LIDILEM

UNIVERSITÉ Grenoble Alpes

Aligner production et normalisation : une première approche pour l'étude d'écrits scolaires

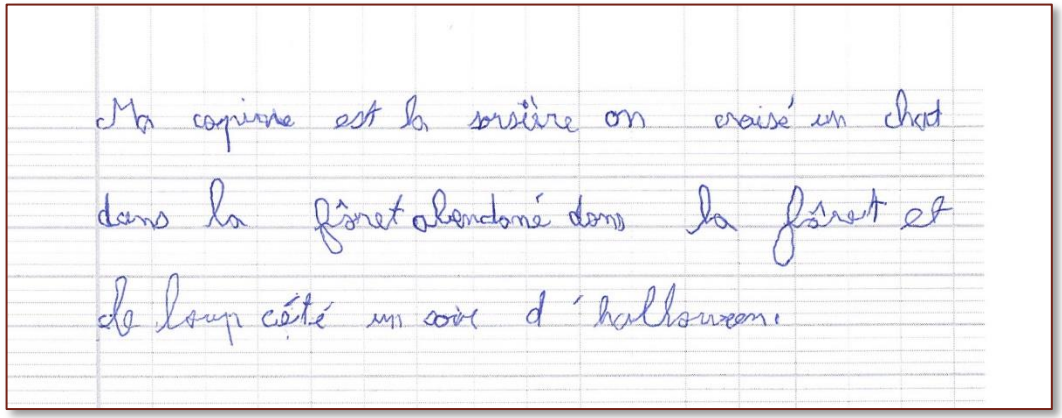
Claire Wolfarth

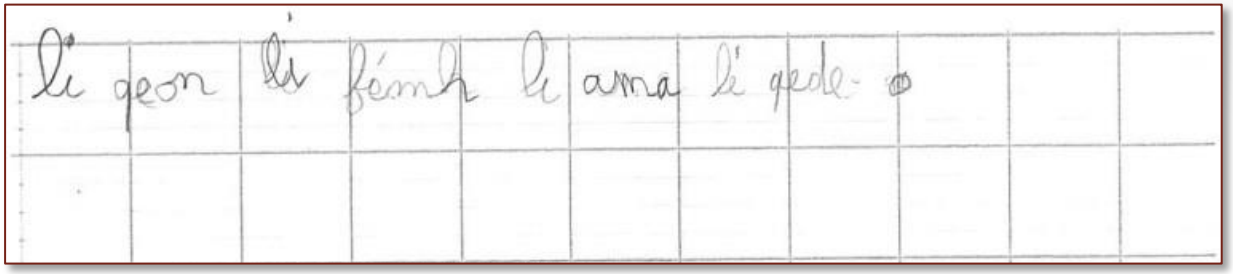
Lidilem

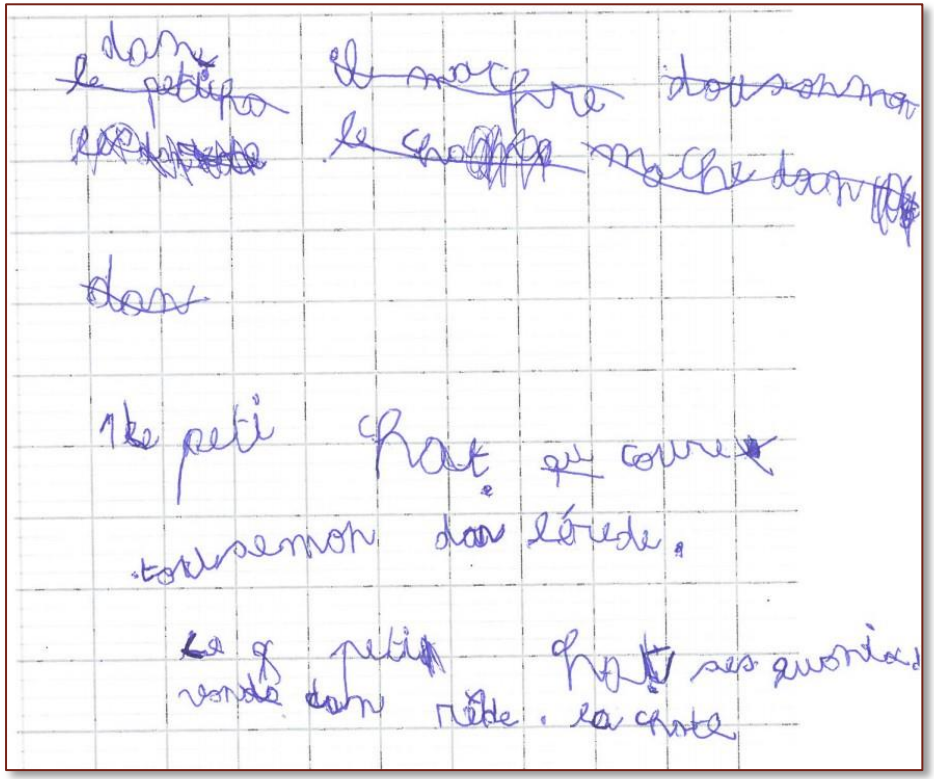
Université Grenoble Alpes

Quelques exemples

Il était une fois un loup qui voiaient un petit chaperon rouge ~~sortais~~ qui de sa maison le petit chaperon rouge allez voir sa grand-mère qui était malade et lui apporter du lait et un gâteau le petit chaperon rouge rencontre un loup le loup dit quest ce que tu fais la le petit chaperon rouge dit j'apporte du lait et un gâteau a ma ~~maman~~ grand-mère qui est malade le loup dit je passe par se chemin et toi l'autre le petit chaperon rouge dit d'accord le petit chaperon rouge ses fait avoir le loup et passé par le chemin le plus cour et le petit chaperon rouge a prise le plus long le loup arriva il ~~dista~~ toc toc toc toc la grand-mère dit qui est la le loup dit ses moi le petit chaperon rouge qui t'apporte du lait et un gâteau la grand-mère dit rentre donc mon enfant le loup rentra et mangent la grand mère il se dégisent an grand-mère le petit chaperon rouge arriva et dit toc toc toc le loup dit qui est la le petit chaperon rouge dit ses moi le petit chaperon rouge qui t'apporte du lait et un gâteau le loup dit entre mon enfant le petit chaperon rouge dit bien manger sa va te guérire le petit chaperon rouge dit que vous avez de grande dent que vous avez de grandes oreilles que vous avez de grandes bouche et oppe le petit chaperon rouge sais fait mangaien par le loup.


Ma copine est la sorcière on croisé un chat
dans la fôret abandoné dans la fôret et
de loup côté un cœur d'halloween.


li geon li fémh li ama li gede


le petit Rat qui court
dans la forêt.

Un chat marche sur la route.

Le chat tombe sur le mur.

Le chat pleure.

La maman dit à la maman par la bouche.

le b a é r t r a t i

g l i c e → « glisse »

p l e r → « pleure »

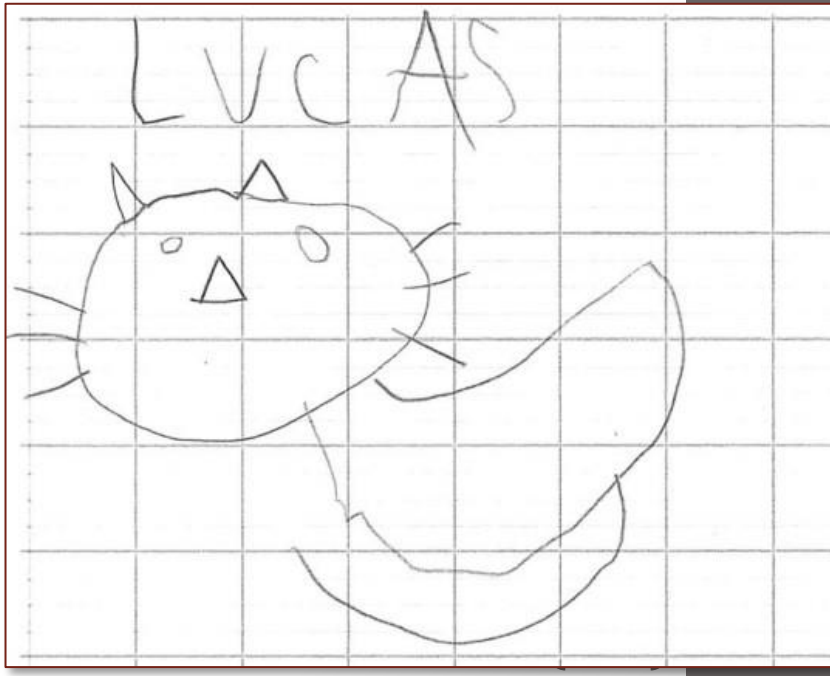
→ « s'est fait mal »

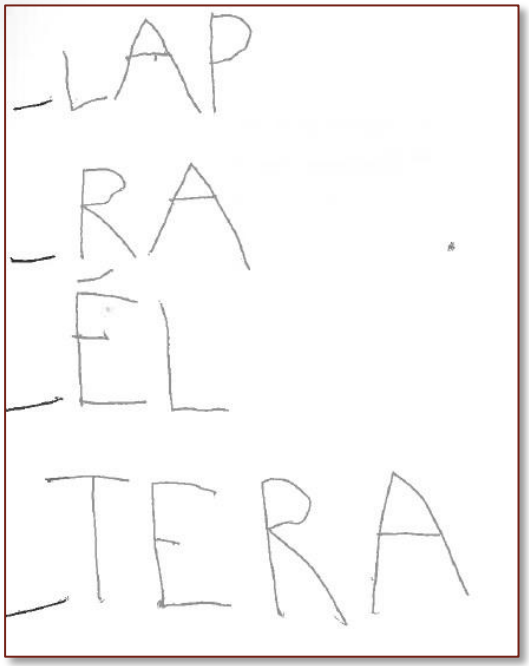
é r t c é s t f é m a e e

ē e t f e m a l o r i é

↳ « s'est fait mal au pied »

47





1/ ~~petit~~ pattes

2/ Patison

3/ cadre

4/ cadeau récra

5/ ~~récréation~~ charitable

6/ manifik

les salade salade vert - les jardin

les jeune caton picore le blé a vec les poule noirs.

iAH
AHU
MNAA
HHUUA

1/ pattin

2/ pattson

3/ capuchon

4/ récréation

5/ charitabl

6/ manifique

En été, les salade ~~vert~~ pousse dans les jardins

Les jeune caton ~~pico~~ picore le blé avec la poule noir.

Corpus scolaires : état des lieux

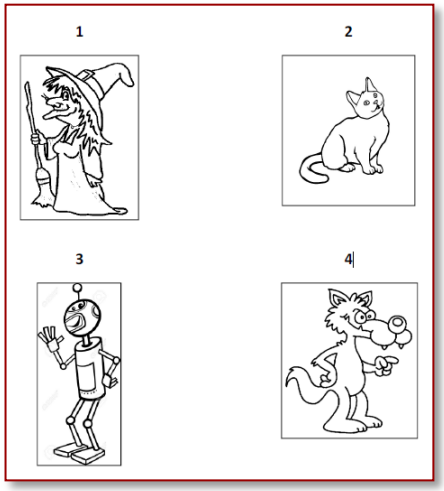
- Développement des recherches sur l'écriture
- Emergence de projets de constitution de grands corpus
  - Elalouf et al., 2005 ;
  - Gunnarsson-Largy et Auriac-Slusarczyk, 2013 ;
  - Garcia-Debanc, 2014 ;
  - David et Doquet, 2016;
  - *Scoledit*
- Annotation manuelle des corpus scolaires

I – Corpus scolaire : *Scoledit*

Scoledit

- Corpus longitudinal de textes scolaires et de dictées
  - 5 ans (CP – CM2)
  - 5 académiques, 40 écoles
  - 800 – 1200 élèves par an
  - Texte + dictée

  http://otus.u-grenoble3.fr/scoledit/



Enjeux

- Pour la linguistique :
  - Description en synchronie et en diachronie des caractéristiques des écrits d'apprenants de 6 à 11 ans
  - Meilleure connaissance des phénomènes d'acquisition de l'écriture (évolution de l'acquisition de la syntaxe, de l'orthographe, de la ponctuation, …)
- Pour la didactique:
  - Développement de séquences et de dispositifs à partir de ces connaissances
- Pour le traitement automatique des langues (TAL) :
  - Type de textes encore peu étudiés

Méthodologie

Constats :

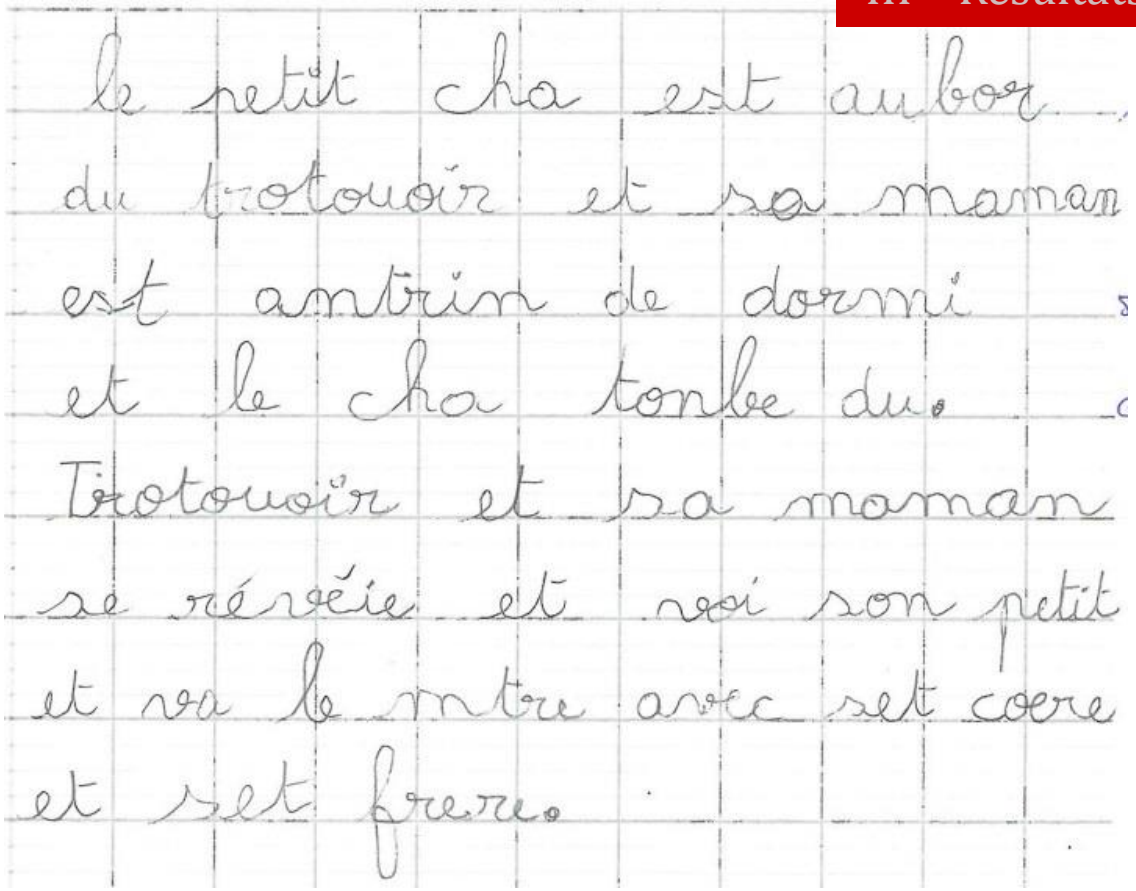
- Erreurs nombreuses
- Erreurs peu régulières

Hypothèse :

- Comparer la production des apprenants avec une norme de cette production devrait permettre de faciliter son annotation

➢ Développement d'un aligneur tokens par tokens des productions transcrites et des productions normées

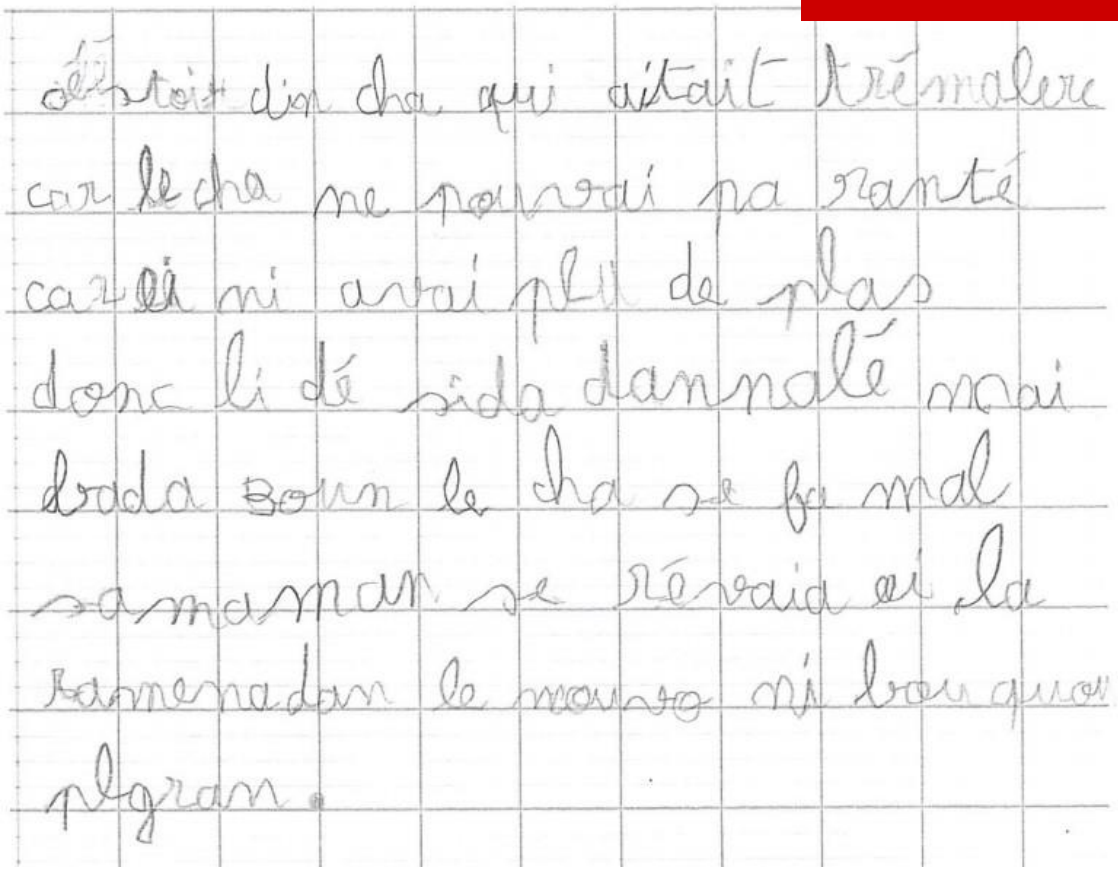
Spécificités du corpus



Production de l'élève 96 en fin de CP

Version orthographiée selon la norme : « Le petit chat est au bord du trottoir et sa maman est en train de dormir et le chat tombe du trottoir et sa maman se réveille et voit son petit et va le mettre avec ses sœurs et ses frères. »

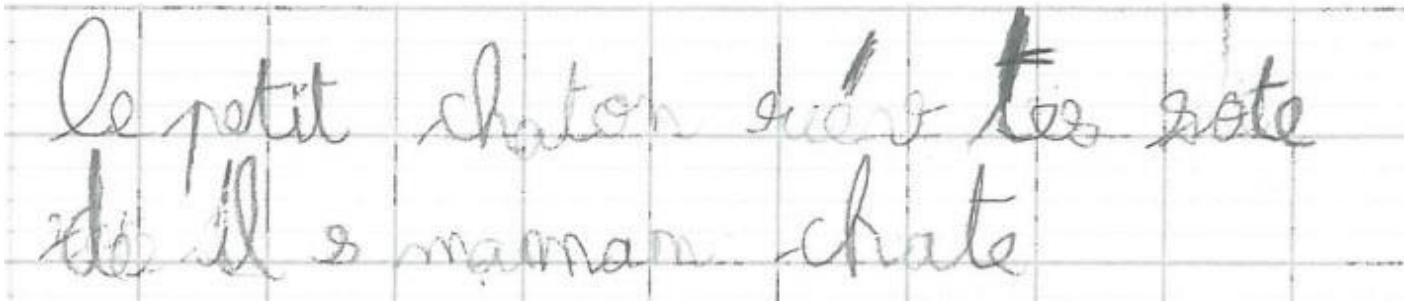
Spécificités du corpus



Production de l'élève 2972 en fin de CP

Version orthographiée selon la norme : « C'est l'histoire d'un chat qui était très malheureux car le chat ne pouvait pas rentrer car il n'y avait plus de place donc il décida de s'en aller mais badaboum le chat se fait mal <segmentation/> sa maman se réveilla et le ramena dans le nouveau nid beaucoup plus grand. »

Spécificités du corpus



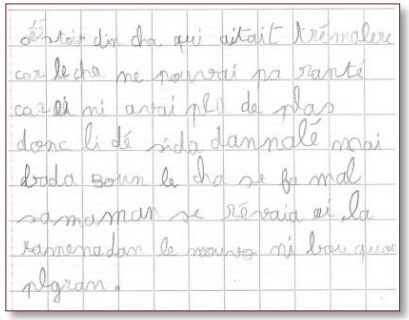
Production de l'élève 2453 en fin de CP

Version orthographiée selon la norme : *« le petit chaton se réveille et saute de […] maman chatte. »*

Composition du corpus

Pour chaque production :

- Scan



- Transcription

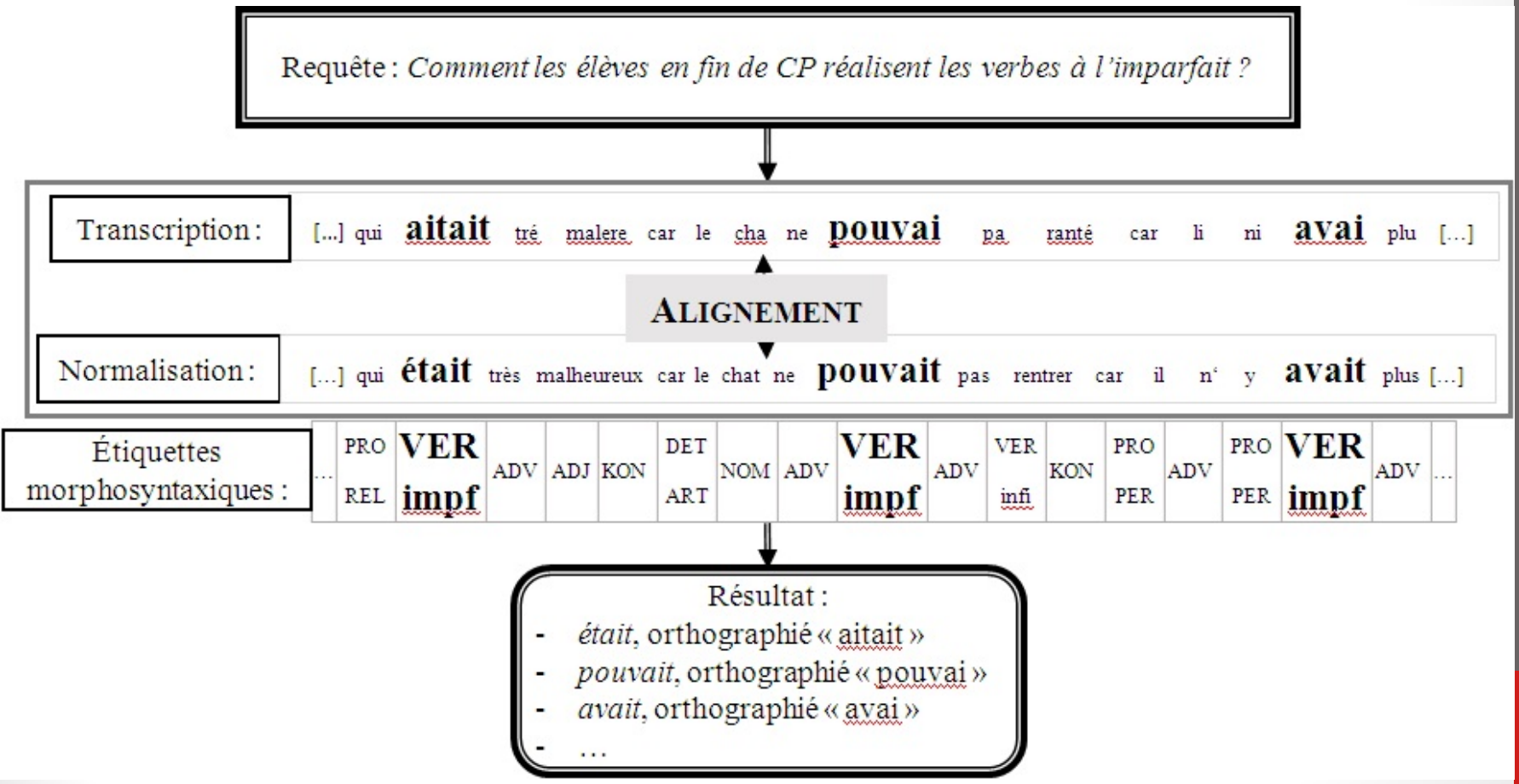
cé <revision/> stoir din cha qui aitait tré malere / car le cha ne pouvai pa ranté / car li ni avai plu de plas / donc le dé sida dannalé mai / bada Boum le cha se fa mal / sa maman se révaia <letMF>a</letMF>i la / ramena dan le nouvo ni bou quou / plgran.

- Normalisation

c'est l'histoire d'un chat qui était très malheureux car le chat ne pouvait pas rentré car il n'y avait plus de place donc il décida de s'en allé mais badaboum le chat se fait mal sa maman se réveilla et le ramena dans le nouveau nid beaucoup plus grand.

II – Principe d'alignement

Alignement : enjeux



Etat de l'art

- Dictées
  - Santiago-Oriola (1998)
  - Beaufort et Roekhaut (2011)
- Traduction
  - Desmet, Hérogue, 2005

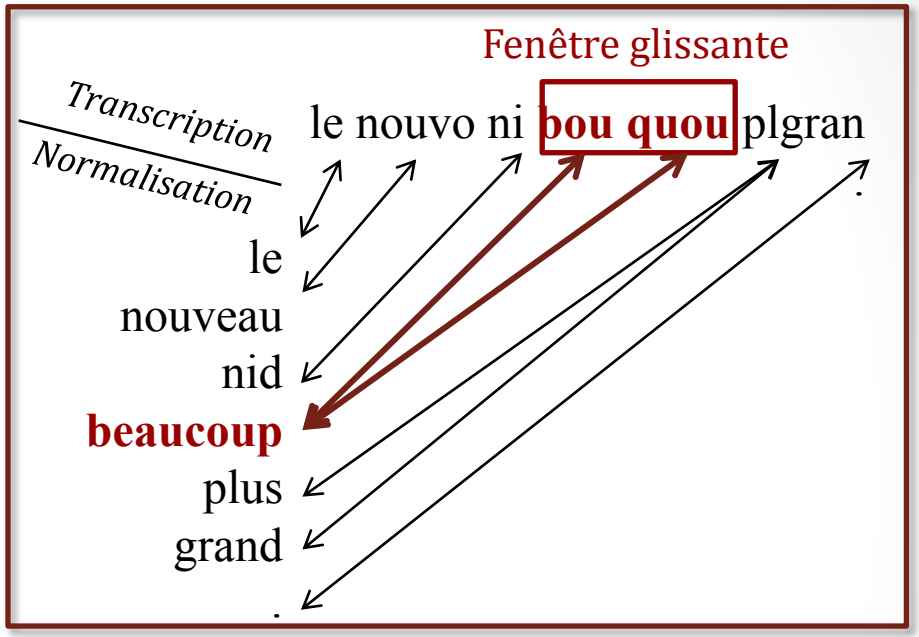
Principes d'alignement

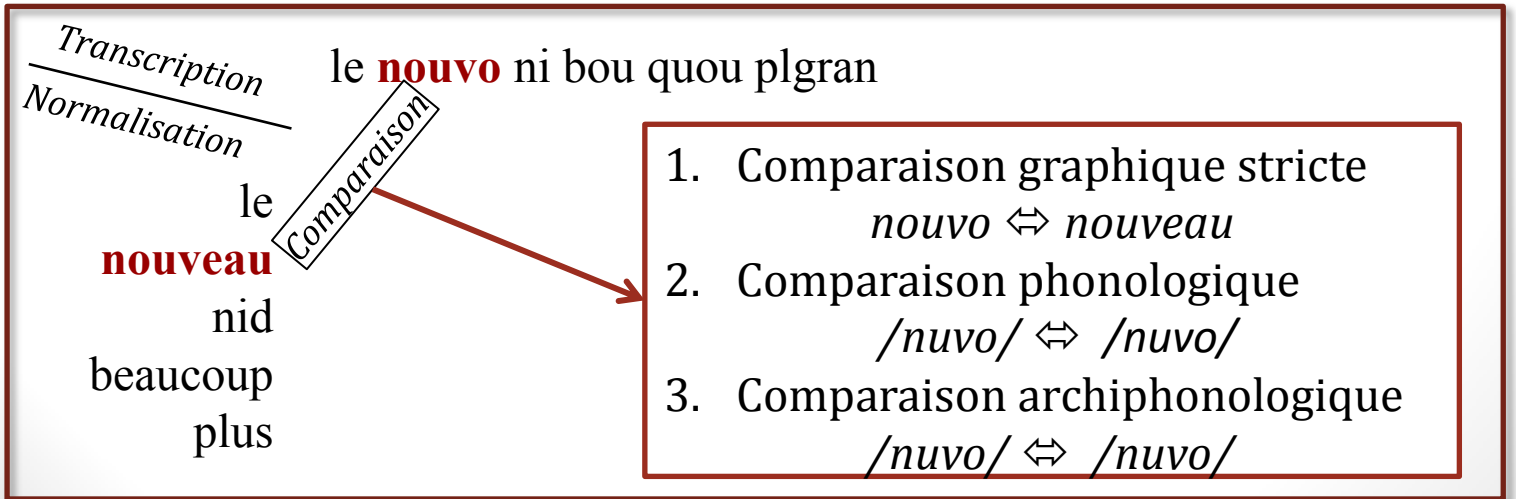
Unité de comparaison

- Alignement de tokens
- Fenêtre glissante

Modes de comparaison

- Indices phonologiques
- Indices phonologiques avec relâchement de contraintes





Exemple

Transcription		*Normalisation*
é	⟷	et
il		il
miole		miaule
sa		sa
maman		maman
li		lui
di		dit
plere		ne
pa		pleure
petit		pas
chaton		petit
et		chaton
elle		et
latrape		elle
.		l'
		attrape
		.

Exemple

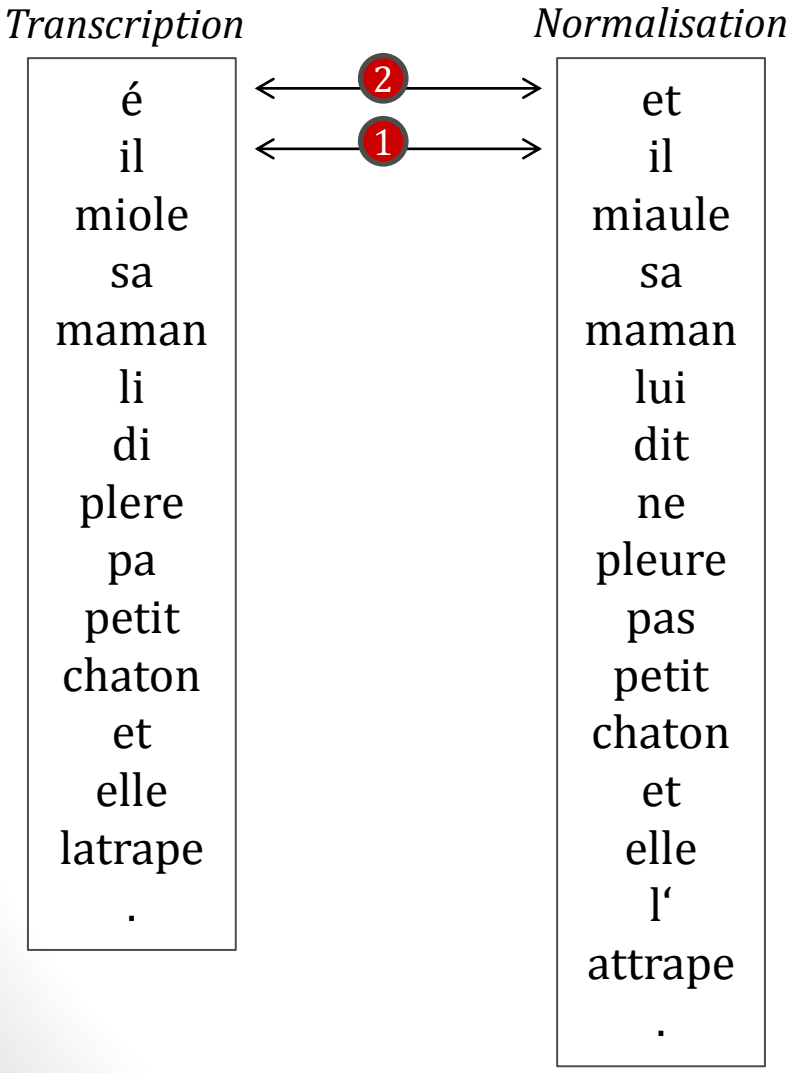
1 Comparaison graphique stricte

Exemple

1. Comparaison graphique stricte
2. Comparaison phonologique

é	et
/e/ ⟷	/e/

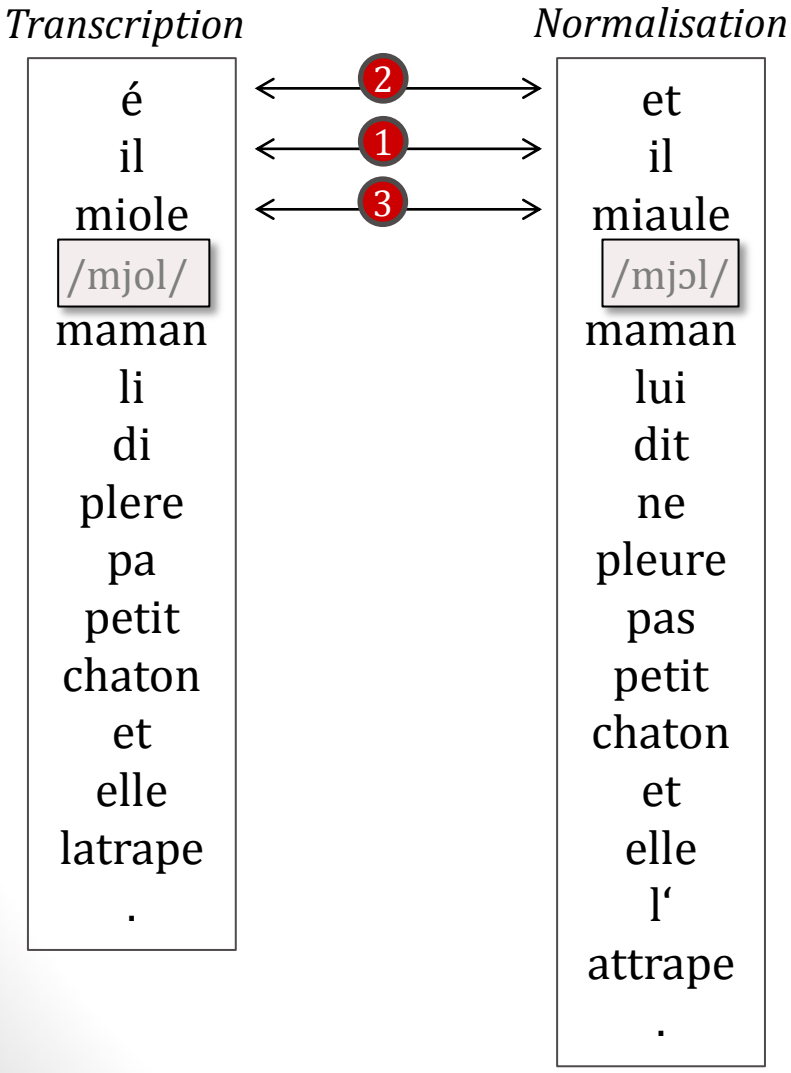
Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

é	et
/e/ ⟷	/e/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

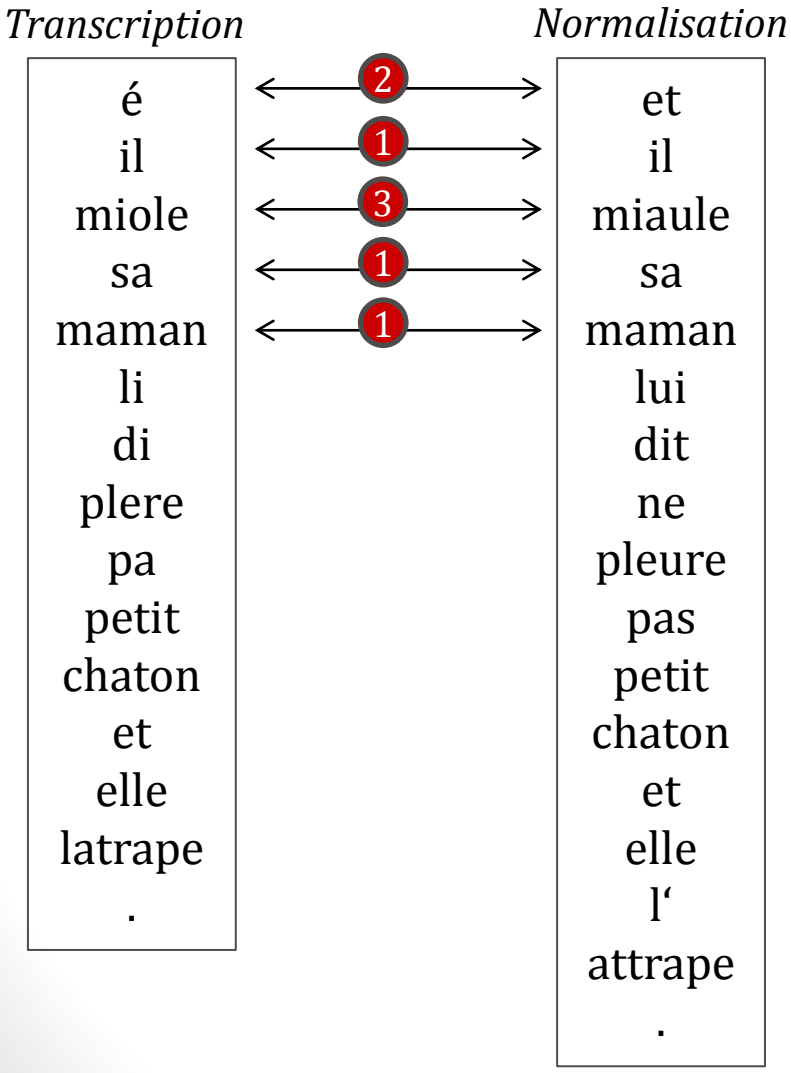
é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

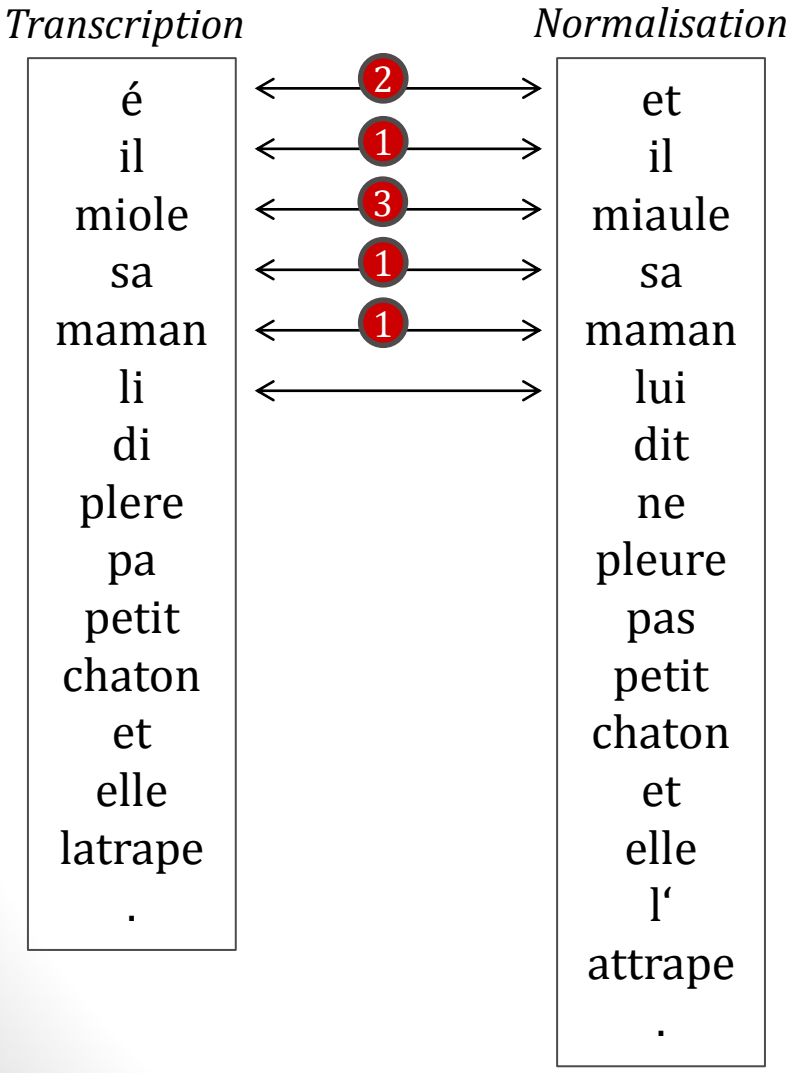
é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

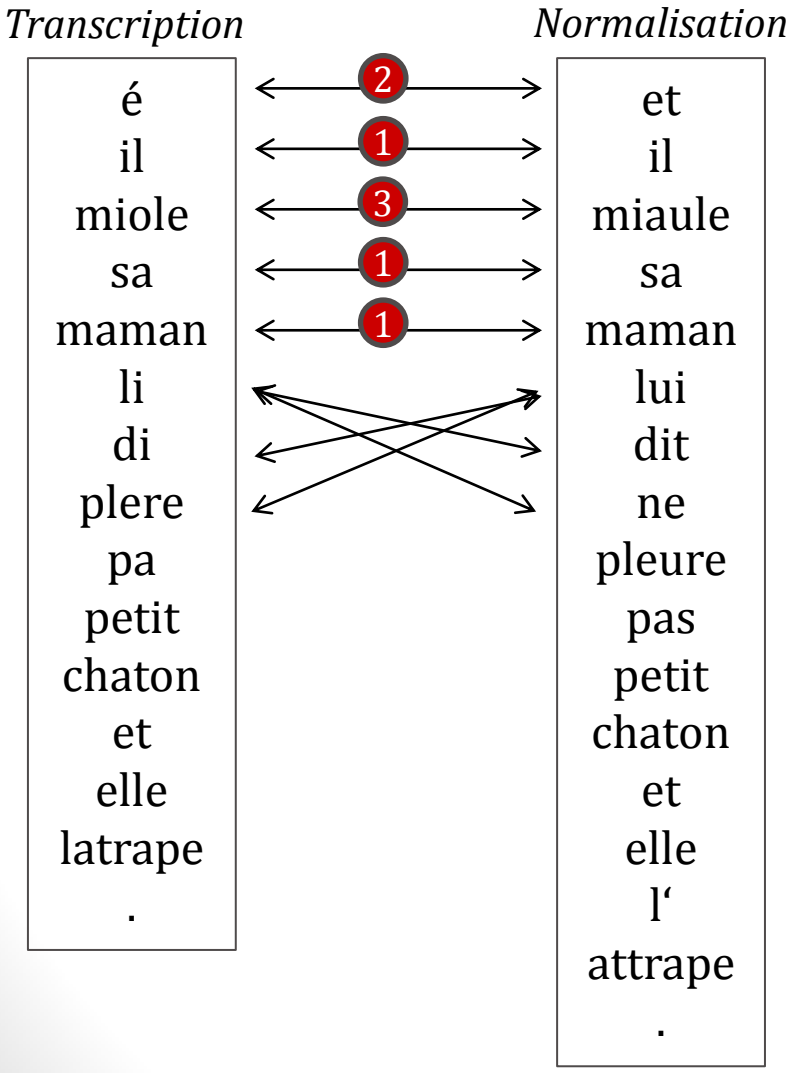
é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

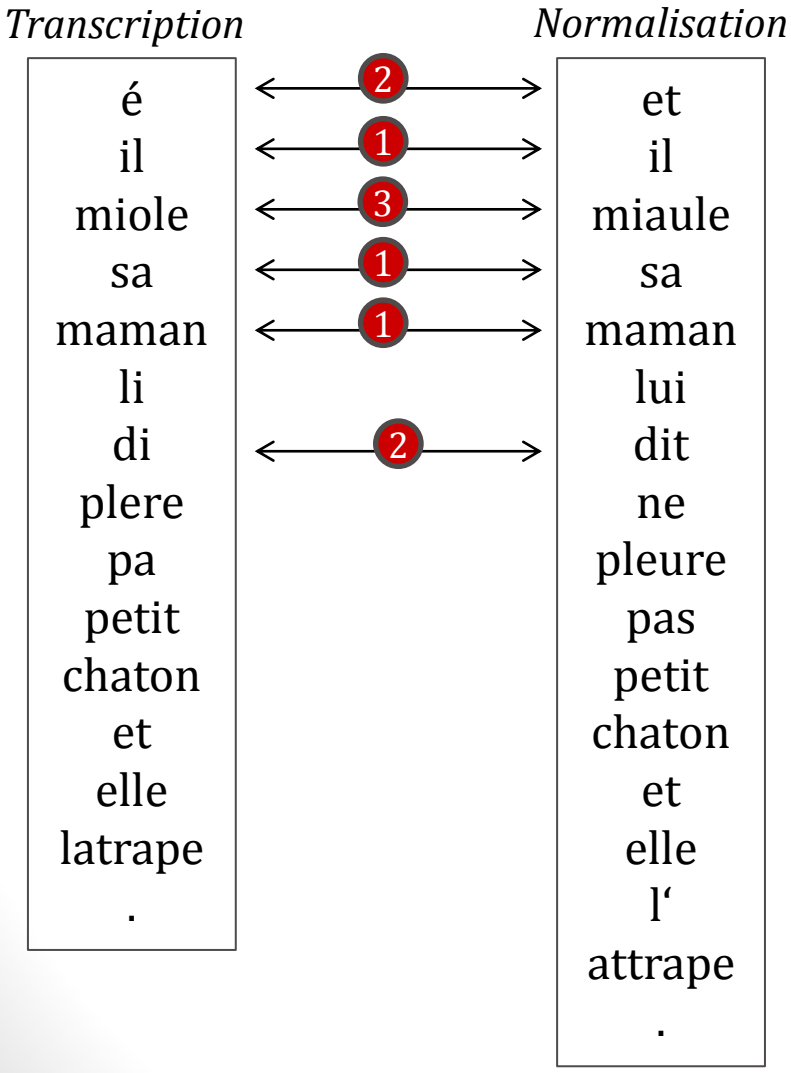
é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

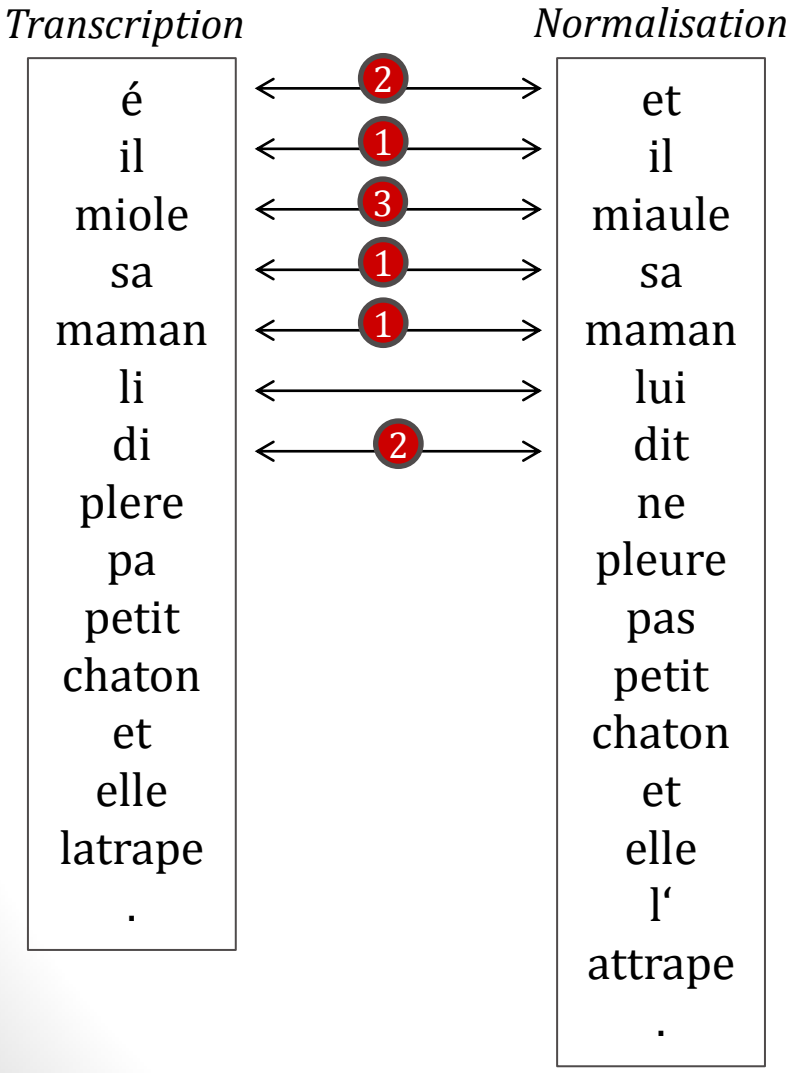
é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

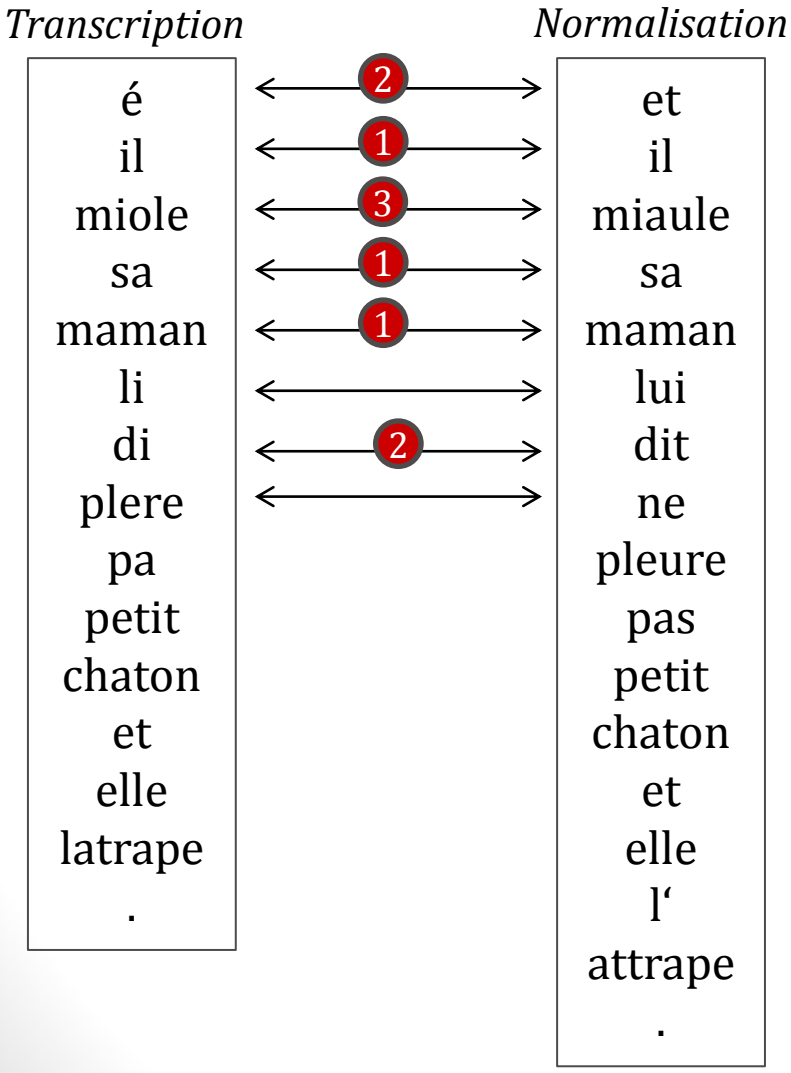
é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

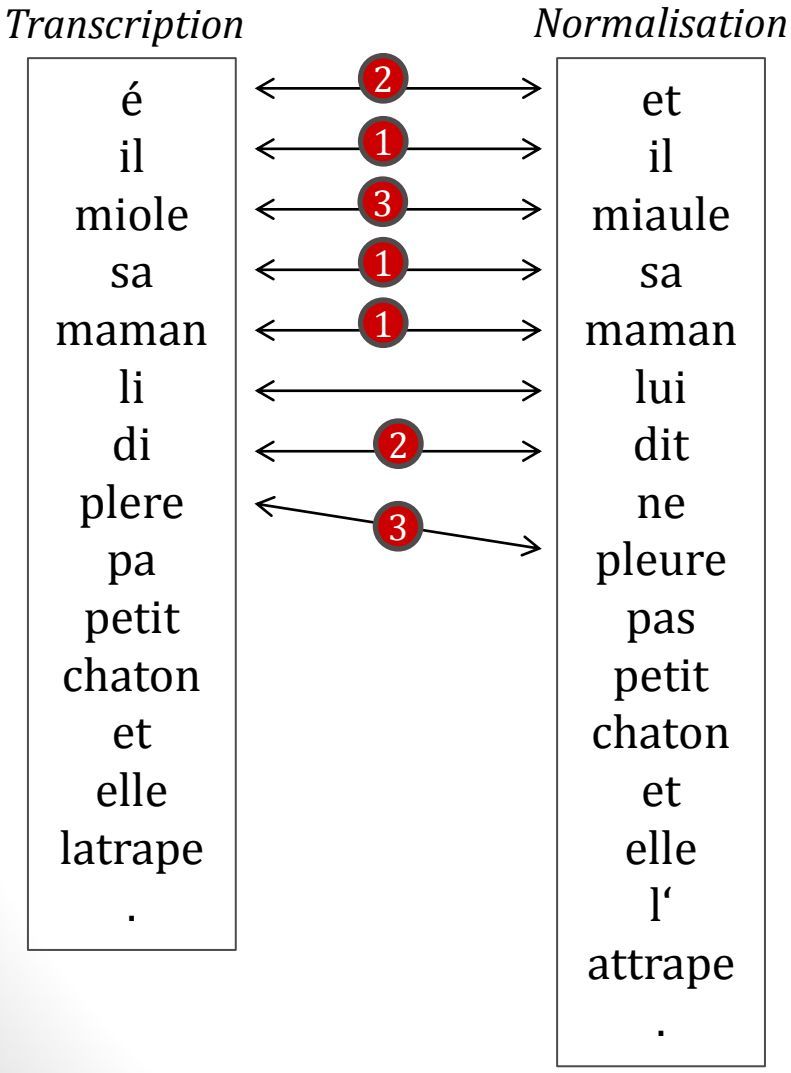
é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

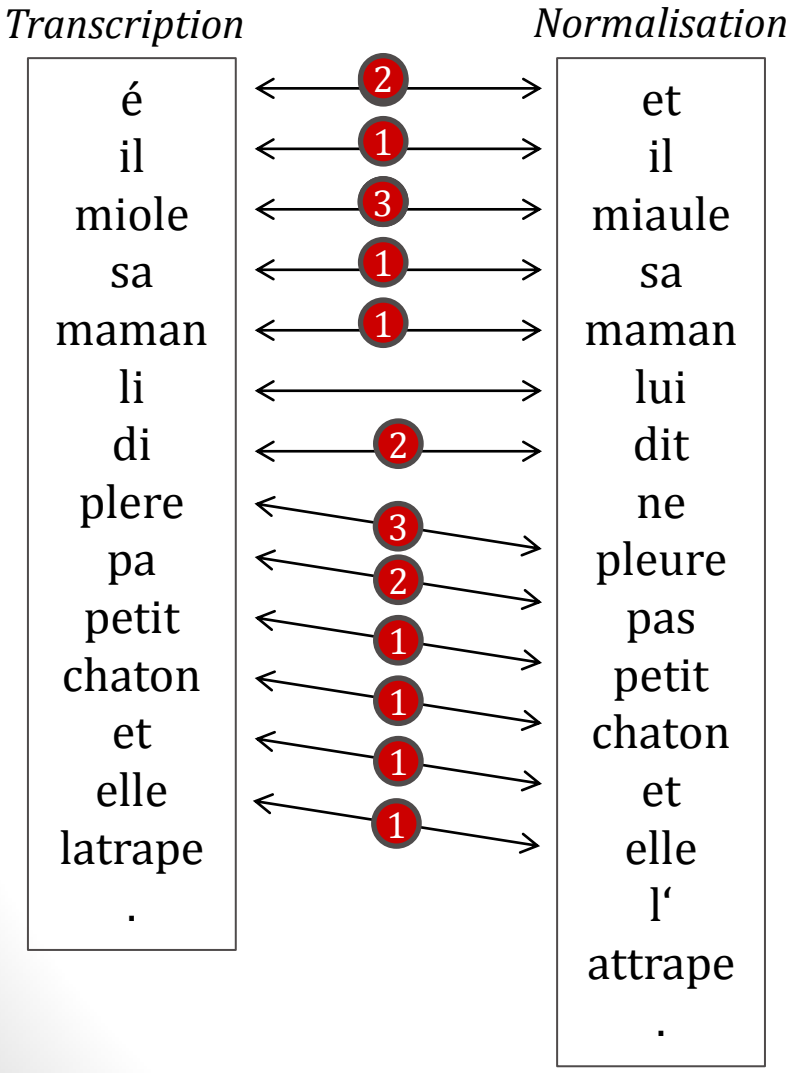
é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

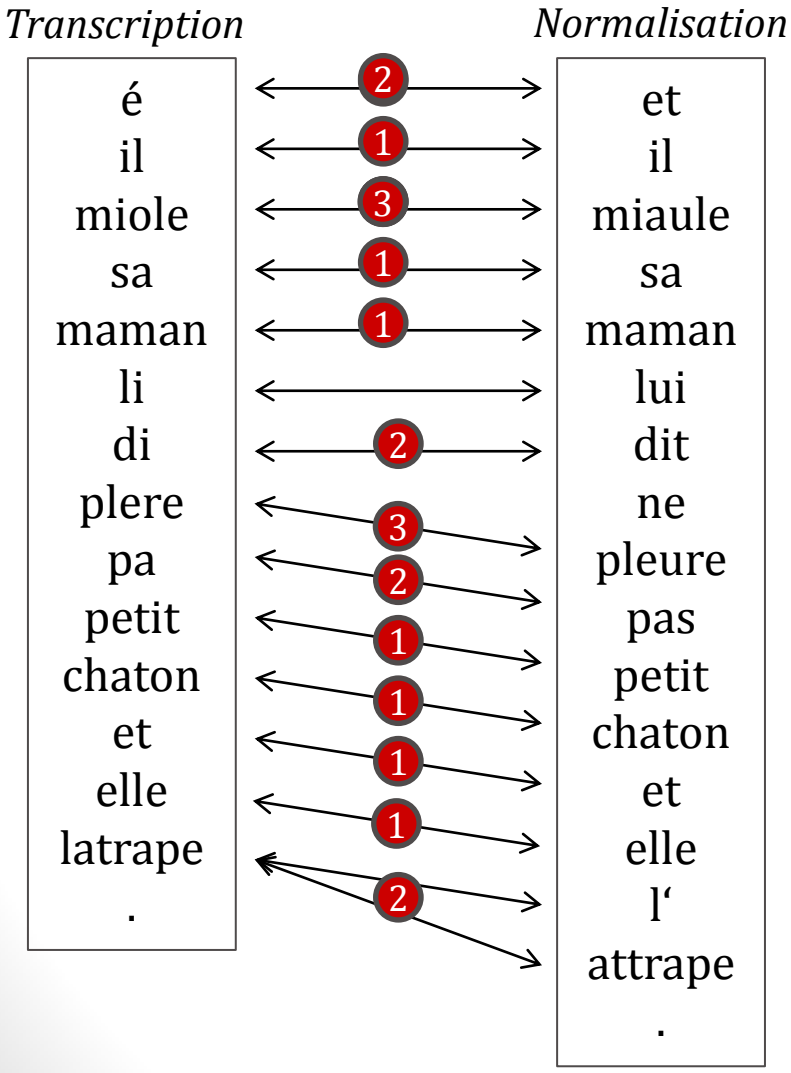
é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

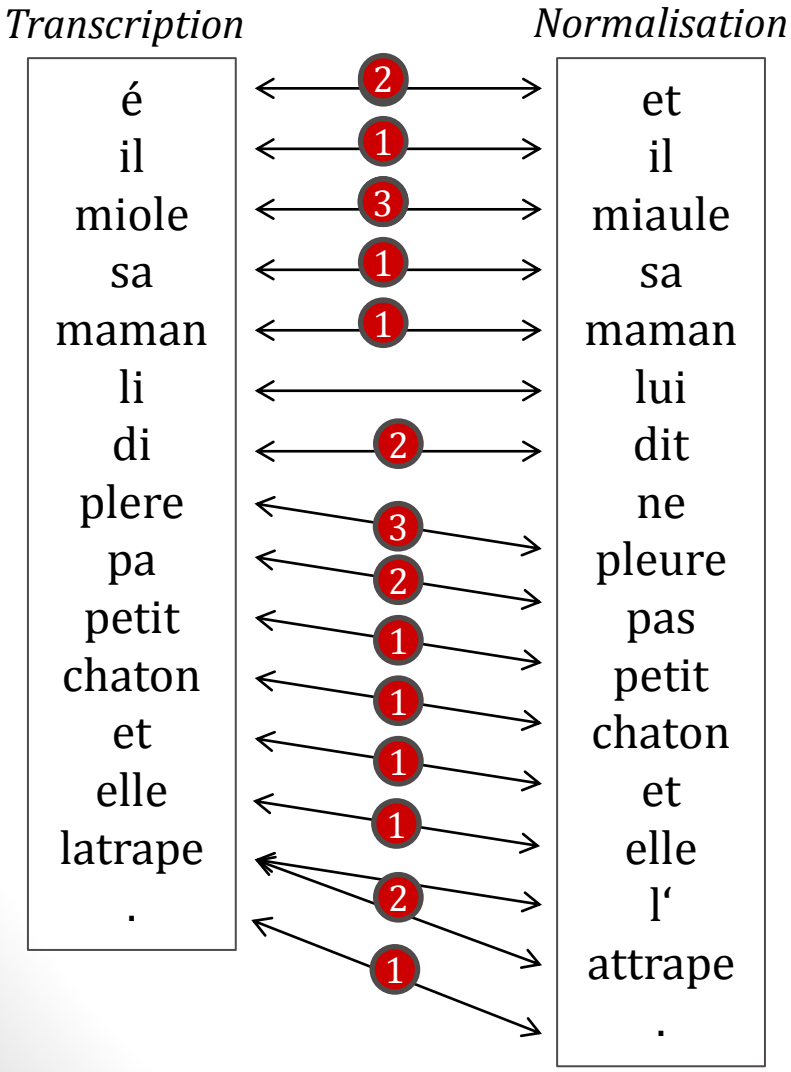
é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

Exemple



1. Comparaison graphique stricte
2. Comparaison phonologique

é	et
/e/ ⟷	/e/

3. Comparaison phonologique avec relâchement de contraintes

e	ə	œ̃	ɔ	Ø final
ɛ	ø	ɛ̃	o	schwa

miole	miaule
/mjol/	/mjɔl/
/mjol/ ⟷	/mjol/

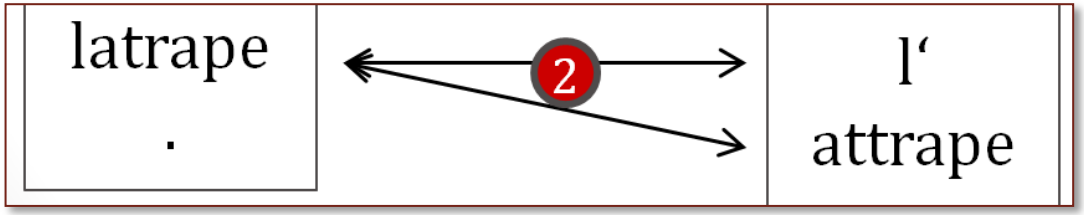
III – Résultats et perspectives

Résultats

Mesure	Alignement par comparaison graphique stricte	Alignement par comparaison phonologique	Alignement par comparaison phonologique avec relâchement de contraintes
Précision	83,3 %	86,3 %	**86,7 %**
Rappel	59,9 %	72,6 %	**74 %**
F-mesure	69,7 %	78,9 %	**80 %**

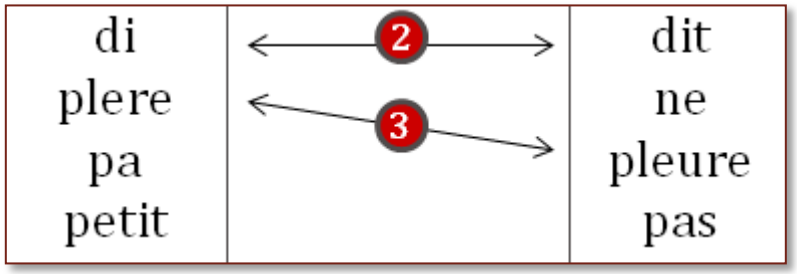
Résultats

Taille de la fenêtre	Aucune concaténation	Concaténation n.(n+1)	Concaténation n.(n+1).(n+2)	Concaténation n.(n+1).(n+2).(n+3)
Précision	87,82%	88,22%	**88,17%**	88,17%
Rappel	71,74%	75,48%	**75,60%**	75,60%
F-mesure	78,97%	81,35%	**81,41%**	81,41%



Résultats

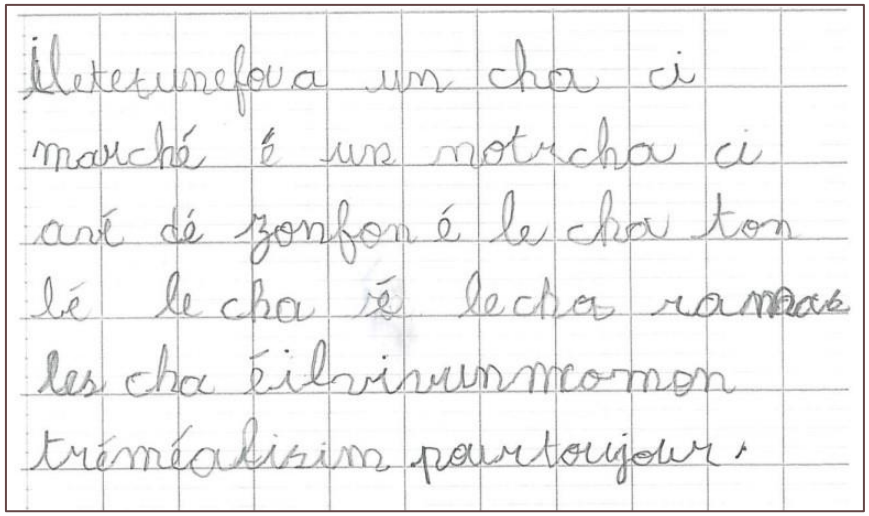
Distance	Comparaison à n+1	Comparaison à n+2	Comparaison à n+3
Précision	86,89%	**88,17%**	86,93%
Rappel	66,35%	**75,60%**	74,21%
F-mesure	75,24%	**81,41%**	80,07%



Formes non reconnues

de	de
dormire	dormir
a#lor	alors
le	le
petit	petit
des#den	descend
dans	dans
lescalie	l'#escalier
boum	boum
.	.
il#miel	Il#miaule#<dialogue>
miaou	miaou
	</dialogue>
la	la
maman	maman
chat	chat
se	se
revei#re#mei	réveille#remet
son	son
petit	petit
é	et
il	il
sen#dors	s'#endort
.	.

Formes non reconnues



le	le
cha	chat
sé#la#cha#rama	<incomprehensible/>
les	les
cha	chats
éilvivunmcomon#tréméalisim#pourtoujour#.	et#ils#vivent#en#commun#très#<incomprehensible/>
	pour
	toujours#.#

lien

UNIVERSITÉ Grenoble Alpes

Scoledit

Tableau de bord Accueil Visualisation Recherche Commentaires A propos

Login : Claire | Déconnexion

Visualisation Corpus CP

Afficher le CE1

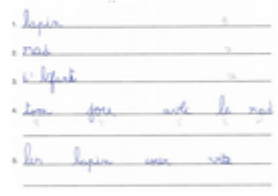
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Rechercher un élève : N° d'élève ok

Il y a actuellement à l'étude : 1400 élèves.

Elève 47

CP CE1

SCAN DICTÉE SEPTEMBRE	DICTÉE SEPTEMBRE	DICTÉE JUIN	SCAN DICTÉE JUIN
	LP RAPH LPIH LLULHÏH	lapin rad é léfant tom jou avec le rad loc lapin cour vite	

PRODUCTION	SCAN PRODUCTION
le chet et tonbe dent les zozelié et il pleur perse que il ses fe tre tre mele et se memen le pratg et il le loch.	

Elève 49

CP CE1

UNIVERSITÉ Grenoble Alpes

Scoledit

Tableau de bord Accueil Visualisation Recherche Commentaires A propos

Login : Claire | Déconnexion

Lexique 476 mots distincts toutes classes confondues.

Afficher les formes produites par les élèves

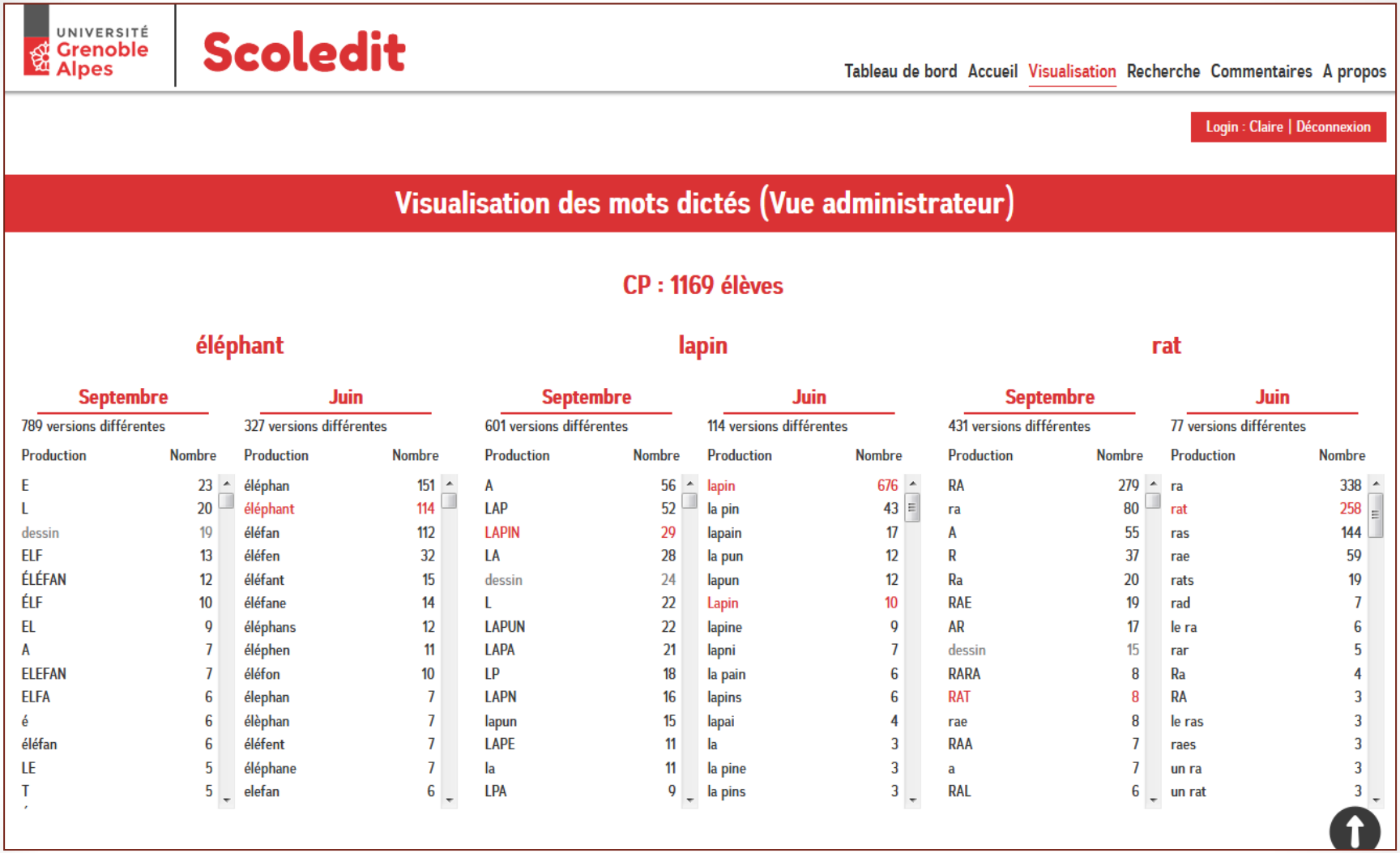
Afficher les formes rectifiées (normées)

Formes du dictionnaire (vocabulaire)

Vocabulaire CP

476 mots distincts sur 12686 répertoriés à ce jour (3,75%).

Mots	Nb d'occurrences	Mots	Nb d'occurrences
le	1565	chat	1052
il	892	et	822
petit	483	maman	482
son	454	un	438
être	398	se	393
tomber	393	pleurer	230
faire	174	avoir	172
miaou	156	réveiller	149
de	147	qui	139
marche	120	dormir	119
chaton	118	mal	116


UNIVERSITÉ Grenoble Alpes
Scoledit
Tableau de bord Accueil Visualisation Recherche Commentaires A propos
Login : Claire | Déconnexion
Visualisation des mots dictés (Vue administrateur)
CP : 1169 élèves
éléphant
Septembre
789 versions différentes
Production Nombre
E 23
L 20
dessin 19
ELF 13
ÉLÉFAN 12
ÉLF 10
EL 9
A 7
ELEFAN 7
ELFA 6
é 6
éléfan 6
LE 5
T 5
Juin
327 versions différentes
Production Nombre
éléphan 151
éléphant 114
éléfan 112
éléfen 32
éléfant 15
éléfane 14
éléphans 12
éléphen 11
éléfon 10
élephan 7
élèphan 7
éléfent 7
éléphane 7
elefan 6
lapin
Septembre
601 versions différentes
Production Nombre
A 56
LAP 52
LAPIN 29
LA 28
dessin 24
L 22
LAPUN 22
LAPA 21
LP 18
LAPN 16
lapun 15
LAPE 11
la 11
LPA 9
Juin
114 versions différentes
Production Nombre
lapin 676
la pin 43
lapain 17
la pun 12
lapun 12
Lapin 10
lapine 9
lapni 7
la pain 6
lapins 6
lapai 4
la 3
la pine 3
la pins 3
rat
Septembre
431 versions différentes
Production Nombre
RA 279
ra 80
A 55
R 37
Ra 20
RAE 19
AR 17
dessin 15
RARA 8
RAT 8
rae 8
RAA 7
a 7
RAL 6
Juin
77 versions différentes
Production Nombre
ra 338
rat 258
ras 144
rae 59
rats 19
rad 7
le ra 6
rar 5
Ra 4
RA 3
le ras 3
raes 3
un ra 3
un rat 3

Bibliographie

- Beaufort R., Roekhaut S. (2011). Automation of dictation exercises. *Computational Linguistics in the Netherlands Journal*, (1), 1-20.
- Béchet F. (2001). LIA_PHON - Un système complet de phonétisation de textes, *Traitement Automatique des Langues (T.A.L.)* 42(1), 47-67.
- Boré C., Elalouf M.-L. (2017). Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. In Doquet C., David J., Fleury S., *Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement*, Corpus, 16 | déc. 2016, 185-214.
- Catach N. (1980, 3e édition, 1995). L'orthographe française : traité théorique et pratique avec des travaux d'application et leurs corrigés (avec la collaboration de Gruaz C. et Duprez D.). Paris : Nathan.
- David J., Doquet C. (2016). Les écrits d'élèves: un corpus de référence pour le français contemporain. Actes de *SHS Web of Conferences* (Vol. 27, p. 11001). EDP Sciences.
- Desmet P., Héroguel A. (2005). Les enjeux de la création d'un environnement d'apprentissage électronique axé sur la compréhension orale à l'aide du système auteur IDIOMA-TIC. Actes d *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, 8(1), 281–303.
- Elalouf M.-L. (dir) (2005). *Écrire entre 10 et 14 ans. Un corpus, des analyses, des repères pour la formation*, SCérén, CRDP de Versailles.
- Garcia-Debanc C., Bonnemaison K. (2014). La gestion de la cohésion textuelle par des élèves de 11-12 ans: réussites et difficultés. Actes de *SHS Web of Conferences* (Vol. 8, pp. 961-976). EDP Sciences.
- Santiago Oriola C. (1998). Système vocal interactif pour l'apprentissage des langues. La synthèse de la parole au service de la dictée. Thèse. Toulouse III.
- Wolfarth C., Ponton C., Totereau C. (2017). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique. *Corpus 16*, janv. 2017, 185-214.

Merci de votre attention

LIDILEM

UNIVERSITÉ Grenoble Alpes

Aligner production et normalisation : une première approche pour l'étude d'écrits scolaires

Claire Wolfarth

Lidilem

Université Grenoble Alpes