

LeXiK — Construction et provenance

Version 1.0, octobre 2026 (base LeXiK 1.2.3 du 14/10/2026 13:25).

Ce document décrit comment LeXiK a été construit et comment lire les deux fichiers de ce dossier. Pour l'usage de la base elle-même (colonnes, notations), voir le guide à la racine du dépôt (DOCUMENTATION.md), dont les numéros de section sont cités ici sous la forme « guide, §6 ».

Sommaire

1. Contenu du dossier
2. Les sources
3. Le nettoyage des transcriptions
4. Le choix d'une prononciation
5. Les prononciations qui ne s'alignent pas
6. L'indice de confiance en détail
7. Le fichier `lexik_provenance.tsv`
8. L'origine des propriétés
9. La ressource de contrôle : Morphalou
10. Les divergences avec les sources
11. Annexe A : matrices de consensus phonétique

1. Contenu du dossier

Fichier	Contenu
<code>lexik_provenance.tsv.gz</code>	pour chaque entrée du fichier des formes, l'origine de sa prononciation : sources, propositions de chaque source, choix, corrections
<code>lexik_lemmes_provenance.tsv.gz</code>	pour les lemmes concernés, l'origine de leurs propriétés et les avis divergents des sources
<code>DOCUMENTATION_PROVENANCE.md</code> / <code>.pdf</code>	ce document

2. Les sources

Source	Licence	Usage
Lexique 3.83 et 4.10	CC BY-SA 4.0	formes, analyses, prononciations
GLAFF 1.2.2	CC BY-SA 3.0	formes, analyses, prononciations
Wiktionnaire, via Kaikki	CC BY-SA 3.0	formes, analyses, prononciations, propriétés des lemmes
Grammalecte / Dicollecte 7.7	MPL 2.0	formes, analyses, propriétés des verbes
Wikidata, lexèmes	CC0	formes, analyses, prononciations
ipa-dict fr_FR	MIT	prononciations (confirmation)
INSEE, prénoms et noms de famille	Licence Ouverte 2.0	noms propres
GeoNames FR	CC BY 4.0	noms de lieux
Treebanks UD GSD, ParisStories, Rhapsodie, PUD	CC BY-SA	analyses des mots grammaticaux, composés à espace, contrôle
Morphalou 3.1	LGPL-LR	contrôle, sans redistribution (§9)
Treebanks UD Sequoia, FQB	LGPL-LR	contrôle, sans redistribution
Treebank UD ParTUT	CC BY-NC-SA	exclu

Les sources sont regroupées en **familles** : deux sources d'une même famille se recopient et ne comptent pas comme deux avis indépendants.

Famille	Sources
Lexique	Lexique 3.83, Lexique 4

Famille	Sources
Wiktionnaire	GLAFF, Kaikki
Wikidata	Wikidata
ipa-dict	ipa-dict (confirmation seulement, compte pour une demi-voix)

Lexique, Wiktionnaire et Wikidata sont dites **importables** : leurs prononciations peuvent être reprises telles quelles.

Une dernière origine, cadre, n'est pas une ressource extérieure : ce sont les analyses que LeXiK pose lui-même pour les mots grammaticaux. Les déterminants, pronoms, prépositions et conjonctions reçoivent les analyses relevées dans les treebanks libres (GSD, ParisStories, Rhapsodie, PUD), complétées par celles des sources libres et justifiées une à une ; les éléments postposés (*-ce*, *-ci*, *-là*, *-t-...*) viennent d'une courte liste fixée par la spécification. Leur prononciation reste votée par les sources.

3. Le nettoyage des transcriptions

Avant toute comparaison, les transcriptions des sources sont nettoyées et ramenées à l'inventaire de LeXiK : notations converties, marques retirées, consonnes doubles réduites, rejet de tout caractère hors inventaire (le détail est dans le guide, §6.5).

Puis chaque transcription est ramenée aux **conventions** de LeXiK (règles C1 à C12 ; guide, §6.3). Une fois ces conventions neutralisées, l'accord entre sources passe de 57-99 % à 97,6-99,5 % selon les paires : la plupart des désaccords apparents n'étaient que des différences de notation (annexe A).

4. Le choix d'une prononciation

4.1 Recevable ou rejetée

Chaque prononciation nettoyée est soumise à l'**aligneur** (guide, §7). Si elle s'aligne sur l'orthographe, elle est **recevable** : elle est gardée, même minoritaire, comme variante attestée. Sinon, elle est **rejetée** : elle ne subsiste que dans la provenance, avec le diagnostic de l'aligneur.

4.2 Le vote

Parmi les prononciations recevables, on regroupe celles qui donnent la même forme après normalisation, puis **les familles votent** : chaque famille compte pour une voix (ipa-dict pour une demi-voix, et seulement en appui d'une autre). La forme qui a le plus de voix est retenue. En cas d'égalité, un ordre de priorité entre sources tranche ; quand l'égalité ne porte que sur l'ouverture d'une voyelle (*e/ɛ*, *o/ɔ*, *ø/œ*), les deux formes sont gardées comme variante {a|b}.

4.3 Après le vote

Le vote se fait entrée par entrée. Plusieurs étapes rendent ensuite la base cohérente d'une entrée à l'autre ; chacune est tracée dans la colonne corrections de la provenance :

- **paradigme** : un pluriel dont la prononciation contredit celle de son singulier de même genre est aligné sur lui (confiance P) ;
- **ouverture des voyelles** : deux analyses d'une même graphie qui ne diffèrent que par *e/ɛ*, *o/ɔ* ou *ø/œ* sont alignées sur la plus sûre ;
- **h aspiré** : rendu cohérent dans un lemme (*haltes* comme *halte*), entre les catégories d'une même graphie, et dans les composés, qui le prennent de leur premier segment ; une liste relue retire les marques fautives des sources (*hyène*, *humeur*) ;
- **majuscules** : une forme à majuscule prend la prononciation de la même forme en minuscules ;
- **règles finales** : *-enne* jamais nasal (*bolivarienne* /...ʁjɛn/) ; *ex-* initial toujours /ɛks/ ou /ɛgz/ ;
- **relecture** : des surcharges manuelles, chacune avec son motif (confiance R).

4.4 Quand aucune source ne prononce le mot

Le mot est prononcé par le **modèle graphème → phonème de Lectura**, un modèle entraîné sur une version de travail antérieure de la base, qui prédit une prononciation à partir de la seule orthographe (confiance G). Un mot composé sans prononciation propre reçoit celle de ses segments mis bout à bout (confiance S).

5. Les prononciations qui ne s'alignent pas

Un refus de l'aligneur est un signal fort : dans la grande majorité des cas, la transcription de la source est fautive (phonème en trop, transcription d'un autre mot ou d'une partie du mot). Mais ce n'est pas une preuve : les emprunts (*checkout*, *Kōji*) ne suivent pas l'orthographe française.

Quand **aucune** prononciation d'une entrée ne s'aligne, les candidates sont votées telles quelles. Puis la **règle des non-alignées** s'applique à une gagnante qui ne vient que d'une source : elle n'est gardée que si

1. le modèle graphème → phonème la propose parmi ses dix meilleures prononciations, ou
2. le mot, son lemme ou sa base sans préfixe ni terminaison française (*recoacher* → *coach*) figure dans un lexique anglais.

Sinon, elle est remplacée par la prédiction du modèle (confiance G) et marquée rejetée (motif « règle M ») dans la provenance.

Les entrées gardées ont la confiance N (confirmée par deux sources ou par le modèle) ou M (une seule source, retenue par la règle).

6. L'indice de confiance en détail

Code	Sens	Analyses
A	au moins deux familles importables donnent la même forme normalisée	181 927
B	une seule famille importable, confirmée par ipa-dict, par le modèle graphème → phonème ou par la ressource de contrôle (§9)	1 907 614
C	une seule famille, non confirmée (le modèle diverge ou ne s'applique pas)	197 257
R	surcharge de relecture, avec son motif	30 295
P	choisie par le contrôle de paradigme contre le vote	679
K	prononciation de la phrase donnée par le Wiktionnaire, sans vote	1 549
T	prononciation posée par l'inventaire des formes spéciales (tiré du Wiktionnaire, complété et relu à la main)	2 230
E	sigle épilé : aucune prononciation de source, lu lettre par lettre (<i>SNCF</i> / <i>ɛs.en.se.ef</i>)	4 195
N	non alignable, confirmée par deux familles ou par le modèle	996
M	non alignable, une seule source, gardée par la règle des non-alignées (§5)	69
S	sans source : composée des segments, ou reprise de la forme en minuscules	106 312
G	prononciation du modèle graphème → phonème : mot sans source, mot attesté sans prononciation, ou candidate écartée par la règle des non-alignées	587 436
D	aucune prononciation	414
–	forme relevée dans les corpus, sans prononciation ni analyse	92 603

7. Le fichier `lexik_provenance.tsv`

7.1 Colonnes

Une ligne par entrée (graphie, lemme, catégorie). Le fichier se joint au fichier des formes par (ortho, lemme, upos). Deux homonymes numérotés de même graphie (*livre#NOUN#1*, *livre#NOUN#2*) ont la même prononciation et partagent une ligne. Les fichiers des locutions et des formes spéciales n'ont pas de colonne lemme : ils se joignent à la provenance par (ortho, upos) ; pour une forme spéciale, le lemme de la provenance est la forme elle-même.

Colonne	Contenu
ortho, lemme, upos	l'entrée
phone	la prononciation retenue, forme brute (guide, §6.4)
confiance	l'indice de confiance (§6)
sources	les sources qui attestent l'analyse
candidates	les prononciations proposées, chacune avec ses sources et son statut (§7.2)

Colonne	Contenu
regles	les conventions (C1 à C12) qui ont servi à normaliser les candidates
corrections	les modifications après le vote, avec leur motif (§4.3)
marques	les marques ajoutées à la prononciation (consonne latente, h aspiré) et leur origine
controle_morphalou	identique, différent ou absent : accord de la ressource de contrôle avec la prononciation retenue (§9)
confiance_sans_controle	la confiance recalculée sans la ressource de contrôle (§9)

7.2 Lire la colonne candidates

Exemple : *petit* adjectif, retenu /pə.ti/.

```
pə.ti ← glaff,kaikki,lexique383,lexique4 [canonique]
pti ← kaikki [canonique]
pət.si ← kaikki [rejetee] (pos=3/5 reste='it' phone=/s/ ... attendu=[s,ss,c,ç,sç,t...])
ptsi ← kaikki [rejetee] (...)
```

Chaque proposition est suivie des sources qui la donnent et de son statut :

Statut	Sens
[canonique]	une fois normalisée, elle donne la prononciation retenue : elle a voté pour elle
[attestee]	recevable, mais elle donne une autre forme, qui a perdu le vote : variante attestée
[rejetee]	refusée par l'aligneur (le diagnostic suit) ou contenant un symbole hors inventaire
[liaison]	forme de liaison donnée par la source (<i>dix</i> /diz/) : elle ne vote pas, elle sert à poser la consonne latente

Lecture : quatre sources donnent /pə.ti/. Le Wiktionnaire donne aussi /pti/, sans *e* muet ; la règle C4 (le *e* muet intérieur est toujours noté) le ramène à /pə.ti/, il compte donc pour la même forme, d'où [canonique]. Le Wiktionnaire donne enfin deux transcriptions où le *t* se prononce [ts] (sans doute la prononciation québécoise) ; l'aligneur les refuse, et le diagnostic dit où : à la lettre 3 sur 5, il reste *it*, et /s/ ne peut pas s'écrire *i*.

Quand aucune candidate ne s'aligne, la gagnante est [canonique] mais garde le diagnostic de l'aligneur entre parenthèses (§5).

8. L'origine des propriétés : `lexik_lemmes_provenance.tsv`

Une ligne par lemme dont une propriété vient d'une source ou d'une règle (839 263 lemmes en tout, dont une minorité concernée). Les propriétés lues sur les formes elles-mêmes (genre, pluriel, féminin...) n'ont pas de provenance à tracer.

Colonne	Contenu	Exemple (<i>manger</i>)
id_lemme	le lemme	manger#VERB
provenance	l'origine de chaque propriété, en propriété=origine, plusieurs origines concordantes jointes par +	groupe=calcul;transitivite=grammalecte+kaikki;pronominal=kaikki;auxiliaire=grammalecte
conflits	les avis divergents des sources, non retenus	invariable:kaikki=oui (pour <i>Avesta</i>)
cellules	pour un verbe, le nombre de cases de conjugaison attestées sur 48 (24 à l'indicatif, 12 au subjonctif, 6 au conditionnel, 3 à l'impératif, l'infinitif et les deux participes)	48/48

Les origines possibles : kaikki (Wiktionnaire), grammalecte, calcul (déduite par LeXiK, par exemple le groupe d'après la terminaison de l'infinitif), tete:<verbe> (héritée du verbe de base d'un composé : *auto-endurcir* → tete:endurcir) et défaut (aucune source, valeur par défaut).

Les locutions, les formes spéciales et les formes non classées n'ont pas de fichier des lemmes : l'origine de leurs propriétés est dans la colonne `provenance_proprietes` de `lexik_provenance.tsv`, sur la ligne de l'entrée, dans le même format que la colonne `provenance` ci-dessus.

9. La ressource de contrôle : Morphalou

Morphalou 3.1 (ATILF / CNRS) est diffusé sous licence LGPL-LR, incompatible avec la licence CC BY-SA de LeXiK. Il n'est donc **jamais une source** de LeXiK : il sert de **ressource de contrôle**, c'est-à-dire que ses valeurs sont comparées aux valeurs retenues, sans jamais être reprises.

- Morphalou ne vote pas : une prononciation ou une analyse qu'il est seul à donner n'est pas retenue.
- Ses marques de h aspiré et ses formes de liaison ne sont pas lues.
- Le modèle graphème → phonème ne lit que l'orthographe et la catégorie, jamais les propositions des sources.
- Vérification faite sur les entrées dont la prononciation est votée entre les sources (confiances A, B, C, M, N) : chacune a au moins une proposition d'une source libre qui vote pour la prononciation retenue.

Ce que Morphalou apporte, c'est une **confirmation** : une prononciation donnée par une seule famille libre est B si Morphalou ou le modèle la confirme, C sinon. Le code de confiance peut donc dépendre de Morphalou ; la prononciation, non. La colonne `confiance_sans_controle` donne la confiance recalculée sans aucune ressource LGPL-LR : elle ne diffère de confiance que pour 5 309 entrées (B → C), les autres confirmations de Morphalou étant aussi données par le modèle.

Dans le fichier de provenance, Morphalou est absent de sources et de candidates : aucune de ses valeurs n'est reproduite. La colonne `controle_morphalou` dit seulement s'il confirme (identique), s'il diffère (different) ou s'il n'a pas l'entrée (absent).

Si l'ayant droit de Morphalou estimait que cet usage excède la comparaison, l'auteur de LeXiK s'engage à retirer la colonne `controle_morphalou` et à remplacer `confiance` par `confiance_sans_controle`. Contact : voir `LISEZ-MOI.md`.

10. Les divergences avec les sources

LeXiK ne corrige pas ses sources en silence. Pour chaque source, la liste des entrées où LeXiK diverge d'elle, avec la valeur de la source, la valeur de LeXiK et le motif, est **transmise aux auteurs de la source**. Ce sont des signalements, pas des corrections : dans une divergence de fond, la source peut avoir raison.

Les divergences sont classées, de la plus anodine à la plus suspecte :

Type	Sens
convention	même prononciation, notée autrement (syllabation, ouverture d'une voyelle, consonne double) : pas une erreur
correction	la valeur de la source a gagné le vote, puis a été modifiée (§4.3)
variante	la valeur s'aligne sur l'orthographe, mais le vote en a retenu une autre
rejet	la valeur ne s'aligne pas sur l'orthographe, ou a été écartée par la règle des non-alignées : la plus probablement fautive
analyse	une analyse grammaticale de la source a été écartée (forme verbale sans son paradigme, lemme minoritaire, suite de mots prise pour un mot)

Trois signaux ont servi à repérer les erreurs :

1. **le croisement des sources** : chaque prononciation est votée entre familles indépendantes ; une valeur minoritaire n'est jamais perdue, la provenance la garde ;
 2. **le refus d'alignement** : une transcription que l'aligneur ne parvient pas à mettre en correspondance avec l'orthographe est suspecte (§5) ;
 3. **le contrôle par le modèle** : pour une entrée à source unique, la prédiction indépendante du modèle graphème → phonème confirme la valeur ou la met en doute ; les désaccords ont été relus par classes.
-

Annexe A. Matrices de consensus phonétique

Ces mesures justifient les conventions C1 à C12 (guide, §6.3) : pour chaque question, la part des prononciations de chaque source qui suit chaque option. Elles ont été faites le 1^{er} octobre 2026 sur les sources normalisées vers l'inventaire, avant tout arbitrage. Les règles sont repérées à partir de l'orthographe seule, donc sur des sous-ensembles où leur application est certaine ; la syllabation utilisée est une approximation. Les effectifs sont des nombres de prononciations. GLAFF et Kaikki, tous deux tirés du Wiktionnaire, ne sont pas indépendants. La ressource de contrôle (§9) n'est pas reproduite ici.

C2 — o/ɔ en syllabe ouverte non finale (graphie o)

Source	ɔ (graphique)	o (position)	n
lexique	39,0 %	61,0 %	20 113
glaff	82,2 %	17,8 %	116 864
kaikki	78,5 %	21,5 %	210 411
ipadict	93,4 %	6,6 %	23 575

C2 — e/ɛ pour « ai » en syllabe ouverte non finale

Source	ɛ (graphique)	e (position)	n
lexique	84,5 %	15,5 %	600
glaff	79,2 %	20,8 %	2 287
kaikki	84,4 %	15,6 %	2 895
ipadict	78,8 %	21,2 %	697

C3 — schwa final (graphie -e)

Source	ə final absent	ə final noté	n
lexique	99,9 %	0,1 %	26 720
glaff	99,9 %	0,1 %	74 475
kaikki	100,0 %	0,0 %	138 371
ipadict	100,0 %	0,0 %	22 300

C4 — schwa interne facultatif

Source	ə interne noté	ə interne absent	n
lexique	90,1 %	9,9 %	30 386
glaff	90,1 %	9,9 %	367 573
kaikki	93,5 %	6,5 %	472 588
ipadict	91,9 %	8,1 %	56 054

C5 — schwa initial de re- noté ø ou ə

Source	ø	ə	n
lexique	4,4 %	95,6 %	7 806
glaff	0,0 %	100,0 %	190 831
kaikki	0,0 %	100,0 %	201 876
ipadict	0,0 %	100,0 %	9 214

C6 — i + voyelle après consonne + liquide

Source	diérèse (i)	diérèse (ij)	synérèse (j)	n
lexique	13,0 %	78,4 %	8,6 %	1 302
glaff	5,0 %	67,5 %	27,5 %	8 155
kaikki	3,7 %	90,2 %	6,1 %	10 825

Source	diérèse (i)	diérèse (ij)	synérèse (j)	n
ipadict	2,1 %	62,1 %	35,8 %	1 991

C7 — u + voyelle (hors qu, gu, ou)

Source	semi-voyelle u	voyelle y	n
lexique	89,2 %	10,8 %	3 319
glaff	91,0 %	9,0 %	16 265
kaikki	91,1 %	8,9 %	26 188
ipadict	94,2 %	5,8 %	4 651

C8 — -isme

Source	izm	ism	n
lexique	72,6 %	27,4 %	833
glaff	1,8 %	98,2 %	2 657
kaikki	5,4 %	94,6 %	6 167
ipadict	0,0 %	100,0 %	936

C9 — b devant s/t

Source	assimilée (ps/pt)	non assimilée (bs/bt)	n
lexique	88,5 %	11,5 %	607
glaff	45,4 %	54,6 %	3 357
kaikki	40,7 %	59,3 %	4 550
ipadict	99,0 %	1,0 %	821

C10 — gn intervocalique

Source	gn	autre (ŋ, gn)	nj	n
lexique	91,6 %	4,5 %	3,9 %	2 462
glaff	94,4 %	5,1 %	0,5 %	14 519
kaikki	94,3 %	4,9 %	0,8 %	21 330
ipadict	95,0 %	4,8 %	0,2 %	3 164

C11 — rr des futurs/conditionnels de courir, mourir, acquérir

Source	rr	rrr	n
lexique	70,8 %	29,2 %	89
glaff	67,2 %	32,8 %	235
kaikki	66,4 %	33,6 %	232
ipadict	100,0 %	0,0 %	102

C12 — finales verbales -ai / -ais

Source	-ai → e	-ais → e	-ai → e	-ais → e	n
lexique	24,7 %	47,4 %	27,9 %	0,0 %	5 585
glaff	0,0 %	50,1 %	49,8 %	0,0 %	52 194
kaikki	0,0 %	50,2 %	49,8 %	0,0 %	69 619
ipadict	49,2 %	50,2 %	0,5 %	0,0 %	10 709

C1 — voyelles moyennes en syllabe finale : désaccords entre sources

Paire	mots comparés	désaccord sur la voyelle finale	exemple
lexique – glaff	83 286	3,3 % (2 720)	embrouillerai ['äbɣujəɛ'] / ['äbɣujəɛ', 'äbɣujɛ']
lexique – kaikki	91 002	3,3 % (2 972)	embrouillerai ['äbɣujəɛ'] / ['äbɣujəɛ']
lexique – ipadict	69 779	2,7 % (1 887)	atomiserai ['atomizəɛ'] / ['atomizəɛ']
glaff – kaikki	442 916	0,1 % (540)	hydrospeeder ['idɣospidəɛ'] / ['idɣospidəɛ']
glaff – ipadict	107 571	8,9 % (9 613)	défavorisai ['defavəɣize'] / ['defavəɣize']
kaikki – ipadict	112 611	8,7 % (9 745)	défavorisai ['defavəɣize'] / ['defavəɣize']

Couverture : part des prononciations qui rejoignent la classe majoritaire après neutralisation des différences dérivables (R1-R9)

Source	prononciations comparées	majoritaire	minoritaire
lexique	180 726	99,03 %	0,97 % (1 761)
glaff	395 078	98,00 %	2,00 % (7 886)
kaikki	413 371	99,30 %	0,70 % (2 897)
ipadict	243 002	99,38 %	0,62 % (1 507)

Synthèse

Règle	Consensus mesuré	Décision
C2 o non final	ɔ majoritaire (79-97 %) ; Lexique partagé (39 / 61)	ɔ d'après la graphie
C2 ai non final	ɛ 79-98 %	ɛ
C3 schwa final	absent 99,9 %	jamais noté
C4 schwa interne	noté 90-94 %	noté
C5 schwa de <i>re-</i>	ə ≥ 96 %	ə
C6 diérèse après consonne + liquide	62-96 %	diérèse
C7 semi-voyelle	ɥ 89-94 %	semi-voyelle
C8 <i>-isme</i>	ism 95-100 % (Lexique izm 73 %)	ism
C9 assimilation de voisement	assimilée chez Lexique et ipa-dict ; Wiktionnaire partagé	assimilée, alternative notée
C10 <i>gn</i>	ɲ 92-95 % ; le reste = mots savants en gn	ɲ et liste d'exceptions
C11 consonnes doubles	ɸ simple 66-100 %	jamais doublée
C12 <i>-ai</i> / <i>-ais</i>	<i>-ais</i> → ɛ 100 % ; <i>-ai</i> → e chez la majorité, ɛ chez ipa-dict	e / ɛ