% of the world's population live with disabling hearing loss, including 34 million children (World Health Organization, 2021). Most of them (90%) have hearing parents (Jones et al., 1989): an oral language is used for everyday communication. Failure to hearing impacts learning speech and its intelligibility. It also has an impact in learning to read. Lip reading is not enough to disambiguate some sounds, such as the visemes in '***p**ain*' (bread), '***b**ain*' (bath), and '**m**ain' (hand) in French.

Cued Speech (CS) is a visual communication system designed to improve spoken language comprehension for people with hearing impairments by using handshapes, positions around the face, and mouthshapes (called 'cue' or 'key') to disambiguate phonetic information. Oral sounds can be represented with this code, originally developed for American English by Cornett (1967), and adapted to about 65 languages, including French with *Langue française Parlée Complétée* (LfPC or LPC). The term 'French CS' is used hereafter.

Although its usefulness is recognized (Leybaert and LaSasso, 2010), there are few studies on its learning and its complexity. Gala et al. (2024) were the first to mention that readability and phonemic variables could be combined to automatically estimate the complexity for French CS. They cited several readability and phonemic features but they only annotated the CS key frequency without establishing any correlations. This paper investigates these parameters to propose a first method for the automatic classification of French CS resources graded by level of learning complexity.

There are many studies on readability and textual complexity, including several on Alector, the French corpus analyzed in this paper. Javourey-Drevet et al. (2022) proposed an analysis of this corpus, but on a small part and not adapted for the target public here. Ormaechea and Tsourakis (2024) also analyzed this corpus with other features and a different research purpose: the comparison of automatic text simplification systems. Listenability, which aims to assess listening difficulty of spoken materials, is also relevant for us. Researchers in this field have generally investigated combining textual and phonemic features to explain listenability of materials for language learners (Kotani et al., 2014; Kotani and Yoshimi, 2017), including for French (Ozawa et al., 2024). Such phonemic variables also appears relevant to model CS communication.

As part of a wider project aimed at creating an online learning platform with automatically generated videos in French CS from spoken texts, we propose to identify, extract, and analyze their textual and phonemic features that might be more complex to (de)code in French CS. These features are



*Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026*, pages 61–73
11 May 2026. 

combined in a model unsupervised to classify automatically the learner texts for the future learning platform. The paper is organized as follows. Section 2 introduces French CS. Section 3 describes the project. Section 4 presents the corpus to be annotated. Section 5 defines the variables and methods used for corpus analysis. Section 6 reports the results. Section 7 concludes the paper.

## 2. French Cued Speech

CS is not a language and is not intended to replace natural sign languages. Rather, it complements them by supporting access to a spoken language when such access is required or preferred (e.g., alongside French Sign Language, LSF, for French). A CS key does not encode a whole word; it encodes phonological information. This makes initial learning relatively fast (some studies report about ten hours, although details are limited), but regular practice remains necessary. Once the system is mastered, it can be used to cue any utterance in the target spoken language, including proper nouns, neologisms, and foreign words. Parents and families need to learn CS to facilitate the child's language immersion and so that the child can benefit from the contribution of CS. There are several associations that promote and teach the French CS to different audiences, such as the ALPC in France,[1] and in Belgium.[2]

CS is based on phonemes. There are 3 possible key structures: Consonant (C), Vowel (V), or CV. In French CS, there are 8 handshapes representing (semi-)consonants (Figure 1) and 5 positions near the face representing vowels (Figure 2),[3] each representing several sounds because the moutshape disambiguates them. The side position is also used for single consonant. The handshape 5 is used for single vowel. There is also a neutral handshape and a neutral position, both used during silences.

In order to facilitate corpus annotation and automatic processing, each position and handshape is assigned a symbolic label, following conventions from prior work on automatic key generation for French CS (Bigi, 2023, 2025a; Lancien and Bigi, 2025). Handshapes are labeled using numbers as shown in Figure 1, and positions with lowercase letters as shown in Figure 2.

Only a few studies have investigated the development of technologies for the generation and learning CS keys. The Swiss A Capella Foundation developed the Text2LPC system,[4] a French-focused online tool that provides text-to-LPC conversion. To the best of our knowledge, no scientific publication describing its design or evaluation is currently available. Piquard-Kipffer (2016) presented a digital album that uses a 3D avatar as narrator in French CS for children with language difficulties and learning challenges between 3 and 12 years. The texts of the stories told by the talking and coding head are organized into several levels of linguistic complexity, based on lexical, morphological, syntactic, and semantic properties. The author only gives a few examples of sentences and complexity criteria without details. Sankar (2024) also worked on an automatic system for recognition and generation of French CS. To our knowledge, these systems are not available. However, Bigi (2025a) has proposed AutoCS in TextCueS, an open source automated system for CS key generation, also available online for French and American English.[5]

The corpus and resources for learning to cue

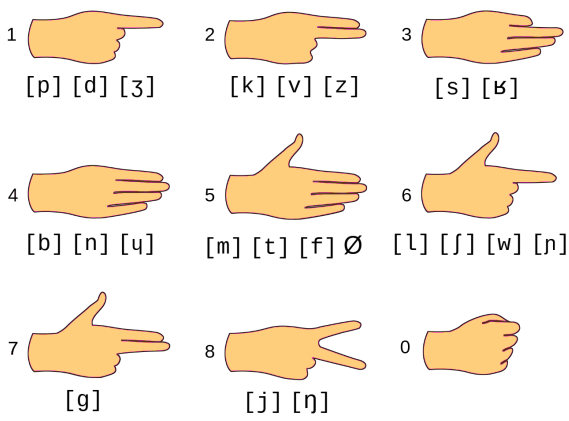


Figure 1: Handshapes representing consonants


Figure 2: Positions representing vowels

[1] https://alpc.asso.fr

[2] http://www.lpcbelgique.be

[3] For example, the fourth figure can represent the words or sounds in '*pain*' (bread), '*peu*' (few), '*deux*' (two), '*jeu*' (game), etc. because the handshape 1 is used on the cheekbone, the mouthshape makes the distinction.

[4] https://text2lpc.a-capella.ch

[5] https://auto-cuedspeech.org/textcues.html

are also scarce. Recently, Sankar (2024) built corpora in French CS, but without annotation. Another corpus is the CLeLfPC, available online on the ORTOLANG platform (Bigi et al., 2022).[6] It contains 4 hours of audio/video recordings, partly annotated. People who want to learn French CS use various resources that are not always easily available: games with speech therapists (Artaz, 2011; Olhagaray, 2013), learning CDs, short videos (e.g. some can be found on various websites and YouTube), books (Sabbagh, 2012a,b), or courses with associations.

## 3. The VizLector Project

The aim of the VizLector project is the creation of freely accessible resources in French CS. An online learning platform will be deployed, including video supports with humans enabling to practice and learn French CS for different audiences (e.g. children or adults with hearing impairment and their family), thus bridging the gap in resources for learners. Producing CS videos is both time-consuming and costly, as it requires the involvement of trained human cuers. Video recordings of people reading texts are processed with SPPAS (Bigi, 2015),[7] together with its AutoCS spin-off (Bigi, 2025a).[8] Using the video signal, the associated audio track, and an orthographic transcription, SPPAS automatically performs text normalization, phonetic conversion, forced alignment, and CS cue generation (keys, timing, hand positions/angles, and coordinates); the resulting hand cues are then overlaid onto the video. It can also generate the CS keys from a written text for French and American English. For written texts, the processing pipeline includes normalization, phonetization, and CS key generation (Gala et al., 2024; Bigi, 2023).

## 4. The Alector Corpus

For the online learning platform, it is necessary to have a written corpus with the view of automatically generating it in French CS. The 200 texts from the French Alector corpus (Gala et al., 2020)[9] were annotated and analyzed. It contains 100 original texts and their corresponding 100 adapted versions (Table 1), manually simplified by humans following the simplification guidelines developed in their project. Half of them are in the literary genre, while the other half are scientific. These texts are intended for children between 7 and 11 years (2nd to 5th grade). There are 5 text levels: IReST, CE1 (2nd), CE2 (3rd), CM1 (4th), and CM2 (5th). The first group is a set of 10 standardized, easy texts usually used for assessment of reading performances from the French version of the IReST corpus (Vital-Durand, 2011). There is a disproportion of scientific texts compared to literary texts for this level, even though there are fewer texts. We will investigate whether the level of the texts is correlated with the variables that will be extracted.

| | **Original** | | **Simplified** | | **Total** |
|---|---|---|---|---|---|
| | lit | sci | lit | sci | |
| IReST | 1 | 9 | 1 | 9 | 20 |
| CE1 | 15 | 10 | 15 | 10 | 50 |
| CE2 | 14 | 10 | 14 | 10 | 48 |
| CM1 | 10 | 10 | 10 | 10 | 40 |
| CM2 | 10 | 11 | 10 | 11 | 42 |
| **Total** | 50 | 50 | 50 | 50 | 200 |

Table 1: Number of texts per level, type and genre

## 5. Features Extraction

In order to develop an automatic text classification model to predict the complexity level, relevant text representations need to be extracted. In this study, two types of information were combined: readability characteristics and phonemic features. For the automatic extraction of readability features (Section 5.1), the API of FABRA (Wilkens et al., 2022) was used. FABRA is a readability toolkit that offers a large number of readability predictors for French.[10] It allows to calculate 509 language variables, most of which can be represented through 18 statistical aggregators (e.g., sum, median, average, variance, etc.). This tool can be used, for example, for corpus analysis related to readability, text simplification, automatic genre identification, etc. SPPAS v.4.29 was used for annotation of phonemic features (Section 5.2). Finally, the two types of features were combined (Section 5.3).

### 5.1. Readability Features

In FABRA, 30 variables were selected: 3 variables based on length (section 5.1.1), 23 lexical variables (section 5.1.2), and 5 syntactic ones (section 5.1.3). Note that 1 length-based variable not from FABRA has been added for correlations. The full list of variables used is shown in Appendix A (Table 7). It should also be mentioned that the average was used as aggregator in FABRA, except for LEXdvr-FLC, that has no underlying distribution and therefore is scalar.

[6] https://hdl.handle.net/11403/clelfpc
[7] https://sppas.org
[8] https://auto-cuedspeech.org
[9] The corpus is available on demand (https://corpusalector.huma-num.fr).
[10] https://cental.uclouvain.be/fabra

#### 5.1.1. Length-based Variables

Length-based variables were historically the first to be used in readability to assess the difficulty of a text (Flesch, 1948; Kandel and Moles, 1958). They are still employed, despite their lack of causal relation with reading difficulty.

**Word length** (1 variable) and **Sentence length** (1). The number of syllables per word (LENwrdSYL), and the number of tokens per sentence (LENsntWRD) were the only length-based variables used from FABRA because the others are based on letter count, which is less relevant for CS keys (phoneme-based). More syllables and tokens require generating more CS keys, potentially resulting in more cognitive load.

**Text length** (1). The non-normalized length of texts (not directly output from FABRA, but reused from Alector) was added, i.e. the total number of words per text without punctuation (LENtxtWRD).

All length-based variables are used to perform correlation analyses with phonemic features.

#### 5.1.2. Lexical Variables

Several families of lexical variables are tested.

**Content overlap** (1 variable). It measures repetitions of lemmas (LEXcovLGAL), we can assume that the text will be easier to (de)code because more easier to predict.[11]

**Lexical diversity** (1). The CTTR (Corrected Type Token-Ratio) is linked to the hapax legomena (LEXdvrFLC), the words that only occur once in a document, about 40-60% of a text (Kornai, 2007). These rare words could be more complex to (de)code. They are also longer because they are included in open word classes (nouns, verbs, etc.).

**Lexical frequency** (4). These variables captures the frequency of lexical words in the text, based on the lexical databases of CHILDES (LEXfrqCCS), FLELex (LEXfrqFCL), and Lexique3 (LEXfrqLCL). In addition, LEXfrqLWL considers of words in the text. More frequent words are likely to be known by a greater number of people than other words even if they do not necessarily belong to the informal language register (e.g. register used in family or classroom).

**Graded lexicons** (10). As with the previous variables, we can assume that texts with words assigned to the CEFR levels A1 or A2 (LEXgrd[BA1/BA2/FA1/FA2] in two lexicons) may be easier to (de)code than texts with words of level B1, B2 (LEXgrd[BB1/BB2/FB1/FB2]) or C1 and C2 (LEXgrd[FC1/FC2]). These measures are based on the Beacco's French Reference Level Descriptors and the FLELex lexicon.

**Lexical norms** (4). Psycholinguistic features are included, such as the age of acquisition of each word (LEXnrmAOA), word level of concreteness (LEXnrmCNCR), word familiarity or subjective frequency (LEXnrmFAM), and word imageability (LEXnrmIMG).

**Lexical sophistication** (3). The lexical sophistication is measured as the proportion of words in the text belonging to the first frequency bands of 1,000 words of the three following frequency lists: CHILDES (LEXsopCK1), Gougenheim vocabulary list (LEXsopGK1), and Lexique3 (LEXsopLWK1).

#### 5.1.3. Syntactic Variables

**Language development** (5 variables). Since deaf students sometimes have difficulty identifying verbs in a sentence (Leitao et al., 2021), the number of words before the main verb (SYNdevBFR) is calculated. Several structural features are included: the number of constituents/phrases (SYNdevNPHRS), as well as those of the internal conjugate clause type (SYNdevNPRSSINT), of the relative clause type (SYNdevNPRSSREL), and of the subordinate clause type (SYNdevNPRSSSUB). The stories told in French CS with the avatar from the digital album by Piquard-Kipffer (2016) were also classified by level of syntactic complexity, one criterion being the greater number of complex sentences.

### 5.2. Phonemic Features

In total, 9 phonemic features were automatically extracted, and 1 manually annotated due to a lack of annotation tools. All phonemic variables were normalized as explained in Appendix A (Table 8).

**CS key frequency** (4 variables). As with Gala et al. (2024), the Alector corpus was automatically annotated with SPPAS to get the number of CS keys (PHOkey). With SPPAS, it is possible to compute the proportion of the different types of transition between face positions. We assume that transitions between more distant positions would be more complicated, such as side compared to throat and conversely (PHOpositionST) or cheekbone compared to throat and conversely (PHOpositionBT). As regards the structure of CS keys, they have three possible configurations: C, V, or CV. The proportion of CV clusters (PHOclusterCV) is automatically extracted. The CV seems to be the most complex of the three to (de)code. We assume that it is because it is necessary to consider the position of the vowel. MarsaTag (Rauzy et al., 2014) has been used to get the Part-of-Speech (POS) and observe if there are links between the most (or least) frequent tags and the number of CS keys.

[11] This was the case, for example, during the French CS training of one of the authors of the paper when it was necessary to code children's nursery rhymes containing repetitions. The participants found it easier to code.

**Phonemic frequency** (6). Some phonological phenomena that have an impact on the lack of phoneme-grapheme consistency – such as glides, liaisons, etc. – may seem complex to (de)code (Gala et al., 2024). Therefore, the proportion of each of the three French glides (see Table 2) was computed, i.e., the semi-consonants or semi-vowels w, ɥ, j (PHOglideW, PHOglideH, PHOglideJ). The total number of glides normalized by the number of CS keys was also considered (PHOglide). In addition, one author of the paper annotated the obligatory liaisons based on the previous study (Gala et al., 2024), their proportion was computed (PHOliaison). Liaison is a phenomenon where an orthographically-final consonant is mute except in certain environments (Table 3), i.e., when it precedes a vowel, a mute h or a glide. To our knowledge, there is no sufficiently reliable automatic tool for annotating French liaisons, in part due to the arbitrary application of liaison rules. Finally, the number of Consonant-Consonant (C-C) clusters in a same word (PHOclusterCC) was also extracted. We then assume that CS key splitting can be more complex if there is a double consonant.

| Pho. | Example | Phonemes (keys) |
|---|---|---|
| w | ***oi**seau* (bird) | wa.zo (**6**s.2s) |
| ɥ | *n**ui**t* (night) | n.ɥi (4s.**4**m) |
| j | *feu**ille*** (leaf) | fœ.j (5s.**8**s) |

Table 2: Examples of glides in French

| Pho. | Example | Phonemes (keys) |
|---|---|---|
| z | *le**s** autres* (the others) | le.zo.t.ʁə (6t.**2s**.5s.3s) |
| t | *tou**t** à coup* (suddenly) | tu.ta.ku (5c.**5s**.2c) |
| n | *u**n** été* (one summer) | œ̃.ne.te (5t.**4t**.5t) |
| ʁ | *dernie**r** étage* (top floor) | dɛ.ʁ.n.je.ʁe.ta.ʒ (1c. 3s.4s.8t.**3t**.5s.1s) |
| p | *tro**p** agressif* (too aggressive) | t.ʁɔ.pa.g.ʁɛ.si.f (5s. 3s.**1s**.7s.3c.3m.5s) |

Table 3: Examples of obligatory liaisons in French

### 5.3. Combining the Features

As previously mentioned, there is a lack of available training data on French CS comprehension collected from the deaf population. Therefore, unsupervised model had to be used to automatically classify texts by level of complexity. As a first step before clustering, feature selection was performed using the Minimum Redundancy Maximum Relevance (MRMR) algorithm (Ding and Peng, 2005). This method aims to identify a subset of variables that are maximally informative with respect to the target variable while minimizing redundancy among the selected features. The target variable was the text level encoded numerically to preserve its inherent ordinal nature. In v.4.3.3 of R (R Core Team, 2021), the mRMRe package (v.2.1.2.2) was used (De Jay et al., 2013), as well as the NbClust package (v.3.0.1) for determining the relevant number of clusters (Charrad et al., 2014), before applying the K-means algorithm.

## 6. Results

Readability features are analyzed first, then phonemic features, and finally both.

### 6.1. Readability Features

The number of words per text (LENtxtWRD) increases monotonically across levels of texts. Median values rose from IReST (Mdn = 127.5) to CE1 (Mdn = 244), CE2 (Mdn = 307.5), CM1 (Mdn = 370), and CM2 (Mdn = 513.5). The (average) number of syllables per word (LENwrdSYL) shows a small but consistent increase across levels. Median values ranged from 1.41 in IReST to 1.44 in CM2, with substantial overlap between distributions. Sentence length (LENsntWRD) showed moderate variation across levels. While median values increased slightly from CE1 (12.24 words) to CM2 (14.07 words), distributions largely overlapped, indicating that sentence length contributed only modestly to level differentiation.

The correlations of readability features were calculated (Figure 3). The lexical variables, especially the proportions of words in graded lexicons are strongly correlated with each other (in yellow). There are also extremely strong negative correlations (in dark blue) between the numbers of complex clauses and the proportions of words in graded lexicons. Logically, the results show a perfect positive correlation (in yellow) between the number of tokens per sentence and the number of constituents/phrases (correlation of 1). There is a strong positive correlation (in light green) between the number of words per text and the measure of lexical diversity (correlation of 0.74).

### 6.2. Phonemic Features

A total of 129,608 CS keys are obtained, compared with 91,786 keys reported by Gala et al. (2024) on the same Alector corpus. In that study, sentences containing more than 100 phonemes were excluded; this constraint is not applied here.

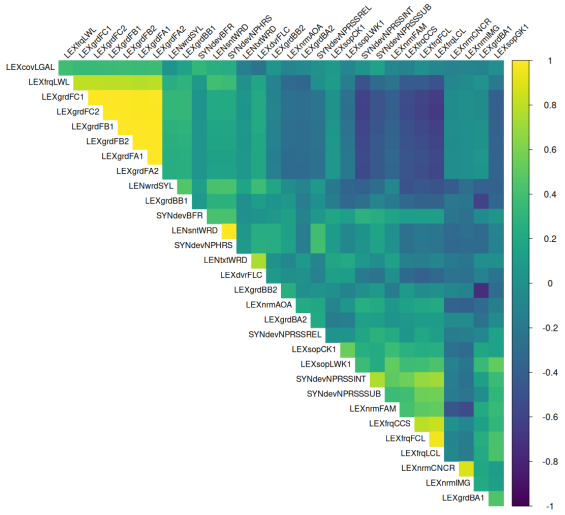


Figure 3: Correlation heatmap among the readability features on the Alector texts

| Position | # | % | Vowels |
|---|---|---|---|
| Side (s) | 65,625 | 50.60 | a o œ ə ⊕ |
| Chin (c) | 22,869 | 17.60 | ɛ u ɔ |
| Mouth (m) | 22,354 | 17.20 | i ɑ̃ ɔ̃ |
| Throat (t) | 15,523 | 12.00 | e y œ̃ |
| Cheekbone (b) | 3,237 | 2.50 | ø ɛ̃ |

Table 4: Total number of keys by position (vowels) in the CS annotated Alector corpus

| Handshape | # | % | Conson. |
|---|---|---|---|
| 5 | 30,596 | 23.61 | m t f ⊕ |
| 3 | 30,135 | 23.25 | s ʁ |
| 1 | 20,359 | 15.70 | p d ʒ |
| 6 | 19,673 | 15.17 | l ʃ w ɲ |
| 2 | 14,289 | 11.02 | k v z |
| 4 | 9,043 | 6.97 | b n ɥ |
| 8 | 3,941 | 3.04 | j ŋ |
| 7 | 1,572 | 1.21 | g |

Table 5: Total number of keys by handshape (consonants) in the CS annotated Alector corpus

Among those 129,608 CS keys, the most frequent position for vowels is the side of the face (Table 4), as it is the placement which has the most vowels and the most frequent ones in French, it also codes the silent -e as well as the absence of vowels – e.g. the single consonant t, 5**s** in '*les autres*' (cf. Table 3).

In some French CS training programs, we start by learning all the positions of the vowels in the first lesson in order to be able to gradually learn the handshapes for the consonants: one to two per lesson by combining them with all the vowels in short words, then longer and longer ones, which use the handshapes seen in the previous lessons.

As regards the distribution of handshapes, the most frequent one is the shape 5 (Table 5), which includes many coding possibilities (along with shape 6), and it is the one learned first in training (it is the shape that allows the easiest transition to the others – the open palm –, and which codes the single vowel as '*un*', **5**t in Table 3). Conversely, shape 7 only includes the sound g (rare in French) and is the only shape with one consonant). It is always the least frequent handshape (the second to last shape seen in training). The side position and the shape 5 both encode two different key structures (V and CV clusters or C and CV clusters).

Although the handshape 8 is second to last in terms of frequency, it represents the most frequent glide j (3,940 occurrences) – followed by w (1,812), then ɥ (952). The other consonant of shape 8 (ŋ) is rare in French and typical of foreign words ('*parki**ng***', '*joggi**ng***', etc.). In Alector, there is a single instance of this glide: the English proper noun 'Grunni**ng**s' in an excerpt from Harry Potter book. Conversely, the second most frequent handshape is 3, which includes the two most commonly used consonants in French. We assume that this shape is probably one of the most complex to make for beginners, unlike 5, which is the open palm. The numbers between some positions and handshapes are extremely close. The number of positions per text is compared using ANOVA: for chin and mouth, the difference is no significant ($p$ = 0.16), as for shapes 5 and 3 ($p$ = 0.21). In contrast, the difference between shapes 1 and 6 is significant ($p < 0.05$).

As the number of words per text (LENtxtWRD), the total number of CS keys per text (unnormalized PHOkey) increases markedly across text levels. Median values rose from IReST (Mdn = 249.5) to CE1 (Mdn = 454), CE2 (Mdn = 597.5), CM1 (Mdn = 728), and CM2 (Mdn = 1,002). The text with the most CS keys is a CM2 text (id_189 with 1,352 keys). Conversely, the one with the fewest keys is an IReST text (id_51 with 203 keys).

For the three possible CS key configurations, there are 66.36% of CV clusters (PHOclusterCV), and respectively 25.79% and 7.83% of C and V clusters. These ratios are close to those of a corpus of 4,143 French CS keys produced by experienced cuers on read texts (Bigi, 2023), i.e., 70.72% for CV, 20.69% for C and 8.59% for V.

The most frequent POS tags are: nouns, determiners, and verbs. The most frequent lemmas are almost all invariable (closed word classes, such as determiners, prepositions, pronouns), and are small words: one or two CS keys for '*il*' (he), or the forms of '*être*' (to be) and '*avoir*' (to have). Hapax words are indeed nouns, verbs, adjectives; they are longer words (with more CS keys), which we

assume to be more complex to (de)code due to cognitive load.

The number of transitions between the side from/to throat positions (PHOpositionST) is much more frequent than between the cheekbone from/to throat positions (PHOpositionBT), respectively with 14,651 and only 665 occurrences in the corpus. Note that the distance between side from/to throat positions is not necessarily the greatest because the possible space for these peripheral positions is large (Bigi, 2025b).

Phonemic variables are generally not strongly correlated with each other (Figure 4), except (in light green) between the number of all glides and the number of each glide (j, w, ɥ) – respectively correlations of 0.78, 0.37, 0.26 –, and between the number of CS keys and the number of syllables per word (correlation of 0.58).

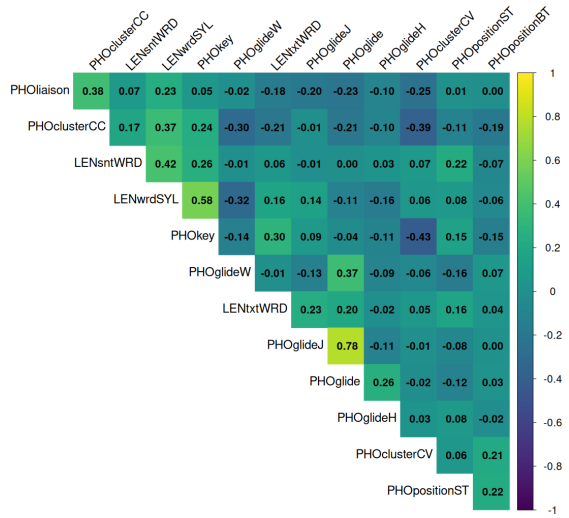


Figure 4: Correlation heatmap among the length-based variables and the phonemic variables on the Alector texts

### 6.3. Combining the Features

A full correlation heatmap shows that readability variables are not strongly correlated with phonemic variables as shown in Appendix B (Figure 6).

The two types of features were extracted to automatically classify texts of Alector according to their level of complexity. In total, 10 variables were automatically selected by MRMR as they provided an optimal trade-off between explanatory power and model parsimony: LENtxtWRD, LENwrdSYL, LEXdvrFLC, LEXgrdFA1, LEXnrmIMG, LEXsopCK1, SYNdevNPRSSREL, PHOkey, PHOpositionST, PHOglideJ; i.e., 7 readability variables (2 length-based, 4 lexical, 1 syntactic), and 3 phonemic variables. There are one or two variables per family in each type even if not all families are represented. In order to assess redundancy among the selected features, pairwise correlations with text level were examined (Table 6). Overall, the selected feature set reflects a balance between highly informative predictors (LENtxtWRD, LEXdvrFLC) and complementary linguistic indicators (PHOkey, PHOglideJ, LENwrdSYL, LEXgrdFA1), ensuring both predictive relevance and reduced redundancy. This supports the use of the selected variables for downstream clustering analyses.

| # | Feature | Corr. with level |
|---|---|---|
| 1 | LENtxtWRD | 0.907 |
| 2 | LEXdvrFLC | 0.712 |
| 3 | PHOkey | 0.429 |
| 4 | PHOglideJ | 0.246 |
| 5 | LENwrdSYL | 0.238 |
| 6 | LEXgrdFA1 | 0.208 |
| 7 | LEXsopCK1 | 0.160 |
| 8 | PHOpositionST | 0.145 |
| 9 | SYNdevNPRSSREL | -0.0118 |
| 10 | LEXnrmIMG | -0.286 |

Table 6: Selected features and their correlation with the text level

The relevant number of clusters suggested by the K-means algorithm applied on our data is 3. The distribution of text levels across clusters reveals a clear gradient of text difficulty (Figure 5). Cluster 1 is dominated by beginner level texts (IReST, CE1), whereas cluster 3 is enriched in advanced level texts (CM1, CM2). Cluster 2 shows a more heterogeneous composition, encompassing primarily intermediate levels texts (CE1, CE2, CM1). Although cluster boundaries do not perfectly align with the predefined text levels, the observed distribution suggests that the clustering captures latent dimensions of complexity. We proposed an initial model for selecting texts to be recorded in CS based on different features. However, the target variable is the text level. These levels are intended for children learning French, but not French CS. This calls into question whether the model actually captures French CS learning difficulty. We acknowledge this limitation. Further research is needed to verify the impact of the text level.

## 7. Conclusion and Discussion

This paper presented the extraction of textual and phonemic features to automatically classify texts for French CS and reading training according to their complexity level. The creation of freely accessible resources in CS addresses an important societal need aimed to improving the inclusion and comprehension of people with hearing loss.

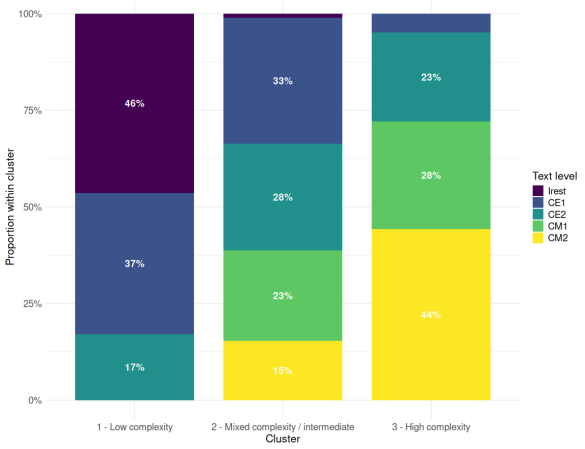


Figure 5: Proportional distribution of text levels within each cluster

In total, 41 variables potentially linked to CS complexity were inspected. Several variable types were considered, including both textual- and phoneme-level measures, which are rarely examined jointly. Because the literature on complexity variables in French CS remains limited, some hypotheses cannot yet be verified within the scope of this paper. The results showed positive correlations between readability variables, especially the length-based and lexical. The phonemic features are poorly correlated with each other and with readability variables, except the number of CS keys, glides or syllables per words. The analysis of key frequency appears consistent with previous studies (Bigi, 2023; Gala et al., 2024), with reported French syllable frequencies, and with the key learning order commonly taught; further surveys of the target audience are needed to better document current CS teaching practices.

Modeling complexity in terms of readability or listenability traditionally relies on the use of data evaluated by the target audience, which we didn't have. Therefore, we proposed an initial exploratory unsupervised model to classify texts that will be generated in French CS. We identified, extracted and combined readability and phonemic features. Future work includes choosing texts and words from Alector corpus to be recorded. The videos will then be annotated with a CS coding hand using the SPPAS tool, before being evaluated by deaf or hard-of-hearing CS users. The learning complexity of the recorded texts will be assessed to refine the features. This evaluation will allow to propose customization options for users of the learning platform, such as adding specific filters (by CS keys, text length, with or without glides, etc.).

Additional factors may also contribute to CS complexity and deserve further investigation, including fluency-related measures – e.g., syllables per minute, pause duration, pauses per minute as in Ozawa et al. (2024) –, and learner-specific features. Further work could also incorporate phoneme-grapheme consistency criteria that may complicate (de)coding in French CS, such as graphemes with context-dependent pronunciations governed by more or less regular rules.

## Lay Summary

In this work, we analyze a corpus of texts that will be translated into Cued Speech. Cued Speech is used with deaf and hard-of-hearing people to improve their understanding of spoken language. This communication mode combines gestures and speech. There are several hand movements in different positions around the face to represent the sounds. Although this method is known to be helpful, there are still not many tools available to learn Cued Speech, especially in French. In order to create learning videos, we propose to extract and analyze 41 textual and phonemic features that might be more complex in French Cued Speech. We mainly used two tools: FABRA and SPPAS. The results show some correlations. A first model is proposed for selecting texts to be recorded for learning French Cued Speech.

## Acknowledgments

This work received support from the French government under the France 2030 investment plan, as part of the *Initiative d'Excellence d'Aix Marseille Université* - AMIDEX (AMX-22-RE-AB-022), VizLector project. It is also a part of the Automatic Cued Speech project (APa2022_022), supported by FIRAH (*Fondation Internationale de la Recherche Appliquée sur le Handicap*, International Foundation of Applied Disability Research). Finally, the authors thank the reviewers of the first version of the paper and the members of the French ALPC association for the LfPC lessons.

## 8. Bibliographical References

Mélody Artaz. 2011. L'apprentissage de la Langue française Parlée Completée par les enfants sourds d'âge scolaire : Conception d'un loto LPC. Master's thesis, Université Stendhal.

Brigitte Bigi. 2015. SPPAS - Multi-Lingual Approaches to the Automatic Annotation of Speech. *The Phonetician. Journal of the International Society of Phonetic Sciences*, 111:54–69.

Brigitte Bigi. 2023. An analysis of produced versus predicted French Cued Speech keys. In *10th*

*Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 24–28, Poznań, Poland.

Brigitte Bigi. 2025a. Bridging the Gap: Design and Evaluation of an Automated System for French Cued Speech. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 8–18, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Brigitte Bigi. 2025b. Spatial Analysis of Hand Positions in French Cued Speech (LfPC). In *16th International Conference on Linguistic Research and Applications*, Paris, France.

Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6):1–36.

R. Orin Cornett. 1967. Cued Speech. *American Annals of the Deaf*, 112(1):3–13.

Nicolas De Jay, Simon Papillon-Cavanagh, Catharina Olsen, Nehme El-Hachem, Gianluca Bontempi, and Benjamin Haibe-Kains. 2013. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics*, 29(18):2365–2368.

Chris Ding and Hanchuan Peng. 2005. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205.

Rudolph Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Núria Gala, Brigitte Bigi, and Marie Bauer. 2024. Automatically Estimating Textual and Phonemic Complexity for Cued Speech: How to See the Sounds from French Texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1817–1824, Torino, Italia. ELRA and ICCL.

Ludivine Javourey-Drevet, Stéphane Dufau, Thomas François, Núria Gala, Jacques Ginestié, and Johannes C. Ziegler. 2022. Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of French. *Applied Psycholinguistics*, 43(2):485–512.

Elaine Jones, Robert Strom, and Susan Daniels. 1989. Evaluating the Success of Deaf Parents. *American Annals of the Deaf*, 134(5):312–316.

Liliane Kandel and Abraham Moles. 1958. Application de l'indice de Flesch à la langue française. *Cahiers Etudes de Radio-Télévision*, 19:253–274.

András Kornai. 2007. *Mathematical Linguistics*. Springer Science & Business Media.

Katsunori Kotani, Shota Ueda, Takehiko Yoshimi, and Hiroaki Nanjo. 2014. A Listenability Measuring Method for an Adaptive Computer-assisted Language Learning and Teaching System. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 387–394.

Katsunori Kotani and Takehiko Yoshimi. 2017. Effectiveness of Linguistic and Learner Features for Listenability Measurement Using a Decision Tree Classifier. *The Journal of Information and Systems in Education*, 16(1):7–11.

Mélanie Lancien and Brigitte Bigi. 2025. French Cued Speech rhythm: first findings on the relationship between hand position and segments' duration. In *11th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science, Linguistics and Low Resourced Languages*, Poznań, Poland.

Manuel Leitao, Elodie Venti, Thomas Sigiez, Christophe Laroche, Marie Perini, and Agnès Piquard-Kipffer. 2021. Projet LogilecSur: quelles stratégies enseignantes pour guider des élèves sourds vers l'autonomie en compréhension écrite? In *IDEKI 2021 - 4ème colloque international Didactiques et métiers de l'humain*, Pont-à-Mousson, France.

Jacqueline Leybaert and Carol J. LaSasso. 2010. Cued Speech for Enhancing Speech Perception and First Language Development of Children With Cochlear Implants. *Trends in Amplification*, 14(2):96–112.

Aïnizé Olhagaray. 2013. « Le Pirate Codeur » : Élaboration d'un matériel ludique visant à entraîner et automatiser le décodage de la Langue française Parlée Complétée (LPC) : à destination des enfants sourds ayant reçu une introduction tardive du code LPC. Master's thesis, Université Bordeaux Segalen.

Lucía Ormaechea and Nikos Tsourakis. 2024. Automatic text simplification for French: model fine-tuning for simplicity assessment and simpler text generation. *International Journal of Speech Technology*, 27(4):957–976.

Minami Ozawa, Rodrigo Wilkens, Kaori Sugiyama, and Thomas François. 2024. Modéliser la facilité d'écoute en FLE: vaut-il mieux lire la transcription ou écouter le signal vocal ? In

*35èmes Journées d'études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)*, volume 1, pages 549–566. ATALA & AFPC.

Agnès Piquard-Kipffer. 2016. Un album numérique pour raconter une histoire avec un avatar narrateur. In *XVIèmes rencontres internationales en orthophonie - Orthophonie et technologies innovantes*, Paris, France.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.

Stéphane Rauzy, Grégoire Montcheuil, and Philippe Blache. 2014. MarsaTag, a tagger for French written texts and speech transcriptions. In *Second Asian Pacific Corpus linguistics Conference*, pages 220–220, Hong Kong, China.

Valérie Sabbagh. 2012a. *Le Petit Clown 2 Le LPC pour les enfants : Entraînement et perfectionnement*. ALPC, Paris, France.

Valérie Sabbagh. 2012b. *Le Petit Clown Le LPC pour les enfants : L'imagier d'Agathe*. ALPC, Paris, France.

Sanjana Sankar. 2024. *Automatic recognition and generation of French Cued Speech using deep learning*. Ph.D. thesis, Université Grenoble Alpes.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P Yancey, and Thomas François. 2022. FABRA: French Aggregator-Based Readability Assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France. ELRA.

World Health Organization. 2021. *World Report on Hearing*. World Health Organization. ISBN: 978-92-4-002048-1.

## 9. Language Resource References

Brigitte Bigi, Maryvonne Zimmermann, and Carine André. 2022. CLeLfPC: a Large Open Multi-Speaker Corpus of French Cued Speech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 987–994, Marseille, France. ELRA.

Gala, Núria and Tack, Anaïs and Javourey-Drevet, Ludivine and François, Thomas and Ziegler, Johannes C. 2020. *Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers*. ELRA.

Vital-Durand, François. 2011. *International Reading Speed Texts IResT (French version)*.

## A. Overview of all the Variables

All details and references on readability variables from FABRA can be found in Wilkens et al. (2022).

## B. Correlation Matrix

Correlations among all variables are shown as percentages for readability (Figure 6).

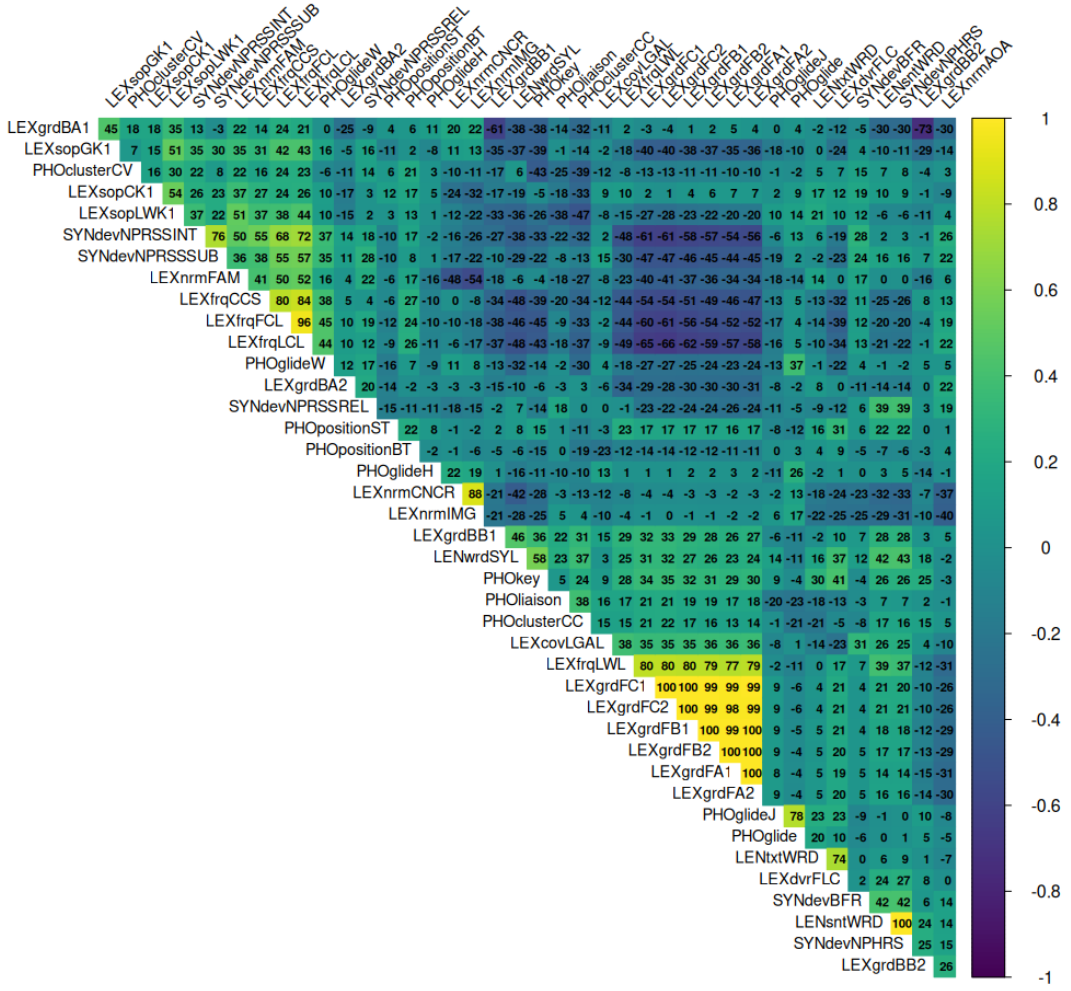


Figure 6: Correlation heatmap among the readability and phonemic features on the Alector texts

| | Variable | Description |
|---|---|---|
| **Length based** | | |
| Word length | LENwrdSYL | Number of syllables per word |
| Sentence length | LENsntWRD | Number of tokens per sentence, excluding punctuation |
| Text length | LENtxtWRD | Number of words per text, excluding punctuation; unnormalized |
| **Lexical Variables** | | |
| Content overlap | LEXcovLGAL | Any lemma is shared in any sentences |
| Lexical diversity | LEXdvrFLC | CTTR of all types of lemma forms of nouns, proper nouns, verbs, adjectives and adverbs in the text, considering all tokens |
| Lexical frequency | LEXfrqCCS | Frequency of surface form of all nouns, proper nouns, verbs, adjectives and adverbs based on the occurrence at CHILDES corpus |
| Lexical frequency | LEXfrqFCL | Frequency of lemma form of all nouns, proper nouns, verbs, adjectives and adverbs based on the occurrence at FLELex corpus |
| Lexical frequency | LEXfrqLCL | Frequency of lemma form of all nouns, proper nouns, verbs, adjectives and adverbs based on the occurrence at Lexique3 corpus |
| Lexical frequency | LEXfrqLWL | Frequency of lemma form of all words based on the occurrence at Lexique3 corpus |
| Graded lexicons | LEXgrd[BA1/BA2/BB1/BB2] | Proportion of words in Beacco's French Reference Level Descriptors for each CEFR level (A1 to B2) |
| Graded lexicons | LEXgrd[FA1/FA2/FB1/FB2 FC1/FC2] | Frequency of words in FLELex resource for each CEFR level (A1 to C2) |
| Lexical norms | LEXnrmAOA | Age of acquisition of each word |
| Lexical norms | LEXnrmCNCR | Words level of concreteness |
| Lexical norms | LEXnrmFAM | Words familiarity, also called subjective frequency |
| Lexical norms | LEXnrmIMG | Imageability of each word |
| Lexical sophistication | LEXsopCK1 | Number of words in the first frequency bands of 1,000 words of CHILDES |
| Lexical sophistication | LEXsopGK1 | Number of words in the first frequency bands of 1,000 words of Gougenheim vocabulary list |
| Lexical sophistication | LEXsopLWK1 | Number of surface form words in the first frequency bands of 1,000 words of Lexique3 |
| **Syntactic Variables** | | |
| Language development | SYNdevBFR | Number of words before the main verb |
| Language development | SYNdevNPHRS | Number of constituents/phrases |
| Language development | SYNdevNPRSSINT | Number of different types of phrases/ constituents of type SINT (internal conjugate clause - *proposition conjuguée interne*) |
| Language development | SYNdevNPRSSREL | Number of different types of phrases/ constituents of type SREL (relative clause - *proposition relative*) |
| Language development | SYNdevNPRSSSUB | Number of different types of phrases/ constituents of type SSUB (subordinate clause - *proposition subordonnée*) |

Table 7: Readability variables description

| | Variable | Description |
|---|---|---|
| CS key frequency | PHOkey | Number of CS keys, (un)normalized per character |
| CS key frequency | PHOpositionST | Number of transitions between side and throat positions (and conversely); normalized per PHOkey-1 |
| CS key frequency | PHOpositionBT | Number of transitions between cheekbone and throat positions (and conversely); normalized per PHOkey-1 |
| CS key frequency | PHOclusterCV | Number of consonant/vowel clusters; normalized per PHOkey |
| Phonemic frequency | PHOglideW | Number of glides (semi-consonants or semi-vowels) $\mathrm{w}$; normalized per PHOkey |
| Phonemic frequency | PHOglideH | Number of glides (semi-consonants or semi-vowels) $\mathrm{ɥ}$; normalized per PHOkey |
| Phonemic frequency | PHOglideJ | Number of glides (semi-consonants or semi-vowels) $\mathrm{j}$; normalized per PHOkey |
| Phonemic frequency | PHOglide | Number of glides (semi-consonants or semi-vowels) $\mathrm{w}$, $\mathrm{ɥ}$, $\mathrm{j}$; normalized per PHOkey |
| Phonemic frequency | PHOliaison | Number of obligatory liaisons; normalized per LENtxtWRD-1 |
| Phonemic frequency | PHOclusterCC | Number of double consonants; normalized per LENtxtWRD |

Table 8: Phonemic variables description