LEXIQUE PHONÉTIQUE ET GRAMMATICAL DU FRANÇAIS

LeXiK

Guide d'utilisation

Version 1.0, octobre 2026 (base LeXiK 1.2.4 du 15/10/2026 11:58).

Maxime CARRIERE

Lectura — lectura.world
lectura.world@gmail.com

Licence Creative Commons Attribution – Partage dans les mêmes conditions 4.0 (CC BY-SA 4.0)

Sommaire

1. Qu'est-ce que LeXiK . . . **3**
1.1 En bref . . . 3
1.2 Quatre sous-lexiques . . . 3
1.3 Ce qui le distingue . . . 4

2. Prise en main . . . **4**
2.1 Contenu du dépôt . . . 4
2.2 Ouvrir les fichiers . . . 4
2.3 Lire une ligne . . . 5
2.4 Quelques cas limites . . . 6

3. Les quatre sous-lexiques . . . **6**
3.1 Mots . . . 6
3.2 Locutions et phrases (MultiMots) . . . 7
3.3 Formes spéciales (Spéciaux) . . . 7
3.4 Formes non classées (Non classé) . . . 8

4. Les fichiers de formes (lexik_formes_*.tsv) . . . **8**

5. Le fichier des lemmes (lexik_lemmes.tsv) . . . **9**

6. La prononciation . . . **10**
6.1 Les phonèmes . . . 10
6.2 Lire la colonne phone . . . 10
6.3 Les conventions de notation . . . 11
6.4 La prononciation brute . . . 11
6.5 D'où viennent les prononciations . . . 11

7. L'alignement graphème–phonème . . . **12**
7.1 Ce que dit la colonne . . . 12
7.2 Les règles de notation . . . 12
7.3 Comment l'alignement est calculé . . . 13

8. La syllabation et l'orthocode . . . **13**
8.1 La syllabation de la prononciation . . . 13
8.2 L'orthocode . . . 13

9. L'analyse grammaticale . . . **14**
9.1 La catégorie (upos) . . . 14
9.2 Le code MULTEXT (multext) . . . 14

10. Les fréquences . . . **15**
10.1 Les trois corpus . . . 15
10.2 La fréquence d'une graphie (freq_*) . . . 15
10.3 La part de chaque analyse (parts_*) . . . 15
10.4 La fréquence d'une locution . . . 16
10.5 La part pronominale d'un verbe . . . 16
10.6 La fréquence d'une forme spéciale . . . 17

11. L'indice de confiance . . . **17**

12. Les sources . . . **17**
Modèles et algorithmes . . . 18

13. Limites connues . . . **18**

14. Format des fichiers . . . **19**

Ce guide explique ce que contient LeXiK et comment le lire. La façon dont la base a été construite (sources, arbitrage entre sources, contrôles) est décrite à part, dans `provenance/DOCUMENTATION_PROVENANCE.md`.

1. Qu'est-ce que LeXiK

1.1 En bref

LeXiK est un **lexique de formes fléchies et de leurs lemmes** pour le français. Une forme fléchie est un mot tel qu'il s'écrit dans un texte (*mangeons*, *petites*, *chevaux*) ; son lemme est la forme sous laquelle on le cherche dans un dictionnaire (*manger*, *petit*, *cheval*).

Pour chaque forme, LeXiK donne :

- sa **prononciation**, en alphabet phonétique international (API), découpée en syllabes ;
- son **analyse grammaticale** : lemme, catégorie, genre, nombre, temps, personne… ;
- son **alignement graphème-phonème** : quelles lettres écrivent quel son ;
- son **orthographe découpée en syllabes** (l'« orthocode ») ;
- ses **fréquences** dans trois corpus (sous-titres de films, Wikipédia, littérature), avec la part de chacune de ses analyses.

Pour chaque lemme, il donne :

- ses **propriétés** : genre, féminin et pluriel, groupe et auxiliaire d'un verbe, type d'un nom propre (prénom, lieu…) ;
- sa **fréquence**, somme de celles de ses formes ;
- et, par le fichier des formes, la liste de toutes ses formes fléchies avec leur prononciation.

1.2 Quatre sous-lexiques

LeXiK se compose de quatre **sous-lexiques** ; chaque entrée appartient à l'un d'eux (§3) :

Sous-lexique	Contenu	Graphies	Entrées	Lemmes
`Mots`	mots simples et composés, communs et propres	2 240 213	2 877 446	752 058
`MultiMots`	locutions et phrases	80 059	80 688	—
`Spéciaux`	sigles, abréviations, symboles, éléments	45 423	45 919	—
`Non classé`	formes qu'on n'a pas pu ranger	103 184	103 501	—

Chaque sous-lexique a son fichier (`lexik_formes_mots.tsv`, `lexik_formes_multimots.tsv`, `lexik_formes_speciaux.tsv`, `lexik_formes_non_classes.tsv` : 2 580 027 lignes en tout), avec les colonnes qui le concernent. Seuls les mots ont un fichier des lemmes, `lexik_lemmes.tsv`, relié aux formes par la colonne `id_lemme`. Les trois autres sous-lexiques ont un fichier qui se suffit à lui-même, une ligne par entrée, avec les propriétés sur la ligne. Pour une locution, une ligne désigne le lemme de la locution, avec un champ (`flechir`) qui indique les mots à fléchir ; pour une forme spéciale, une ligne correspond à sa graphie et à sa forme développée. Les fréquences sont détaillées plus loin (§3.2, §3.3 et §10).

LeXiK fera l'objet d'une publication sur ORTOLANG sous cette forme complète : les quatre sous-lexiques réunis dans un seul dépôt.

1.3 Ce qui le distingue

1. **Au-delà des mots communs : noms propres, sigles et locutions.** 415 707 entrées de noms propres (prénoms, noms de famille, communes, lieux), 45 919 formes spéciales (sigles, abréviations, unités, symboles) et 80 688 locutions et phrases, chacune avec sa prononciation.
2. **Une phonétique enrichie.** La prononciation ne se limite pas aux phonèmes : elle note ce que la forme isolée ne laisse pas deviner, la consonne de liaison (*petit* `pə.ti[t]`), le son qui change en liaison (*bon* `bɔ̃[ɔ̃>ɔn]`), le h aspiré, les variantes libres, et elle est découpée en syllabes (§6.2).
3. **Des mots absents des sources, prononcés.** 587 436 analyses qu'aucune source ne prononçait ont reçu une prononciation du modèle graphème → phonème de Lectura (confiance `G`, §6.5).
4. **Des sources corrigées.** Repérage et correction des entrées fautives des différentes sources, via leur croisement et par alignement ortho-phonétique. Les divergences sont listées entrée par entrée et seront transmises aux auteurs des sources (documentation de la provenance, §10).
5. **Une prononciation, un alignement et une syllabation qui concordent.** Pour chaque syllabe prononcée, on sait quelles lettres l'écrivent.
6. **Des fréquences par analyse.** La fréquence de *mange* est répartie entre l'indicatif (1re et 3e personne), le subjonctif et l'impératif.
7. **Une provenance documentée.** Chaque prononciation est tracée : quelle source l'a proposée et pourquoi elle a été retenue (dossier `provenance/`).

LeXiK n'est pas un dictionnaire : il ne donne ni définitions ni sens, et ne distingue pas deux sens d'un même mot.

L'élaboration d'un dictionnaire est l'objet d'un projet distinct, **WikiLexik**, consultable en ligne sur https://lexique.lectura.world. Il branche sur les lemmes de LeXiK deux autres couches : la couche **sémantique** du Wiktionnaire (sens, définitions, synonymes, via l'extraction Kaikki) et la couche **encyclopédique** de Wikidata (les entités liées à un mot : personnes, lieux, œuvres…, avec leurs articles Wikipédia). LeXiK reste la référence des formes ; WikiLexik y rattache ce qu'elles veulent dire.

2. Prise en main

2.1 Contenu du dépôt

```
LISEZ-MOI.md                  présentation, licence, citation
DOCUMENTATION.md / .pdf       ce guide
LICENSE                       licence CC BY-SA 4.0
lexik_formes_mots.tsv.gz         les mots
lexik_formes_multimots.tsv.gz    les locutions et phrases
lexik_formes_speciaux.tsv.gz     les formes spéciales
lexik_formes_non_classes.tsv.gz  les formes non classées
lexik_lemmes.tsv.gz              le fichier des lemmes des mots
annexes/
    lexik_multext.tsv.gz      les codes grammaticaux MULTEXT et leur sens (§9)
    lexik_pronominaux.tsv.gz  la part pronominale de chaque verbe (§10.5)
provenance/
    DOCUMENTATION_PROVENANCE.md / .pdf   construction de la base
    lexik_provenance.tsv.gz              origine de chaque prononciation (et des propriétés hors mots)
    lexik_lemmes_provenance.tsv.gz       origine des propriétés des lemmes
checksums.sha256              sommes de contrôle
```

2.2 Ouvrir les fichiers

Les fichiers sont des tableaux texte, une colonne par tabulation, compressés en gzip. Le fichier des lemmes s'ouvre dans un tableur après décompression (séparateur : **tabulation** ; délimiteur de texte : **aucun**). Le fichier des formes (2 580 027 lignes) dépasse la limite des tableurs (1 048 576 lignes) pour les mots ; les autres fichiers s'ouvrent dans un tableur. En ligne de commande, en Python ou en SQL :

```
zgrep -P "^petit\t" lexik_formes_mots.tsv.gz
```

```
import csv, pandas as pd
mots = pd.read_csv("lexik_formes_mots.tsv.gz", sep="\t", quoting=csv.QUOTE_NONE,
                   dtype=str, keep_default_na=False)
mots[mots.ortho == "petit"]
```

```
SELECT * FROM read_csv('lexik_formes_mots.tsv.gz', delim='\t', quote='',
                       header=true, all_varchar=true)
WHERE ortho = 'petit';
```

2.3 Lire une ligne

Voici la ligne de la forme *petit* adjectif dans le fichier des formes, colonne par colonne :

Colonne	Valeur	Ce qu'elle dit
`ortho`	`petit`	la forme écrite
`lemme`	`petit`	son lemme
`id_lemme`	`petit#ADJ`	l'identifiant du lemme, lien vers le fichier des lemmes
`upos`	`ADJ`	adjectif
`multext`	`Afpms`	adjectif qualificatif, masculin singulier (§9)
`phone`	`pə.ti[t]`	/pə.ti/, avec un *t* qui se prononce en liaison (*petit ami*) (§6)
`alignement`	`p=p ə=e t=t i=i [t]=t°`	*p* écrit /p/, *e* écrit /ə/…, le *t* final est muet mais porte la liaison (§7)
`orthocode`	`pe.tit°`	deux syllabes écrites, *t* final muet (§8)
`composition`	*(vide)*	rempli seulement pour les mots composés
`freq_opensubtitles`	`485.845`	fréquence de la graphie *petit* (toutes analyses) par million de mots, dans les sous-titres (§10)
`freq_wikipedia`, `freq_gutenberg`	`168.391`, `454.248`	idem dans Wikipédia et dans Gutenberg
`parts_opensubtitles`	`0.944`	part de cette analyse dans la fréquence de la graphie (§10)
`parts_wikipedia`, `parts_gutenberg`	`0.983`, `0.944`	idem
`confiance`	`A`	prononciation confirmée par plusieurs sources (§11)

Et la ligne de son lemme, `petit#ADJ`, dans le fichier des lemmes :

Colonne	Valeur	Ce qu'elle dit
`id_lemme`	`petit#ADJ`	l'identifiant, le même que dans le fichier des formes
`lemme`, `upos`	`petit`, `ADJ`	le lemme et sa catégorie
`proprietes`	`genre:m,f;epicene:non;invariable:non;feminin:petite;pluriel:petits`	adjectif à deux genres, non épicène (féminin *petite*), variable en nombre (pluriel *petits*) (§5)
`freq_opensubtitles`	`905.362`	fréquence du lemme : somme des fréquences de ses quatre formes (*petit*, *petite*, *petits*, *petites*), chacune pour sa part analysée comme cet adjectif (§10)
`freq_wikipedia`, `freq_gutenberg`	`503.394`, `1050.015`	idem dans Wikipédia et dans Gutenberg

Les quatre formes de *petit* se retrouvent dans le fichier des formes par `id_lemme` = `petit#ADJ`.

2.4 Quelques cas limites

Une graphie qui relève de plusieurs lemmes a plusieurs lignes ; un lemme qui a plusieurs analyses pour une même graphie n'en a qu'une.

Graphie	Lignes	Prononciation	Ce que montre l'exemple
livre	`livre#NOUN#1` (masculin) ; `livre#NOUN#2` (féminin) ; `livrer#VERB`	livʁ	deux noms homonymes de genres différents sont deux lemmes, numérotés ; les cinq analyses du verbe (*je livre*, *il livre*, *livre !*, *que je livre*, *qu'il livre*) tiennent sur une ligne
fils	`fils#NOUN` ; `fil#NOUN`	fis / fil[z]	*fils* « enfant » et *fils* pluriel de *fil* : même graphie, deux lemmes, deux prononciations
avocat	`avocat#NOUN`	a.vɔ.ka[t]	l'homme de loi et le fruit : même lemme, même catégorie, même genre, donc une seule ligne
couvent	`couvent#NOUN` ; `couver#VERB`	ku.vɑ̃[t] / kuv[t]	même graphie, prononciations différentes : chaque ligne a la sienne
est	`être#AUX`, `être#VERB` ; `est#NOUN`, `est#ADJ`	{e|ɛ}[t] / ɛst	le verbe et le point cardinal
mousse	`mousse#NOUN#1` (féminin) ; `mousse#NOUN#2` (masculin) ; `mousse#ADJ` ; `mousser#VERB`	mus	quatre lemmes pour une graphie et une prononciation

Le genre n'entre dans l'identifiant du lemme (`#1`, `#2`) que pour départager deux homonymes ; sinon, il est dans les propriétés du lemme (§5).

3. Les quatre sous-lexiques

3.1 Mots

Mots simples

Un mot simple est une suite de lettres sans tiret, apostrophe ni espace à l'intérieur : *maison*, *mangeons*, *Paris*. Les formes élidées (*l'*, *qu'*) et les pronoms postposés (*-elle*, *-nous*) sont aussi des mots simples : le tiret ou l'apostrophe n'y marque que le rattachement au mot voisin.

	Graphies	Analyses	Lemmes
Noms communs et autres	1 693 055	2 257 128	311 294
Noms propres	326 049	352 270	325 741

Mots composés

Un mot composé est formé de segments reliés par un tiret, une apostrophe ou une espace : *arc-en-ciel*, *aujourd'hui*, *pomme de terre*.

- Un composé à **tiret** ou à **apostrophe** est un mot dès qu'une source l'atteste.
- Une suite à **espace** n'est un mot que si elle est attestée comme unité par au moins deux sources indépendantes, ou annotée au moins trois fois comme expression figée dans les corpus annotés libres (treebanks Universal Dependencies). Cela donne une liste fermée de 943 lemmes (*pomme de terre*, *alter ego*, *à jeun*, *parce que*, *en fait*, *tout à fait*). Les autres suites à espace sont des locutions (sous-lexique MultiMots).

La colonne `composition` donne la catégorie et le lemme de chaque segment : `arc:NOUN:arc en:ADP:en ciel:NOUN:ciel`. Un segment qui n'a pas d'entrée propre dans LeXiK Mots (un **élément**) est noté `segment:ELEMENT`, sans lemme : *ultra-costaud* `ultra:ELEMENT costaud:ADJ:costaud`, *-t-il* `t:ELEMENT il:PRON:il`. Sa prononciation est donnée dans LeXiK Spéciaux.

	Graphies	Analyses	Lemmes
Noms communs et autres	137 455	179 653	29 030
Noms propres	86 159	88 395	86 118

3.2 Locutions et phrases (`MultiMots`)

Toutes les suites à espace qui ne sont pas des mots au sens du §3.1 sont des **locutions**, quelle que soit leur catégorie : verbales (*avoir lieu*), nominales (*pomme de discorde*), adverbiales (*à la rigueur*)… Les verbes que les sources donnent avec leur pronom (*se battre*, *s'évanouir*) sont aussi des locutions, à l'infinitif.

Une seule forme : la forme de citation

Une locution est enregistrée sous sa **forme de citation** seulement, sans ses flexions : *avoir lieu*, pas *a eu lieu*. La colonne `flechir` donne le ou les mots qui portent la flexion : *avoir lieu* → `avoir` ; *faire attention* → `faire` ; *pomme de discorde* → `pomme`. Pour retrouver une forme fléchie, on fléchit ce mot avec le fichier des formes de LeXiK Mots.

Les phrases

Les phrases du Wiktionnaire (proverbes, formules, répliques figées) n'ont pas d'analyse grammaticale ; leur prononciation est celle que donne le Wiktionnaire (confiance `K`).

Fréquence et part pronominale

La fréquence d'une locution compte **toutes ses formes** : *avoir lieu* compte aussi « a eu lieu », « aura lieu ». Chaque locution est cherchée dans les corpus à partir de sa composition (colonne `composition`) : le verbe peut être conjugué, le nom ou l'adjectif fléchi, le pronom réfléchi ou possessif varier (*je m'en vais* pour *s'en aller*), et une négation ou un adverbe peut s'insérer (*je ne m'en bats pas*). Méthode, exemples et limites : §10.4.

Un verbe pronominal (*se battre*) a ainsi sa propre fréquence, distincte de celle du verbe simple (*battre*). L'annexe `annexes/lexik_pronominaux.tsv` les met côte à côte et donne la **part de pronominalité** de chaque verbe, c'est-à-dire la part de ses emplois qui sont pronominaux : dans les sous-titres, 58 % pour *se battre*, 65 % pour *se taire* (§10.5).

3.3 Formes spéciales (`Spéciaux`)

Les formes spéciales viennent d'un **inventaire** établi à partir des entrées du Wiktionnaire (extraction Kaikki), puis **complété et relu à la main**. Chaque forme y a sa lecture et sa **forme développée** (*km* → *kilomètre*, *SNCF* → *Société nationale des chemins de fer français*).

Une forme entre dans les Spéciaux par sa **forme** (plusieurs majuscules, chiffres, caractères non alphabétiques) ou par sa **lecture**, qui n'est pas celle d'un mot (*km* se lit *kilomètre*). La colonne `type` dit de quoi il s'agit :

type	Entrées	Exemples
`sigle`	3 657	*SNCF* (épelé), *PL*
`acronyme`	1 113	*ONU*, *MOOC* (lu comme un mot)
`abreviation`	721	*M.* (monsieur), *etc.*
`unite`	215	*km*, *kg*
`code`	224	*KRW* (won sud-coréen)
`mixte`	243	*O2* (dioxygène), *GitLab*
`lettre`	133	*ψ* (psi)
`nombre`	88	*Xe* (dixième), *2nde*
`symbole`	62	*%* (pour cent), = (égale)
`prefixe`, `suffixe`, `interfixe`	1 005, 465, 5	*anti-*, *-able*
`element`	37 988	morceau d'un nom composé qui n'existe pas seul : *Séponet* (de *Suc de Séponet*)

Les **éléments** sont les segments de mots composés qui n'ont pas d'entrée propre dans LeXiK Mots. Ils sont publiés pour que chaque segment d'un composé ait une prononciation ; leur développement donne le ou les composés dont ils viennent. Dans la colonne `composition` des composés, ils sont notés `segment:ELEMENT`, sans lemme (*ultra-costaud* : `ultra:ELEMENT costaud:ADJ:costaud`). Ils serviront de réservoir pour les versions suivantes : un élément qui s'avère être un vrai mot pourra recevoir une entrée dans LeXiK Mots.

La fréquence d'une forme spéciale est celle de sa **graphie, tout court** : elle est comptée telle qu'elle s'écrit, majuscules comprises (*CE* et *ce* à part), sans distinguer ses lectures. Les deux lignes de *AB* (*Agriculture Biologique*, *assez bien*) portent le même chiffre (§10.6).

3.4 Formes non classées (`Non classé`)

Ce sont les formes que LeXiK n'a pas pu ranger dans les trois autres sous-lexiques. Elles sont de deux origines :

- **issues des lexiques** (15 038 entrées) : des entrées des sources qui ne sont ni des mots, ni des locutions, ni des formes spéciales au sens de LeXiK (noms de molécules, graphies fautives ou sans accent, entrées incomplètes) ; la colonne `type` en donne parfois la raison (`accent manquant ou fautif`) ;
- **issues des corpus** (13 867 noms communs et 78 736 noms propres candidats) : des formes fréquentes dans les corpus mais absentes de toutes les sources, relevées comme candidates (`type = candidat`) ; elles n'ont ni prononciation ni analyse, seulement leurs fréquences.

À utiliser avec précaution. Ces formes sont publiées telles quelles, sans vérification : une graphie peut être fautive, un lemme proposé par sa source peut être faux, et les candidates issues des corpus n'ont ni prononciation ni analyse. Elles sont tenues à part, dans leur propre fichier (`lexik_formes_non_classes.tsv`), et ne se mêlent pas aux trois autres sous-lexiques : on peut les ignorer sans rien perdre du reste. Elles servent de réservoir pour les versions suivantes, où elles pourront être vérifiées et rejoindre un autre sous-lexique.

4. Les fichiers de formes (`lexik_formes_*.tsv`)

Les quatre fichiers ont la même organisation et des colonnes communes ; chacun n'a que les colonnes utiles à son sous-lexique (dernière colonne du tableau ci-dessous). Une ligne par **entrée**, c'est-à-dire par couple (graphie, lemme). Quand une graphie a plusieurs analyses pour un même lemme, elles sont réunies sur la même ligne : leurs codes `multext` sont séparés par `;`, et les colonnes `parts_*` suivent le même ordre.

Colonne	Contenu	Voir	Fichiers
`ortho`	la graphie		tous
`lemme`	le lemme (pour une forme non classée, tel que proposé par sa source, non vérifié)		mots, non classés
`id_lemme`	`lemme#CATÉGORIE`, plus `#1`, `#2` pour départager deux homonymes ; lien vers le fichier des lemmes	§5	mots
`upos`	la catégorie grammaticale (NOUN, VERB, ADJ…)	§9	tous
`multext`	le code grammatical complet de chaque analyse, séparés par `;`	§9	tous
`phone`	la prononciation, avec ses marques (liaison, h aspiré, variantes)	§6	tous
`alignement`	les couples phonème=graphème	§7	mots, locutions, spéciaux (vide)
`orthocode`	l'orthographe découpée en syllabes	§8	mots, locutions, spéciaux (vide)
`composition`	pour un composé ou une locution, ses segments : `segment:catégorie:lemme`	§3	mots, locutions
`flechir`	pour une locution, le ou les mots qui portent la flexion	§3.2	locutions
`type`	pour une forme spéciale ou non classée, son type	§3.3, §3.4	spéciaux, non classés
`developpement`, `developpement_pluriel`	pour une forme spéciale, ce qu'elle abrège	§3.3	spéciaux
`proprietes`	propriétés de l'entrée (genre, pluriel, groupe du verbe…), comme celles d'un lemme	§5	locutions, spéciaux, non classés
`freq_opensubtitles`, `freq_wikipedia`, `freq_gutenberg`	fréquence de la graphie, par million de mots ; pour une locution, de toutes ses formes ; pour une forme spéciale, de la graphie seule, toutes lectures confondues	§10	tous

Colonne	Contenu	Voir	Fichiers
`parts_opensubtitles`, `parts_wikipedia`, `parts_gutenberg`	part de chaque analyse dans cette fréquence	§10	mots, non classés
`confiance`	d'où vient la prononciation et par quoi elle est confirmée	§11	tous

D'autres informations se calculent à partir de ces fichiers :

- **les homophones** : les entrées de même prononciation brute (§6.4) ;
- **les mots à h aspiré** : les entrées dont `phone` commence par `*h*` ;
- **la fréquence d'une analyse** : fréquence de la graphie × part (§10) ;
- **l'indice Zipf** : $\log_{10}$(fréquence) + 3.

5. Le fichier des lemmes (`lexik_lemmes.tsv`)

Un lemme regroupe toutes les formes fléchies d'un même mot.

Colonne	Exemple (*manger*)	Contenu
`id_lemme`	`manger#VERB`	identifiant, lien vers le fichier des formes
`lemme`, `upos`	`manger`, `VERB`	le lemme et sa catégorie
`proprietes`	`groupe:1;auxiliaire:avoir;pronominal:possible;transitivite:td,i`	propriétés, en `clé:valeur` séparées par `;`
`freq_opensubtitles`, `freq_wikipedia`, `freq_gutenberg`		fréquence du lemme par million de mots (somme de ses formes)

La prononciation d'un lemme, sa confiance et la liste de ses formes se lisent dans le fichier des formes, par `id_lemme`.

Les propriétés possibles :

Clé	Valeurs	Pour
`genre`	`m`, `f`, `m,f`	noms, adjectifs, noms propres, déterminants, pronoms
`epicene`	`oui` : même forme aux deux genres (*artiste*) ; `non` : féminin distinct	lemmes à deux genres
`invariable`	`oui` : même forme au singulier et au pluriel (*souris*)	lemmes à deux nombres
`nombre`	`s` singulier seulement, `p` pluriel seulement (*funérailles*)	
`feminin`, `pluriel`	adjectifs : forme du féminin, du pluriel ; noms : lemme féminin (*acteur* → *actrice*)	
`masculin`	lemme masculin (*actrice* → *acteur*)	noms féminins
`groupe`	`1`, `2`, `3`	verbes
`auxiliaire`	`avoir`, `être`, `avoir,être`	verbes
`pronominal`	`essentiel` (*s'évanouir*), `possible` (*se laver*)	verbes
`transitivite`	`td` direct, `ti` indirect, `i` intransitif	verbes
`impersonnel`, `defectif`	`oui`	verbes
`type`	`prenom`, `patronyme`, `lieu`, `autre` (plusieurs possibles)	noms propres
`sous_type`	`commune`, `relief`, `hydrographie`, `vegetation`, `lieu_habite`, `zone`, `division_administrative`	lieux
`ortho_1990`	`rectifiee`, `traditionnelle`	graphies concernées par les rectifications de 1990
`etranger`	`oui`	emprunts non francisés

Trois conventions :

- un nom féminin est son propre lemme (*actrice*, pas *acteur*) ; le lien entre les deux est dans `masculin` et `feminin` ;
- un verbe essentiellement pronominal est cité sans son pronom (*évanouir*, avec `pronominal:essentiel`) ;
- deux homonymes de même catégorie mais de genres différents sont deux lemmes (`livre#NOUN#1`, masculin ; `livre#NOUN#2`, féminin) ; de même quand leur flexion diffère (`ciel#NOUN#1`, pluriel *cieux* ; `ciel#NOUN#2`, pluriel *ciels*). Une forme commune aux deux (*ciel*) est rattachée au premier.

Le fichier des lemmes ne couvre que les mots. Les locutions, les formes spéciales et les formes non classées n'y figurent pas : leurs propriétés sont dans la colonne `proprietes` de leur fichier, dans les mêmes `clé:valeur`.

6. La prononciation

6.1 Les phonèmes

LeXiK utilise 36 phonèmes, notés en alphabet phonétique international :

Classe	Phonèmes
Voyelles orales	`a e ɛ i o ɔ u y ø œ ə`
Voyelles nasales	`ɑ̃ ɛ̃ ɔ̃ œ̃`
Semi-voyelles	`j w ɥ`
Consonnes	`p b t d k g f v s z ʃ ʒ m n ɲ ŋ l ʁ`

Choix retenus :

- **un seul *a*** : le *a* postérieur (*pâte*) est noté /a/, comme celui de *patte* ; peu de locuteurs les distinguent encore ;
- **/œ̃/ distinct de /ɛ̃/** (*brun* /bʁœ̃/, *brin* /bʁɛ̃/) : qui ne fait pas la différence peut fusionner les deux, l'inverse serait impossible ;
- **un seul r**, /ʁ/ ;
- **les sons des emprunts en deux phonèmes** : *tchèque* /tʃɛk/, *jean* /dʒin/ ;
- **caractères** : `ɡ` est le caractère phonétique U+0261, pas la lettre `g` ; une voyelle nasale est la voyelle suivie du tilde combinant U+0303.

6.2 Lire la colonne `phone`

La colonne `phone` contient les phonèmes et quelques signes.

Les frontières

Signe	Sens	Exemple
`.`	frontière de syllabe	*petit* `pə.ti`
`‿`	**pont** : une syllabe enjambe la frontière entre deux segments d'un mot composé	*États-Unis* `e.ta.z‿y.ni`

Le pont est le signe de liaison de l'API. Dans *États-Unis*, le /z/ final du premier segment forme une syllabe avec le /y/ du second : /e.ta.z‿y.ni/. Il sert pour la liaison (*États-Unis*), l'enchaînement (*arc-en-ciel* /aʁ.k‿ɑ̃.sjɛl/) et l'élision (*aujourd'hui* /o.ʒuʁ.d‿ɥi/). Quand les segments ne partagent pas de syllabe, la frontière est un simple point (*pomme de terre* /pɔm.də.tɛʁ/). Dans un mot, les tirets, apostrophes et espaces n'apparaissent pas dans la prononciation ; l'alignement dit où ils sont (§7). Dans une **locution** ou une **phrase**, au contraire, chaque mot garde sa prononciation et les mots sont séparés par une espace (*avoir lieu* /a.vwaʁ ljø/) ; le pont remplace l'espace quand une syllabe relie deux mots (*faire attention* /fɛ.ʁ‿a.tɑ̃.sjɔ̃/).

Les marques : elles notent ce que la forme isolée ne laisse pas deviner.

Marque	Sens	Exemple
`[t]`, `[z]`, `[n]`…	**consonne latente** : muette quand le mot est seul, prononcée en liaison devant une voyelle	`pə.ti[t]` : *petit* /pə.ti/, *petit ami* /pə.ti.ta.mi/
`[a>b]`	la liaison **remplace** un son : `a` seul, `b` devant voyelle	`bɔ̃[ɔ̃>ɔn]` : *bon* /bɔ̃/, *bon ami* /bɔ.na.mi/

Marque	Sens	Exemple
`[a>b;a>c]`	deux contextes : devant voyelle, puis devant consonne	`dis[s>z;s>]` : *di**x*** /dis/, *di**x** amis* /di.za.mi/, *di**x** tomates* /di.tɔ.mat/
`{a|b}`	**variante libre**, la plus courante en premier	`{ɛ|e}g.zɔ̃.de` : *exonder* /ɛg.zɔ̃.de/ ou /eg.zɔ̃.de/
h au début	**h aspiré** : ni liaison ni élision (*le héros*, *la hache*)	`*h*e.ʁo[z]`

Le h muet ne se note pas : il ne bloque rien (*l'homme*, *les hommes*).

6.3 Les conventions de notation

Les dictionnaires ne se contredisent le plus souvent pas sur la prononciation : ils l'écrivent de façons différentes (schwa noté ou non, *o* ouvert ou fermé…). LeXiK fixe une **forme unique**, selon un principe : **on note ce qui ne se devine pas, on ne note pas ce qui se calcule**.

Règle	Question	Choix de LeXiK	Exemple
C1	*o/ɔ*, *e/ɛ*, *ø/œ* en dernière syllabe	selon le mot (les sources votent) ; en cas d'égalité, variante `{a|b}`	*pot* /po/, *porc* /pɔʁ/
C2	*o/ɔ*, *e/ɛ*, *ø/œ* ailleurs	d'après l'orthographe : *au*, *eau*, *ô* → /o/, *o* → /ɔ/ ; *é* → /e/, *è*, *ê*, *ai* → /ɛ/	*beauté* /bo.te/, *botté* /bɔ.te/
C3	*e* muet final	jamais noté	*table* /tabl/
C4	*e* muet intérieur	toujours noté (sa chute se calcule, son ajout non)	*samedi* /sa.mə.di/
C5	*e* muet noté /ø/ par certaines sources	toujours /ə/	*renoter* /ʁə.nɔ.te/
C6	*i*, *u*, *ou* + voyelle, après consonne + *l* ou *r*	deux syllabes	*plier* /pli.je/
C7	*i*, *u*, *ou* + voyelle, ailleurs	une syllabe (semi-voyelle)	*distribuez* /dis.tʁi.bɥe/
C8	finale *-isme*	/ism/	*réalisme* /ʁe.a.lism/
C9	consonne qui s'assourdit au contact d'une autre	notée assourdie ; la forme d'origine en variante	*absent* /ap.sɑ̃/, noté `a{p|b}.sɑ̃`
C10	*gn* et *ni* + voyelle	/ɲ/ pour *gn*, /nj/ pour *ni* ; /gn/ pour quelques mots savants	*agneau* /a.ɲo/, *panier* /pa.nje/, *diagnostic* /djag.nɔs.tik/
C11	consonnes doubles	jamais à l'intérieur d'un mot, même quand l'orthographe double la lettre	*illégal* /i.le.gal/, *courrait* /ku.ʁɛ/
C12	finales *-ai* / *-ais*	/e/ pour *-ai*, /ɛ/ pour *-ais*	*mangeai* /mɑ̃.ʒe/, *mangeais* /mɑ̃.ʒɛ/

Pour C11, à la frontière de deux segments d'un composé, la consonne se prononce bien deux fois et reste double (*bonnes nuits* /bɔn.nɥi/).

Les sons facultatifs ne sont pas marqués. Un son qui peut tomber (*samedi* /sa.mə.di/ ou /sam.di/) change la syllabation : le *m* passe dans la syllabe précédente. LeXiK garde une seule forme syllabée : le *e* muet intérieur toujours noté, le *e* muet final jamais. Une marque `( )` est prévue pour une version ultérieure (`sa.m(ə).di`) ; la forme sans le son s'obtiendrait en retirant le contenu des parenthèses puis en refaisant la syllabation (§8.1).

6.4 La prononciation brute

Pour comparer des prononciations (chercher des homophones, par exemple), on utilise la forme **brute** : on retire ce qui est entre `[ ]` et `* *`, et on garde la première valeur des `{ }`.

`dis[s>z;s>]` → `dis` ; `*h*e.ʁo[z]` → `e.ʁo` ; `{ɛ|e}g.zɔ̃.de` → `ɛg.zɔ̃.de`.

6.5 D'où viennent les prononciations

Les prononciations ont été établies en deux temps :

1. **Compilation des sources.** Les prononciations données par les dictionnaires sont nettoyées (voir ci-dessous), ramenées aux conventions du §6.3, vérifiées par l'aligneur (§7) et départagées par un vote entre sources (détail dans la documentation de la provenance).
2. **Complément par le modèle.** Un modèle graphème → phonème de Lectura a été entraîné sur cette première version de la base, puis a prononcé les formes qu'aucune source ne prononce.

Le nettoyage des transcriptions. Avant toute comparaison, les transcriptions des sources sont ramenées à l'inventaire de LeXiK (§6.1) :

- **notations converties** : le SAMPA de Lexique (`E`, `O`, `@`, `R`…) en API ; `ɑ` en `a` ; les nasales écrites sur une autre voyelle (`ã`, `õ`, `ẽ`) en `ɑ̃`, `ɔ̃`, `ɛ̃` ;
- **marques retirées** : accent tonique, longueur (`ː`), points de syllabe et liaisons des sources (LeXiK refait la syllabation), diacritiques de transcription fine ;
- **consonnes doubles réduites** à l'intérieur d'un mot (règle C11) ;
- **rien n'est deviné** : une transcription qui contient un caractère hors inventaire, non couvert par ces conversions, est rejetée (le rejet est tracé dans la provenance) plutôt que corrigée au hasard.

Les prononciations prédites par le modèle sont **moins fiables** que celles des sources, en particulier pour les noms propres et les mots étrangers. Elles portent la confiance `G` (§11) : on peut les isoler ou les écarter selon l'usage.

7. L'alignement graphème–phonème

7.1 Ce que dit la colonne

La colonne `alignement` dit quelles lettres écrivent quel son. C'est une suite de **couples** `phonème=graphème`, séparés par une espace, dans l'ordre du mot :

Mot	Prononciation	Alignement
chapeau	ʃa.po	`ʃ=ch a=a p=p o=eau`
femme	fam	`f=f a=e m=mm =e°`
oiseaux	wa.zo	`wa=oi z=s o=eau [z]=x°`
exister	ɛg.zis.te	`ɛ=e gz=x i=i s=s t=t e=er`
petit	pə.ti	`p=p ə=e t=t i=i [t]=t°`
héros	e.ʁo	`*h*=h° e=é ʁ=r o=o [z]=s°`
homme	ɔm	`=h° ɔ=o m=mm =e°`
grand-père	gʁɑ̃.pɛʁ	`g=g ʁ=r ɑ̃=an =d° =- p=p ɛ=è ʁ=r =e°`
aujourd'hui	o.ʒuʁ.d‿ɥi	`o=au ʒ=j u=ou ʁ=r d=d ‿=' =h° ɥi=ui`
États-Unis	e.ta.z‿y.ni	`e=É t=t a=a =t° z=s ‿=- y=U n=n i=i =s°`

7.2 Les règles de notation

1. **Le phonème d'abord**, puis le graphème qui l'écrit : `ʃ=ch`.
2. **Un couple par graphème** : un graphème est la lettre ou le groupe de lettres qu'on enseigne comme une unité (`ch`, `ou`, `eau`, `gn`, et les consonnes doubles qui valent un seul son : `mm`, `ss`).
3. **Un graphème garde tous ses phonèmes** : `ks=x` (*taxi*), `wa=oi` (*oiseau*).
4. **Une lettre muette est un couple à part**, sans phonème, marquée `°` : `=e°`, `=t°`, `=h°`.
5. **Les marques de la prononciation vont sur la lettre qui les porte** : la consonne latente sur la lettre muette (`[t]=t°`), le h aspiré sur le *h* (`*h*=h°`).
6. **Un tiret, une apostrophe ou une espace est aussi un couple**. Son côté phonème dit ce qui se passe à la frontière : rien (`=-`, `='`, `=␣`, où ␣ représente l'espace) quand les deux segments se prononcent séparément ; le pont (`‿=-`, `‿='`, `‿=␣`) quand une syllabe enjambe la frontière. Dans une locution, l'espace entre deux mots prononcés séparément est noté `␣=␣` (*avoir lieu* : `a=a v=v wa=oi ʁ=r ␣=␣ l=l j=i ø=eu`).

Lecture de *héros* : le *h* ne se prononce pas (`=h°`) mais il est aspiré (`*h*`) : pas de liaison, *les héros* /le.e.ʁo/. Le *s* final est muet mais porte un /z/ de liaison (*héros aimés*). Lecture de *homme* : le *h* est muet et sans marque, donc *l'homme*, *les hommes*.

Deux vérifications tiennent sur toutes les lignes alignées : les graphèmes mis bout à bout redonnent l'orthographe, et les phonèmes mis bout à bout redonnent la prononciation.

7.3 Comment l'alignement est calculé

L'aligneur de Lectura part d'une **table de correspondances** : pour chaque phonème, les graphèmes qui peuvent l'écrire (/o/ → *o*, *au*, *eau*, *ô*…). Il avance en même temps dans la prononciation et dans l'orthographe, essaie à chaque pas tous les graphèmes possibles, et revient en arrière quand il est bloqué. Une lettre qui n'écrit aucun son est déclarée muette. De tous les alignements complets trouvés, il garde celui qui a **le moins de lettres muettes**, puis celui qui les place le plus tôt dans le mot (*Strazeeloise* : `z=z =e° ə=e`, le premier *e* muet), puis celui qui utilise le moins de graphèmes absents de la table, et enfin celui qui a le moins de couples.

La table est volontairement large : un alignement réussi prouve seulement que la prononciation est compatible avec l'orthographe. Un **échec** est en revanche un signal fort d'erreur dans la prononciation ; LeXiK s'en sert pour écarter les transcriptions fautives des sources (voir la documentation de la provenance).

L'alignement est **laissé vide** quand l'aligneur ne trouve pas de correspondance connue entre les lettres et les sons : il renonce plutôt que de produire un alignement faux. C'est surtout le cas des **noms propres étrangers**, dont les graphies n'existent pas en français (*Krajewski*, *-wski* prononcé /fski/ ; *Morici*, *ci* prononcé /tʃi/ ; *Neuenschwander*, *eu* prononcé /ɔj/), et de formes rares ou fabriquées (*smsserais*, *burn-outé*). La **prononciation de ces formes reste renseignée et valable** ; beaucoup ont été vérifiées à la main (confiance `R`). Seuls l'alignement et l'orthocode manquent.

Les formes spéciales et non classées n'ont pas d'alignement.

8. La syllabation et l'orthocode

8.1 La syllabation de la prononciation

Les syllabes sont découpées selon la règle classique du français : **la syllabe suivante prend le plus de consonnes possible au début, tant que le groupe peut commencer une syllabe en français** (principe de l'attaque maximale).

Les groupes qui peuvent commencer une syllabe (les « attaques licites ») forment une liste fermée :

- une consonne seule ;
- une occlusive ou une fricative labiodentale (/p b t d k g f v/) suivie de /l/ ou /ʁ/ : /pl/, /tʁ/, /gʁ/, /fl/…, sauf /tl/ et /dl/ ;
- l'un ou l'autre suivi d'une semi-voyelle (/j w ɥ/) : /pjɛʁ/, /tʁɥi/, /pwa/.

Entre deux voyelles, tout ce qui précède la plus longue attaque licite termine la syllabe précédente :

- une consonne entre deux voyelles va avec la seconde : /mɛ.zɔ̃/ ;
- deux consonnes se séparent, sauf si elles forment une attaque licite : /at.las/, /tɛk.nik/, mais /ta.blo/, /a.pʁɛ/ ;
- devant une attaque licite, le reste va en coda : /ɛs.pʁi/, /abs.tʁɛ/, /kɔ̃s.tʁɥiʁ/ ;
- /ks/, /gz/, /tʃ/, /dʒ/ comptent pour deux consonnes : /tak.si/, /ɛg.zis.te/ ;
- /j/ après voyelle termine la syllabe devant consonne ou en fin de mot (/tʁa.vaj/, /pɛj.zaʒ/) ; entre deux voyelles, il commence la suivante (/gʁi.jaʒ/).

Pourquoi une liste et pas une échelle de sonorité. Le modèle courant classe les consonnes par sonorité (occlusives < fricatives < nasales < liquides < semi-voyelles) et accepte toute attaque dont la sonorité monte vers la voyelle. Pour le français, il accepte trop : /kn/, /tm/, /dm/ ou /tl/ montent en sonorité mais ne commencent pas une syllabe française, et il couperait *technique* en /tɛ.knik/ ou *atlas* en /a.tlas/. L'aligneur de Lectura partait de ce modèle ; LeXiK le remplace par la liste fermée ci-dessus, qui en est la version restreinte au français.

8.2 L'orthocode

L'orthocode découpe l'**orthographe** en suivant la syllabation de la prononciation. Les syllabes sont séparées par `.`, les lettres muettes marquées `°`.

Mot	Prononciation	Orthocode
petit	pə.ti	`pe.tit°`

Mot	Prononciation	Orthocode
femme	fam	`femme°`
Marseille	maʁ.sɛj	`Mar.seille°`
exister	ɛg.zis.te	`e(x).xis.ter`
axe	aks	`ax²e°`
jazz	dʒaz	`j²azz`
grand-père	gʁɑ̃.pɛʁ	`grand°-.père°`
États-Unis	e.ta.z‿y.ni	`É.tat°.s-U.nis°`

Deux notations pour une lettre qui vaut deux sons :

- **à cheval sur deux syllabes** : `(x).x`. Le *x* d'*exister* vaut /gz/, dont /g/ termine la première syllabe et /z/ commence la seconde ; il est écrit entre parenthèses dans la première, en clair dans la seconde ;
- **dans une même syllabe** : `x²`. Le *x* d'*axe* vaut /ks/ dans une seule syllabe : `ax²e°`. De même *jazz* `j²azz` (*j* = /dʒ/), *express* `ex².press`.

Un groupe de plusieurs lettres qui vaut deux sons n'est pas marqué (*roi* : *oi* = /wa/) ; l'alignement en donne le détail.

L'orthocode est la **projection de la syllabation phonétique** sur l'orthographe, par l'alignement : chaque frontière de syllabe de la prononciation est reportée entre les lettres correspondantes. Dans un mot composé, le tiret, l'apostrophe ou l'espace reste à sa place :

- sans pont, la frontière tombe après le séparateur : *grand-père* /gʁɑ̃.pɛʁ/ `grand°-.père°` ;
- avec un pont (§6.2), la syllabe enjambe le séparateur, et l'orthocode aussi : *États-Unis* /e.ta.z‿y.ni/ `É.tat°.s-U.nis°`, *arc-en-ciel* /aʁ.k‿ɑ̃.sjɛl/ `ar.c-en-.ciel` (pont après *arc*, pas après *en*), *aujourd'hui* /o.ʒuʁ.d‿ɥi/ `au.jour.d'h°ui`.

9. L'analyse grammaticale

9.1 La catégorie (`upos`)

La catégorie suit les conventions **Universal Dependencies** (UD), le standard international d'annotation grammaticale.

`upos`	Catégorie	Analyses	Exemples
VERB	verbe	1 724 823	*mange*, *mangeait*, *mangé*
PROPN	nom propre	444 531	*Paris*, *Dupont*, *Marie*
NOUN	nom commun	500 092	*maison*, *eau*
ADJ	adjectif	296 575	*petit*, *bleu*
ADV	adverbe	8 941	*bien*, *très*
autres		138 614	AUX auxiliaire, PRON pronom, DET déterminant, ADP préposition, CCONJ et SCONJ conjonctions, NUM nombre, INTJ interjection, X mot étranger (48 : *the*, *von*…)

9.2 Le code MULTEXT (`multext`)

Le code MULTEXT (jeu GRACE pour le français) résume toute l'analyse en quelques lettres, une position par information :

Code	Lecture
`Afpms`	**A**djectif, quali**f**icatif, **p**ositif, **m**asculin, **s**ingulier
`Ncfs`	**N**om **c**ommun, **f**éminin, **s**ingulier
`Npmp`	**N**om **p**ropre, **m**asculin, **p**luriel
`Vmip3s-`	**V**erbe **m**ain (principal), **i**ndicatif, **p**résent, **3**e personne, **s**ingulier

Code	Lecture
`Vaip3s-`	**V**erbe **a**uxiliaire, même analyse

Le fichier `annexes/lexik_multext.tsv` donne, pour chacun des 387 codes de la base, la catégorie et les traits UD qu'il représente. La correspondance est exacte dans les deux sens : un code = une combinaison de traits, et inversement. Quelques codes ont été ajoutés au jeu GRACE d'origine pour y parvenir (déterminants et pronoms négatifs `Dn`, `Pn` ; exclamatifs `De`, `Pe` ; adverbes interrogatifs, relatifs, négatifs `Rip`, `Rrp`, `Rpn` ; `t` final pour un pronom tonique comme *moi*).

10. Les fréquences

10.1 Les trois corpus

Corpus	Registre	Taille
OpenSubtitles	oral scénarisé (sous-titres de films et de séries)	805 millions de mots
Wikipédia	écrit encyclopédique	841 millions de mots
Gutenberg	littérature du domaine public	45 millions de mots

- **Wikipédia** : articles du 1er juin 2026, texte des paragraphes (les dates, unités et nombres rendus en texte, les références retirées). Licence CC BY-SA 4.0.
- **OpenSubtitles** : corpus monolingue français diffusé par OPUS (fichier du 14 février 2025). OPUS ne précise pas de licence pour ce corpus.
- **Gutenberg** : 4 086 livres en français du Project Gutenberg, sans les en-têtes du projet. Domaine public aux États-Unis.

Les textes sont découpés en mots selon les règles du lexique : *j'aime* fait deux mots (*j'* + *aime*), *veux-tu* aussi (*veux* + *-tu*), *grand-père* et *aujourd'hui* un seul. LeXiK ne publie que des comptages, jamais le texte des corpus.

Le total de mots (la taille du corpus ci-dessus, qui sert de dénominateur) compte chaque unité du découpage une fois : chaque mot, chaque abrègement (*M.*, *SNCF*) et chaque formule (un nombre, une date, une mesure comme « 12 % » compte pour une unité). La ponctuation n'est pas comptée.

Les entrées de plusieurs mots (*pomme de terre*, *parce que*, les locutions) sont comptées **en plus**, sans rien retirer à leurs mots : dans « une pomme de terre », *pomme*, *de* et *terre* comptent chacun une occurrence, et la suite *pomme de terre* en compte une aussi. Le total ne change pas. C'est un comptage de surface : la suite est comptée chaque fois qu'elle apparaît en minuscules, qu'elle forme ou non l'unité (« en fait » compte aussi « il en fait trop »). Les fréquences de toutes les entrées ne s'additionnent donc pas à un million.

10.2 La fréquence d'une graphie (`freq_*`)

C'est le nombre d'occurrences de la graphie dans tout le corpus, **toutes analyses confondues**, ramené à un million de mots. *petit* a la même fréquence sur ses trois lignes (adjectif, nom, adverbe) : c'est la fréquence de la graphie. 33,7 % des analyses ont une fréquence ; les formes rares sont absentes des corpus.

10.3 La part de chaque analyse (`parts_*`)

Les parts disent comment la fréquence d'une graphie se répartit entre ses analyses. Pour les obtenir, un échantillon de chaque corpus a été analysé phrase par phrase par l'analyseur grammatical de Lectura, qui choisit pour chaque mot une des analyses de LeXiK. Cet analyseur est entraîné sur des treebanks et des corpus publics du français ; seules ses décisions, agrégées en parts, sont publiées.

Corpus	Mots analysés	Part du corpus
OpenSubtitles	28,8 millions	4 %
Wikipédia	38,5 millions	5 %
Gutenberg	22,7 millions	50 %

(Échantillons : une ligne sur 2 pour Gutenberg, une sur 28 pour OpenSubtitles, 1,3 million de paragraphes tirés au hasard pour Wikipédia.)

Fréquence d'une analyse = fréquence de la graphie × part.

Fréquence d'une analyse = fréquence de la graphie × part. Exemple, dans les sous-titres :

ortho	id_lemme	multext	freq_opensubtitles	parts_opensubtitles
mange	manger#VERB	`Vmip1s-;Vmip3s-;Vmmp2s-;Vmsp1s-;Vmsp3s-`	68.782	`0.291;0.664;0.010;0.010;0.020`
petit	petit#ADJ	`Afpms`	485.845	`0.944`
petit	petit#NOUN	`Ncms`	485.845	`0.056`
petit	petit#ADV	`Rgp`	485.845	`0.000`

Lecture : *mange* est à 29,1 % une 1re personne de l'indicatif (20,0 par million) et à 66,4 % une 3e personne (45,7 par million) ; l'impératif et le subjonctif se partagent le reste. Les 0,5 % qui manquent pour atteindre 100 % sont des occurrences que l'analyseur a lues autrement (erreur ou analyse absente de LeXiK).

Une part vide ne veut pas dire une fréquence nulle, mais que l'analyse n'a pas été rencontrée dans l'échantillon : sa fréquence est inconnue. 13,3 % des analyses ont une part. Pour un mot rare, la part repose sur peu d'occurrences et reste fragile.

10.4 La fréquence d'une locution

Une locution verbale apparaît rarement sous sa forme de citation : on lit « la réunion a eu lieu » plus souvent que « avoir lieu ». La fréquence d'une locution (`freq_*`) est donc toujours le **cumul de toutes ses formes**, et non le compte de la seule suite exacte.

Chaque locution est cherchée dans une phrase sur deux de chaque corpus sous toutes ses formes : verbe conjugué, y compris aux temps composés ; nom ou adjectif fléchi ; pronom réfléchi et possessif variables (*je m'en vais* pour *s'en aller*) ; une négation ou un adverbe peut s'insérer (*je ne m'en bats pas*) ; à l'impératif, le pronom suit le verbe (*tais-toi*, *dépêche-toi*). Les autres mots doivent être identiques. Le compte est ramené au million de mots de l'échantillon.

Un participe seul après un pronom n'est pas compté : dans « il me traite », *traite* n'est pas une forme de *se traire*. *Rendez-vous* et *voyez-vous* sont aussi écartés.

Locution	Formes comptées	freq_opensubtitles
avoir lieu	*a eu lieu*, *aura lieu*, *avait lieu*…	42,7
faire attention	*fais attention*, *fait attention*, *faisons attention*…	52,2

(Fréquences par million de mots dans les sous-titres.)

Un verbe pronominal est compté pour lui-même : *s'abaisser* (2,9 par million) a sa propre fréquence, distincte de celle du verbe *abaisser* dans LeXiK Mots. Une locution absente des échantillons a une fréquence vide.

Ce comptage a trois limites :

- le verbe n'est pas analysé en contexte : une suite homographe compte aussi (*en faire*, 461,2 par million, compte aussi « je vais en faire un gâteau ») ;
- *s'avoir* et *s'être* sont écartés, leurs formes se confondant avec les auxiliaires ;
- le gérondif après *en* n'est pas compté (*en faisant attention*).

10.5 La part pronominale d'un verbe

L'annexe `annexes/lexik_pronominaux.tsv` compare chaque verbe pronominal à son verbe simple : une ligne par couple, avec le verbe (`id_lemme` de LeXiK Mots), le pronominal, la fréquence de chacun par corpus et la part du pronominal dans les emplois du verbe. La fréquence du verbe (`freq_verbe_*`) est celle de son lemme dans LeXiK Mots, comptée sur les seules occurrences analysées comme ce verbe : les formes écrites seraient faussées par les homographes (*tu*, forme de *taire* mais surtout pronom). La part est le rapport entre la fréquence du pronominal et celle du verbe.

Verbe	Pronominal	Part dans les sous-titres
taire	*se taire*	65 %
battre	*se battre*	58 %

Verbe	Pronominal	Part dans les sous-titres
rendre	*se rendre*	36 %

La part est plafonnée à 1 : les deux fréquences ne sont pas mesurées de la même façon (le pronominal par recherche de la suite, le verbe par l'analyseur), et pour les verbes presque toujours pronominaux (*s'écrier*, *se fier*) le rapport peut dépasser 1. *S'avoir* et *s'être* n'ont pas de ligne. Un verbe peut avoir plusieurs pronominaux (une ligne pour chacun) : leurs parts ne s'additionnent pas forcément, une même occurrence pouvant relever de deux locutions.

10.6 La fréquence d'une forme spéciale

La fréquence d'une forme spéciale (`freq_*`) est celle de sa **graphie, tout court** : la forme est comptée telle qu'elle s'écrit dans les corpus, majuscules comprises (*CE* et *ce* sont comptés à part), et ramenée au million de mots. *km* apparaît 27,1 fois par million de mots dans les sous-titres, 475,1 dans Wikipédia.

Le comptage ne sait pas de quelle lecture il s'agit : **toutes les lectures d'une même forme ont la même fréquence**. Les deux lignes de *AB* (*Agriculture Biologique*, *assez bien*) portent le même chiffre ; *M* (*monsieur* ou la lettre *M*) aussi. Les distinguer demanderait de reconnaître le sens en contexte, ce que ni le comptage ni l'analyseur ne font. Il n'y a donc pas de parts (`parts_*`) pour les formes spéciales.

Seules 5 884 des 45 919 lignes ont une fréquence : les 37 988 éléments (morceaux de noms composés) n'apparaissent pas seuls dans les textes, et beaucoup de sigles et d'abréviations sont absents des échantillons.

11. L'indice de confiance

La colonne `confiance` dit d'où vient la prononciation et par quoi elle est confirmée. Du plus sûr au moins sûr :

Code	Sens	Analyses
`A`	au moins deux sources indépendantes donnent la même prononciation	181 927
`B`	une seule source, confirmée par une vérification indépendante	1 907 614
`C`	une seule source, non confirmée	197 257
`R`	relue et corrigée à la main	30 295
`P`	alignée sur une autre forme du même mot (le pluriel sur le singulier)	679
`K`	phrase reprise telle quelle du Wiktionnaire	1 549
`T`	reprise de l'inventaire des formes spéciales, relu à la main	2 230
`E`	sigle lu lettre par lettre	4 195
`N`	prononciation qui ne suit pas l'orthographe française (emprunt, nom étranger), confirmée	996
`M`	même cas, une seule source	69
`S`	construite à partir des segments d'un composé, ou reprise de la forme en minuscules	106 312
`G`	aucune source : prononciation prédite par le modèle de Lectura	587 436
`D`	aucune prononciation	414
–	sans prononciation (forme relevée dans les corpus, non analysée)	92 603

Les confiances `A` et `B` couvrent 67 % des analyses. Le modèle qui produit les prononciations `G` (un modèle graphème → phonème) est entraîné sur une version de travail antérieure de la base ; il complète les mots qu'aucune source ne prononce, en particulier beaucoup de noms propres.

La définition exacte de chaque code est dans la documentation de la provenance.

12. Les sources

LeXiK est une **compilation enrichie de sources lexicales libres**, croisées entre elles :

Source	Licence
Lexique 3.83 et 4	CC BY-SA 4.0
GLAFF	CC BY-SA 3.0
Wiktionnaire (via Kaikki)	CC BY-SA 3.0
Grammalecte (Dicollecte)	MPL 2.0
Wikidata (lexèmes)	CC0
ipa-dict	MIT
INSEE (prénoms, noms de famille)	Licence Ouverte 2.0
GeoNames	CC BY 4.0
Treebanks Universal Dependencies (GSD, ParisStories, Rhapsodie, PUD)	CC BY-SA

D'autres ressources (Morphalou, et les treebanks Sequoia et FQB, sous LGPL-LR) ont servi au contrôle, sans redistribution de leurs données.

Le dossier `provenance/` donne, pour chaque prononciation, les sources qui l'ont proposée, les variantes écartées et la raison du choix ; sa documentation (`provenance/DOCUMENTATION_PROVENANCE.md`) décrit en détail la construction de la base.

Modèles et algorithmes

LeXiK s'appuie sur deux modèles et deux algorithmes propres à Lectura.

Les modèles, entraînés sur des corpus publics du français :

- le **modèle graphème → phonème** prononce les formes qu'aucune source ne prononce (confiance `G`) et sert de vérification indépendante des prononciations données par une seule source (confiance `B`) ;
- l'**analyseur morphosyntaxique** étiquette les échantillons de corpus dont sont tirées les parts de chaque analyse dans la fréquence (§10.3).

Les algorithmes, dont le code est publié (paquet Python `lectura-aligneur` sur PyPI, https://pypi.org/project/lectura-aligneur/ ; code source sur GitHub, https://github.com/maxcarriere/lectura-modules) :

- l'**aligneur graphème–phonème** : une recherche exhaustive en profondeur avec retour arrière, guidée par une table de correspondances entre sons et graphèmes (§7.3) ;
- le **syllabeur** : la syllabation par attaque maximale (§8.1).

13. Limites connues

1. **Fréquences par analyse** : seules 13,3 % des entrées ont une part (les formes rares sont absentes des échantillons ; les locutions et les formes spéciales n'en ont presque jamais).
2. **Noms propres** : beaucoup sont prononcés par le modèle (confiance `G`) ; les noms étrangers peuvent être mal prononcés.
3. **Alignement** : les entrées des sous-lexiques Mots et MultiMots sans alignement ont une graphie dont l'aligneur ne connaît pas les correspondances, surtout des noms propres étrangers (*Krajewski*, *Morici*) ; leur prononciation reste renseignée (§7.3). Les formes spéciales et non classées n'en ont jamais.
4. **Une prononciation de référence** : les variantes régionales ne sont pas couvertes, sauf les variantes libres notées `{a|b}`.
5. **Fréquence des locutions** : elle compte la locution sous toutes ses formes, mais le verbe n'est pas analysé en contexte : une suite homographe compte aussi ; le gérondif n'est pas compté (§10.4).
6. **Éléments des formes spéciales** : prononcés d'après leur composé, non relus un à un.
7. **Formes non classées** : non vérifiées ; les candidates issues des corpus peuvent être des fautes de frappe, des noms étrangers ou des fragments.
8. **Sens** : LeXiK ne distingue pas les sens d'un même lemme, sauf quand ils se fléchissent différemment (*ciel*, pluriels *cieux* et *ciels*).

14. Format des fichiers

- Texte UTF-8, sans BOM ; une tabulation entre les colonnes ; une ligne d'en-tête.
- Pas de guillemets autour des valeurs ; une valeur vide est une chaîne vide.
- Nombres : point décimal, trois décimales.
- Dans une cellule : `;` sépare les analyses (`multext`, `parts_*`) et les propriétés (`proprietes`) ; `,` sépare les valeurs d'une propriété (`genre:m,f`) ; une espace sépare les couples de `alignement` et les segments de `composition`.
- Fichiers compressés en gzip ; `checksums.sha256` permet de vérifier leur intégrité (`sha256sum -c checksums.sha256`).

Licence, citation et contact : voir `LISEZ-MOI.md`.