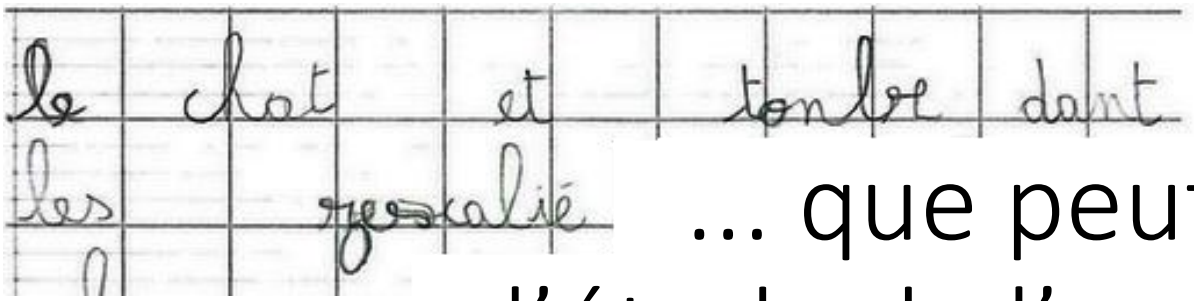


… que peut apporter le TAL à l'étude de l'apprentissage de l'écrit ? *L'approche par comparaison*

Claire Wolfarth et Claude Ponton

Laboratoire LIDILEM – UGA

Séminaire de master et séminaire DéLiCorTal - 16 oct. 2020

Data Institute Univ. Grenoble Alpes

LIDILEM Université Grenoble Alpes

UGA Université Grenoble Alpes

E-Calim Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques — Agence Nationale de la Recherche

Projets en cours

- Projet **Scoledit**
  - Constituer un **corpus scolaire longitudinal** du CP au CM2

- Projet E-Calm
  - Structuration d'un **vaste corpus d'élèves et d'étudiants (CP-Université)**

- Projet Scolinter
  - **Corpus scolaires longitudinaux** en français, italien et espagnol (grade 1 à 5)

Les corpus d'E-Calm

Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques

AGENCE NATIONALE DE LA RECHERCHE

http://e-calm.huma-num.fr/le-projet/

- Scoledit : http://scoledit.org/scoledition
- Ecriscol : http://syled.univ-paris3.fr/ecriscol/CORPUS-TEST/index.html
- Littéracie avancée : https://hdl.handle.net/11403/litteracieavancee/v1 (Ortolang)
- Resolco (CLLES – Toulouse)
- Corpus E-Calm : https://hdl.handle.net/11403/e-calm/v1 (Ortolang)

CLESTHIA EA 7345 | CIRCEFT | LIDILEM Université Grenoble Alpes | Clle

INTRODUCTION

Un format commun de transcription des écrits scolaires : XML-TEI (P5)

e-calm - > Corpus > **Grade_05_(CM2)**

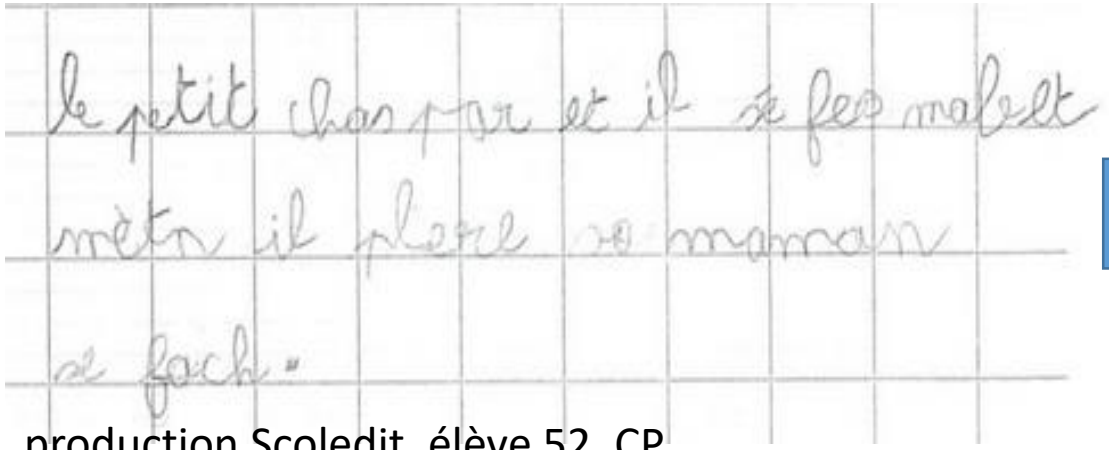
Nom

- Ecriscol
- Resolco
- Scoledit

```
 <settingDesc>
     <name type="region">AURA</name>
     <date>2018</date>
     <locale>école publique (MX) en zone (rurale|urbaine|mixte). (Moins de 100 élève
s|Entre 100 et 200 élèves|...|plus de 500 élèves) </locale>
     <activity>Rédaction</activity>
 </settingDesc>
</profileDesc>
</teiHeader>


<text>
        <body>
                <p><head>B4 le robot</head> Bonjour, je m'appelle B4 je suis un robot
<lb/> programés pour assurer votre sécurité, vous ne <lb/> devez pas trop savoir ou v
ous êtes avec <lb/> toutes ses ruines. Je vais faire un résumé <lb/> de ce qui s'est
passé. Nous sommes en l'an <lb/> 3025 à Berlin. Une guerre a éclaté entre <lb/> les A
mèricains et les nord coréens. Dès <lb/> centaines de bombes nuclaires on éclaté mais
<lb/> eur<mod resp="E" seq="T1"><gap/></mod>eusment pour nous des dizines d'abris on
<lb/> était crée. Ici vous résidès dans l'abri 111 <lb/><pb/> ou vous serez en sécuri
tés. Mai je suis ici <lb/> pour vous aidez... quoi. V<mod resp="E" seq="T1"><gap/></m
od>ous dite. À oui ça. <lb/> Effectivement je suis un peu rouiller ha ! Oups ! <lb/>
mon bras est tombées un soldat coréens ma tirez <lb/> dessu avec son fusil mais tout
va bien. oh </p><p> non c'est le cignal ! Vite ! suivez moi. Voaici <lb/> votre capsu
le les nord coréens nous attaques <lb/> vous serez criogénisée. Bonne chance dans <lb
/> le futur. Et voila comment cette histoire <lb/> ce termine. Pour B4 le gentil robo
t la fin <lb/> sera trajique il ne <mod resp="E" seq="T1"><gap/></mod>réussira pas à
```

Problématique



production Scoledit, élève 52, CP

le petit chas par et il sé fes male et mètn il plere sa maman se fach.

Transcription manuelle

Place, apports du TAL ?

Traitements automatiques

Description linguistique

Quelles sont les compétences des enfants en littéracie ?
Comment ces compétences évoluent ?

Interprétation manuelle

Propositions didactiques

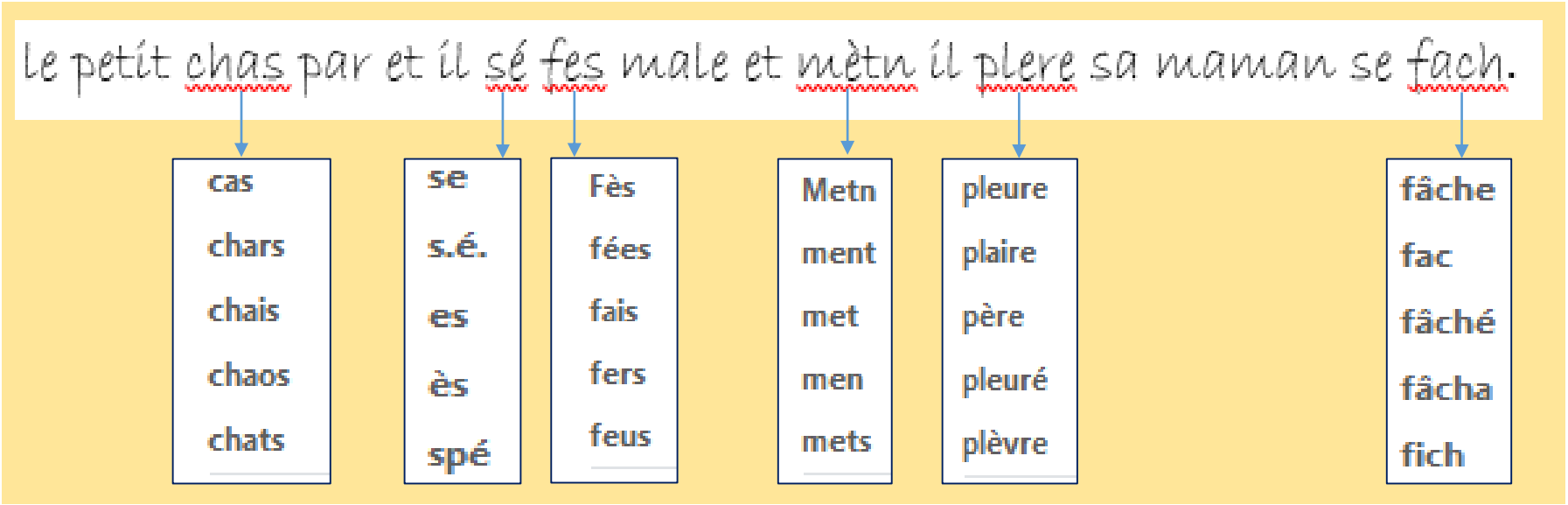
Qualité des données

MÉTHODOLOGIE

TAL et écrits scolaires…

TreeTagger

le	DET:ART	le	mal	ADV	mal
petit	ADJ	petit	et	KON	et
chas	NOM	chas	mètn	NOM	mètn
par	PRP	par	il	PRO:PER	il
et	KON	et	plere	VER:pres	plere
il	PRO:PER	il	sa	DET:POS	son
sé	VER:pper	sé	maman	NOM	maman
fes	VER:pres	fes	se	PRO:PER	se
			fach	NOM	fach



Correction (Word) :

Le petit cas par et il se Fès male et Metn il pleure sa maman se fâche.

Correction humaine :

Le petit chat part et il s'est fait mal et maintenant il pleure. Sa maman se fâche.
Le petit chat part et il se fait mal et maintenant il pleure. Sa maman se fâche.
Le petit chat part et il se fait mal et maintenant il pleure, sa maman se fâche.
...

MÉTHODOLOGIE

Approche par comparaison

- Prendre en compte les atouts et les faiblesses du TAL
  - approche moins-disante [Kraif & Ponton, 2007]
- S'appuyer dans un premier temps sur les traitements les mieux maitrisés du TAL : tokenisation, étiquetage morpho-syntaxique, lemmatisation...
- Se positionner dans une approche d'assistance à la description linguistique
- Proposer des comparaisons descriptives sur différents niveaux
  - graphique, phonologique, morpho-syntaxique...

- Nécessite
  1. un point de comparaison (normalisation)
  2. un alignement des unités

alignement tokens

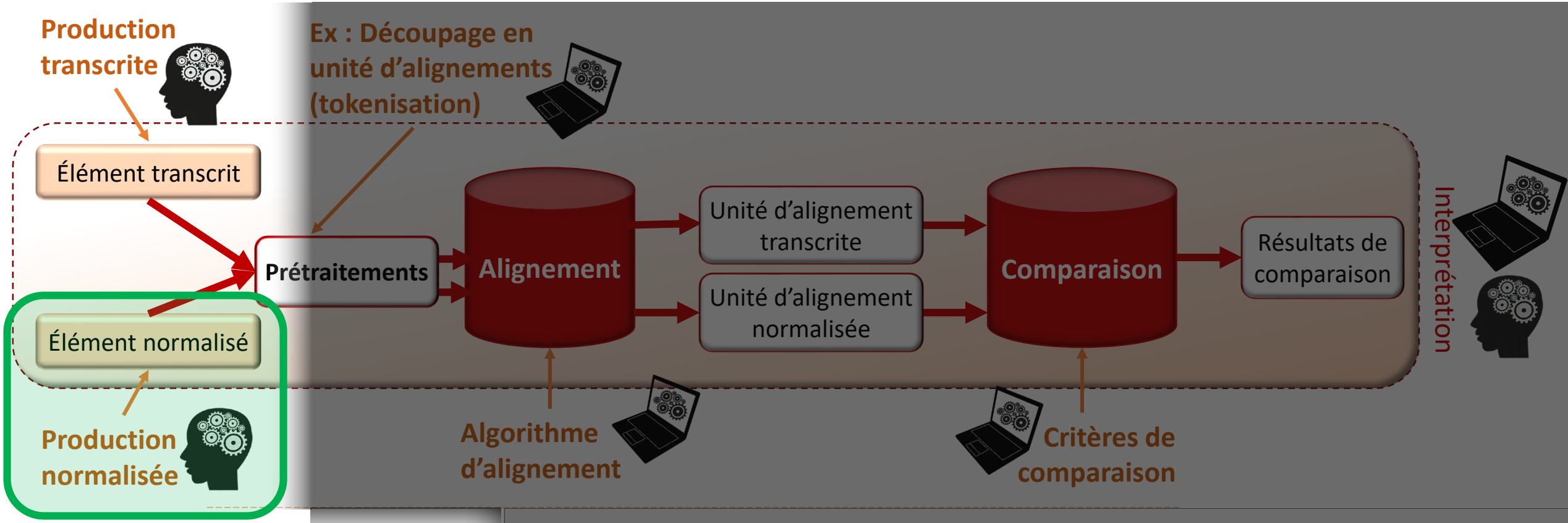
passait ***passai***

description

- différence graphique,
- même phonologie,
- même catégorie (ie. verbe),
- différence de temps verbal,
- personnes différentes,
- même base (*pass-*),
- désinences différentes,
- etc.

MÉTHODOLOGIE

Chaine de traitement : **la normalisation**



Transcription

…	se	**rencontere**	Il	parlais	…

…	se	**rencontrèrent**	Ils	parlaient	…

Normalisation

```
CM1 ...nologisation de la normalisation 53
CM1 :age morphosyntaxique de la production 53
        reading parameters ...
        tagging ...
         finished.
CM1 ...nt de la production 53
```

rencontere ↕ rencontrèrent

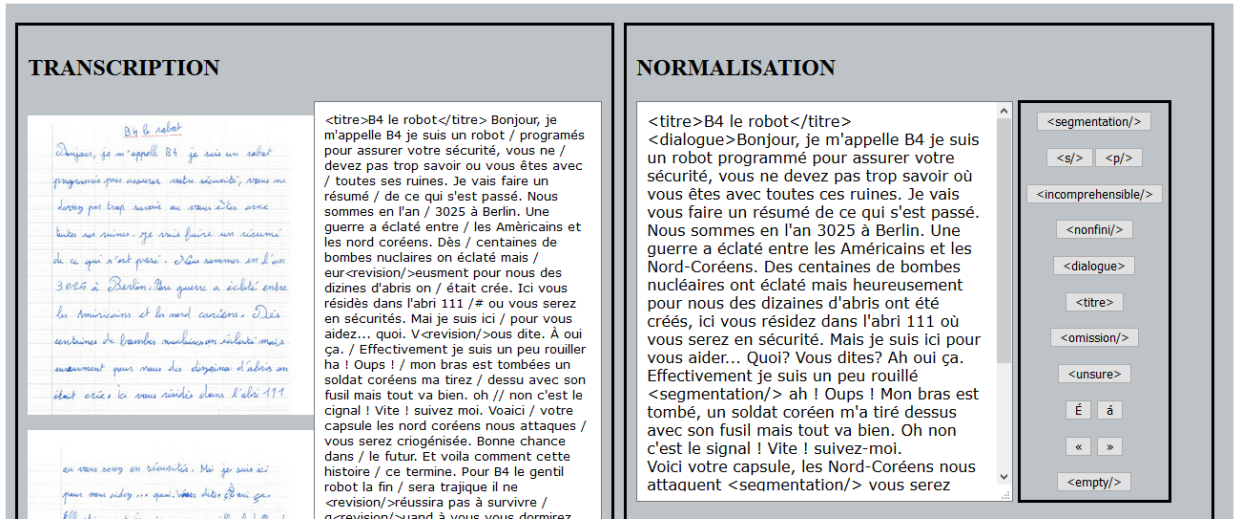
- Orthographe : *non normée*
- Phonologie : *non normée*

Normalisation

- Normaliser pour pouvoir utiliser des outils TAL
- Normaliser pour pouvoir comparer les productions à un « attendu »
- Rester au plus près de la production de l'enfant pour ne pas perdre d'informations
  - Ex. *le petit chas par et il sé fes male ...*
    1. Le petit chat part et il se fait mal
    2. Le petit chat part et il s'est fait mal
    - Choix solution 2 – ne pas perdre l'usage du passé
- Correction orthographe et segmentation des formes
  - hypo/hyper-segmentation
- Correction ponctuation si aucune ambiguïté
- Marquage des oublis de segmentation
- Indication en cas d'oubli de mots

Méthodologie de normalisation

- La normalisation est une étape d'annotation : enrichissement des données
  - Importance de la qualité des données
  - annotation déportée : une version transcrite + une version normalisée
    - faciliter le travail de l'annotateur (écriture d'un texte)
    - utilisation des correcteurs orthographiques (limiter les erreurs de frappe, d'inattention...)
  - guide de normalisation
  - opération manuelle mais outillée (limiter les erreurs de balisage)
- Révision manuelle par les chercheurs du projet de l'ensemble des normalisations
  - Travail long
  - Limiter les erreurs

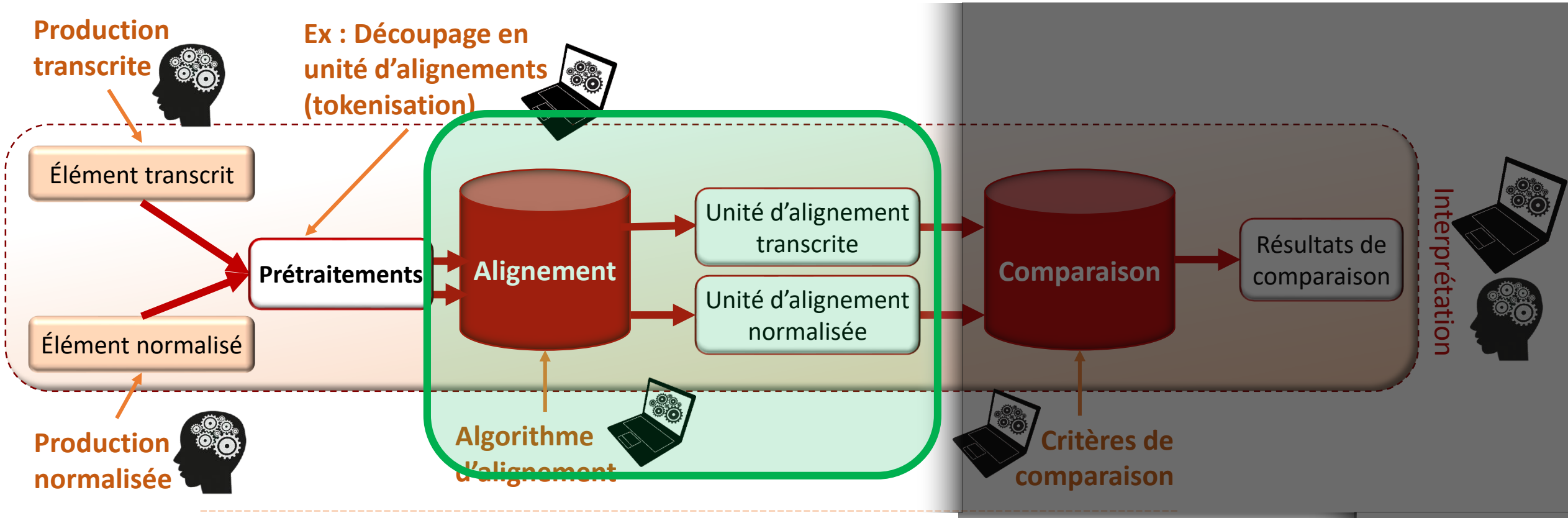


NORMALISATION

Statistiques textuelles

- L'existence d'une version normalisée des textes au format XML permet d'utiliser directement les outils de statistiques textuelles (TXM, Le Trameur...)
- Présentation d'une première étude lors des JLC 2019 (TXM)
  - Corpus Scoledit CP-CM1
  - Métadonnées par texte
    - niveau
    - genre (H/F)
    - type d'école (défavorisée/REP, mixte, favorisée)
- Quelques résultats/pistes
  - évolution constante de la longueur des textes produits
  - lexique mobilisé comparable à Manulex (sous-représentation des adjectifs)
  - les filles écrivent plus que les garçons...
  - forte spécificité du vocabulaire selon le genre (robot/loup/tuer *vs* sorcière/magique/chaton)
  - pas d'impact visible de la composition sociale des classes sur les variables observées
  - à compléter et à affiner avec d'autres niveaux, d'autres consignes, d'autres métadonnées (PCS parents, zone urbaine *vs* rurale par exemple)

Chaine de traitement : l'alignement

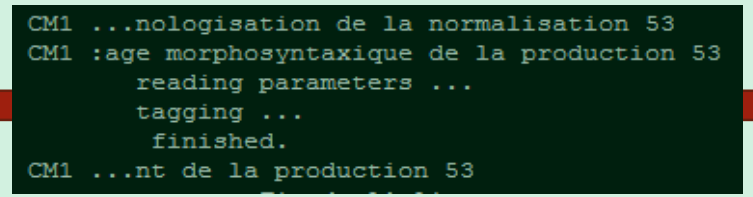


Transcription

… se **rencontere** Il parlais …

… se **rencontrèrent**. Ils parlaient …

Normalisation



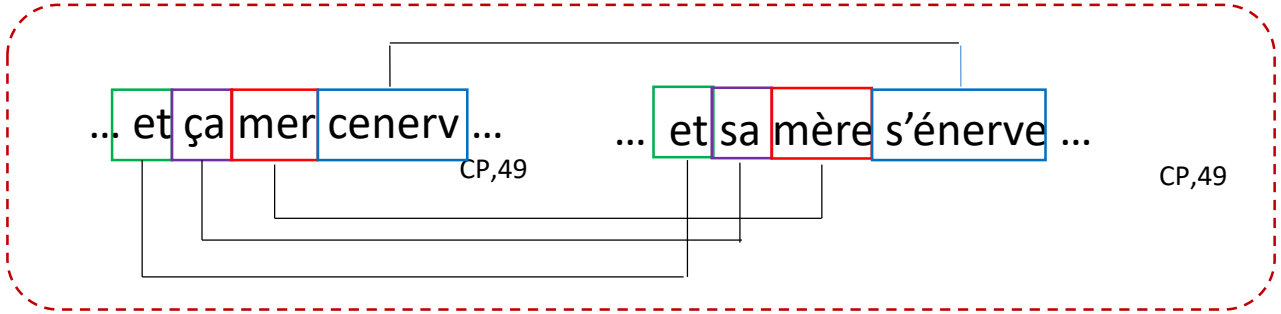
rencontere

↕

rencontrèrent

- Orthographe : *non normée*
- Phonologie : *non normée*

Alignement transcription-normalisation



Unité de comparaison

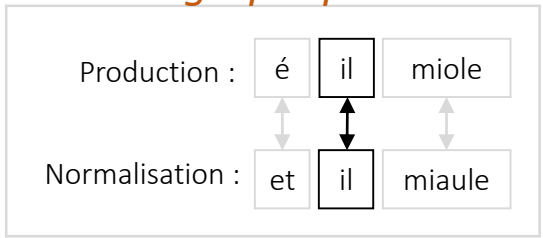
- Alignement de tokens/segments
- Fenêtre glissante

Modes de comparaison

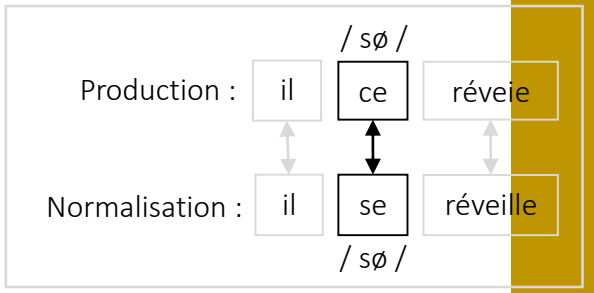
- Indices phonologiques
  - Utilisation de LIA_Phon (Béchet 2001)
- Indices phonologiques avec relâchement de contraintes
  - représentation archiphonologique (Catach 1980)
- Distances d'édition (Levenshtein, 1966)

Indices de comparaison

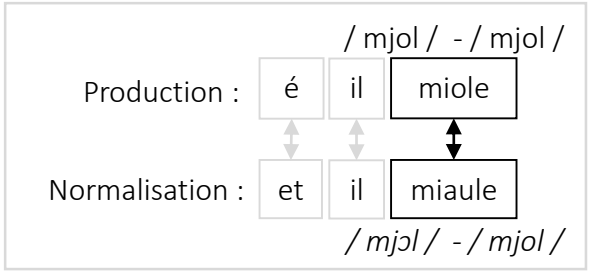
graphiques



phonologiques



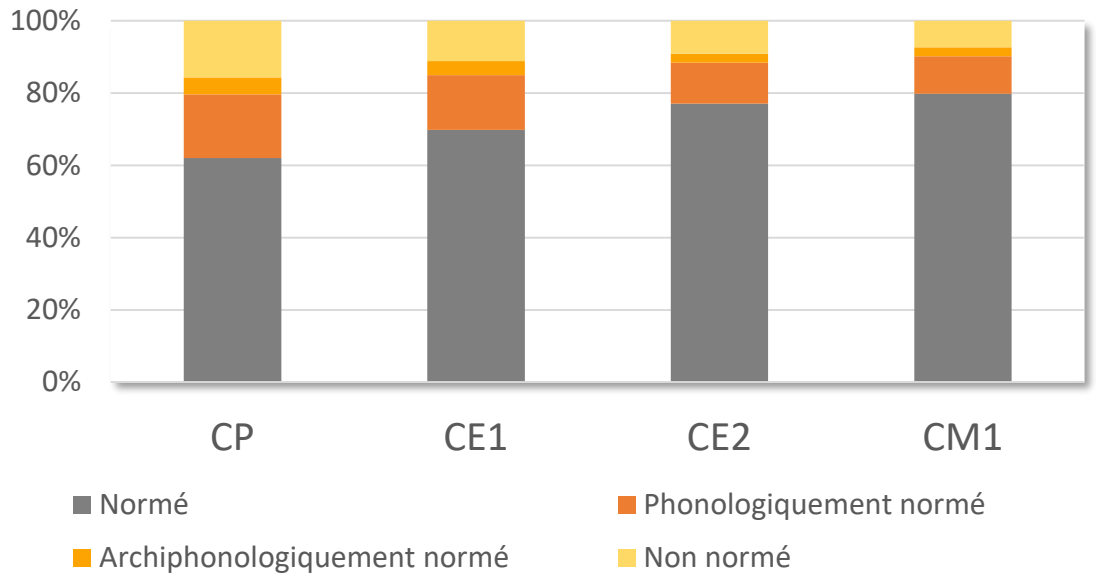
archiphonologiques

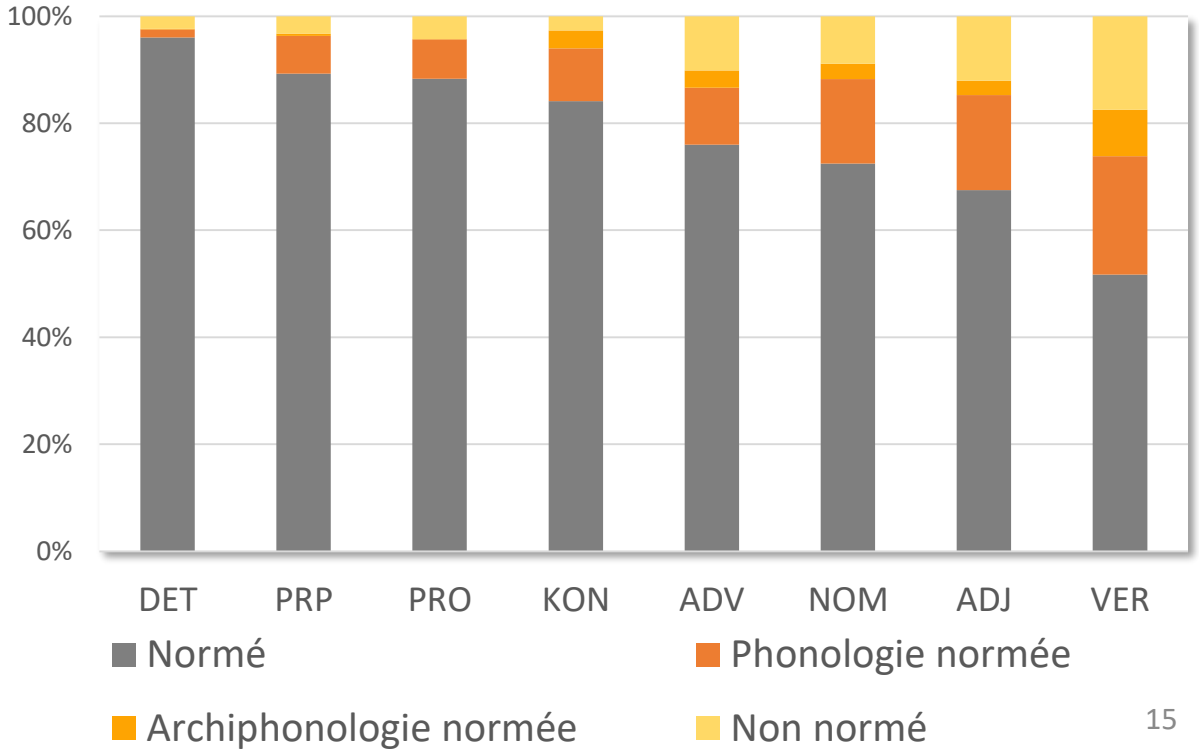


Exemple de sortie de l'aligneur *AliScol*

IdEleve	Niv	SegTrans	SegNorm	Lemme	Catégorie	StatutErreur	StatutSegm
815	CP	**<s>**	**<s>**	*<s>*	*[ZTRM->EXC*	01-Normé	01-Normé
815	CP	**un**	**Un**	*un*	*DET:ART*	01-Normé	01-Normé
815	CP	**petit**	**petit**	*petit*	*ADJ*	01-Normé	01-Normé
815	CP	**cha**	**chat**	*chat*	*NOM*	02-Phono	01-Normé
815	CP	**donbe**	**tombe**	*tomber*	*VER:pres*	04-ApproxGraphique	01-Normé
815	CP	**de**	**d'**	*de*	*PRP*	04-ApproxGraphique	01-Normé
815	CP	**une**	**une**	*un*	*DET:ART*	01-Normé	01-Normé
815	CP	**marque**	**marche**	*marche*	*NOM*	04-ApproxGraphique	01-Normé
815	CP	**é**	**et**	*et*	*KON*	02-Phono	01-Normé
815	CP	**cefaimale**	**se**	*se*	*PRO:PER*	02-Phono	03-HypoSeg
815	CP	**cefaimale**	**fait**	*faire*	*VER:pres*	02-Phono	03-HypoSeg
815	CP	**cefaimale**	**mal**	*mal*	*ADV*	02-Phono	03-HypoSeg
815	CP	**ai**	**et**	*et*	*KON*	03-Archi	01-Normé
815	CP	**plere**	**pleure**	*pleure*	*NOM*	04-ApproxGraphique	01-Normé
815	CP	**et**	**et**	*et*	*KON*	01-Normé	01-Normé
815	CP	**la**	**la**	*le*	*DET:ART*	01-Normé	01-Normé
815	CP	**maman**	**maman**	*maman*	*NOM*	01-Normé	01-Normé
815	CP	**cerai**	**se**	*se*	*PRO:PER*	07-Non normé	05-Non pertinent
815	CP	**véy**	**réveille**	*réveiller*	*VER:pres*	07-Non normé	05-Non pertinent
815	CP	**ai**	**et**	*et*	*KON*	03-Archi	01-Normé
815	CP	**la**	**la**	*le*	*DET:ART*	01-Normé	01-Normé

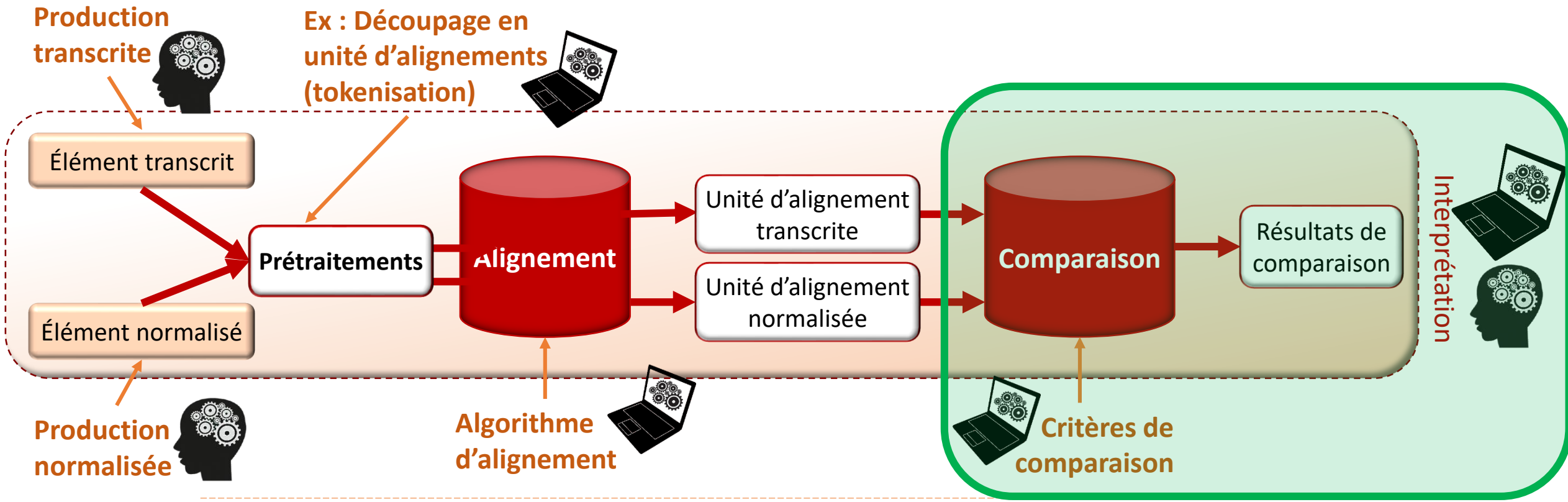
Alignement : *AliScol* - quelques résultats





Wolfarth, C., Ponton, C. & Brissaud, C. (2018). Gestion de la morphographie verbale en production d'écrits : Que peut nous apprendre un corpus longitudinal ? Repères. Recherches en didactique du français langue maternelle, 57, 209–226.

Chaine de traitement : **Analyses**

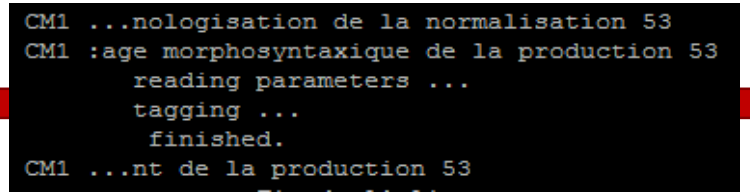


Transcription

…	se	**rencontere**	Il	parlais	…

…	se	**rencontrèrent**	Ils	parlaient	…

Normalisation



rencontere ↕ rencontrèrent

- Orthographe : *non normée*
- Phonologie : *non normée*

Description morphologie verbale

- Comparer les formes verbales produites aux formes verbales normées
- Dans un premier temps, non prise en compte
  - des temps composés (ex. *sé fes* vs *s'est fait*)
  - des cas d'hypo/hyper segmentation (ex. *sé* vs *s'est*)
- Objectifs
  - décrire les verbes utilisés, les tiroirs verbaux mobilisés
  - décrire les lieux d'erreurs : radical, désinence, les deux

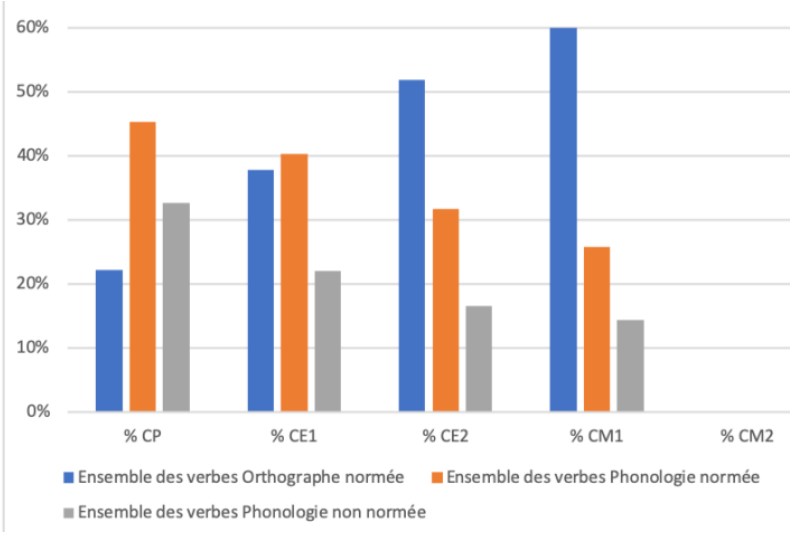
Les verbes dans le corpus

Tableau 2 : Répartition des mots du corpus (hors ponctuation) selon les classes grammaticales et le niveau de scolarité

POS	% CP	% CE1	% CE2	% CM1	% CM2	% Total
Nom	22,14%	20,97%	20,72%	20,00%	19,34%	20,63%
Verbe	19,25%	18,49%	17,94%	18,12%	18,18%	18,40%
Part. passé	2,59%	1,81%	2,37%	2,30%	2,36%	2,29%
Pronom	15,79%	15,31%	15,58%	15,82%	16,19%	15,74%
Déterminant	16,77%	16,13%	15,29%	14,25%	13,37%	15,16%

POS VERBES	% CP	% CE1	% CE2	% CM1	% CM2	Total
Conditionnel	0,06%	0,27%	0,35%	0,72%	0,61%	0,40%
Futur	0,06%	0,66%	1,12%	1,30%	1,12%	0,85%
Impératif	0,00%	0,00%	0,04%	0,00%	0,01%	0,01%
Imparfait	11,94%	26,70%	26,52%	25,07%	23,90%	22,83%
Infinitif	7,86%	15,00%	14,45%	15,58%	17,14%	14,01%
Participe présent	0,47%	0,41%	0,90%	1,04%	1,27%	0,82%
Présent	70,65%	42,33%	40,00%	34,91%	32,55%	44,09%
Passé simple	8,62%	14,10%	15,92%	20,70%	22,75%	16,42%
Subjonctif imparfait	0,17%	0,15%	0,07%	0,17%	0,12%	0,14%
Subjonctif présent	0,17%	0,39%	0,62%	0,51%	0,53%	0,44%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Erreurs selon le tiroir verbal



selon l'année d'apprentissage

		CP	CE1	CE2	CM1	CM2
Ensemble des verbes	Nb total de formes verbales (hors p. pas.)	1455	3691	6583	8254	
	Orthographe normée	25,57%	41,78%	55,28%	62,81%	
	Phonologie normée	45,70%	38,96%	30,08%	24,50%	
	Phonologie non normée	28,73%	19,26%	14,64%	12,70%	
Verbes à l'imparfait	Nombre de formes	173	1008	1786	2107	
	Orthographe normée	13,87%	29,86%	47,98%	57,71%	
	Phonologie normée	59,54%	50,69%	35,55%	29,90%	
	Phonologie non normée	26,59%	19,44%	16,46%	12,39%	
Verbes à l'infinitif	Nombre de formes	119	561	981	1318	
	Orthographe normée	16,81%	47,42%	63,00%	70,56%	
	Phonologie normée	36,13%	29,23%	24,67%	19,88%	
	Phonologie non normée	47,06%	23,35%	12,33%	9,56%	
Verbes au présent	Nombre de formes	1023	1531	2569	2816	
	Orthographe normée	28,64%	49,51%	62,98%	68,93%	
	Phonologie normée	45,65%	36,71%	27,40%	23,30%	
	Phonologie non normée	25,71%	13,78%	9,61%	7,78%	
Verbes au passé simple	Nombre de formes	129	539	1067	1738	
	Orthographe normée	27,13%	38,40%	45,83%	55,01%	
	Phonologie normée	39,53%	33,40%	30,65%	22,61%	
	Phonologie non normée	33,33%	28,20%	23,52%	22,38%	

Modéliser le découpage en base + désinence

- décidaient vs déssidaient (CE2, Scoledit, 2531)
  - décid + aient vs déssid + aient => erreur sur la base
- était vs étaient (CE2, Scoledit, 1845)
  - ét+ait vs ét+aient => erreur sur désinence
- perdit vs pèrdat (CE2, Scoledit, 2970)
  - perd + it vs pèrdat => erreur base et désinence
  - Découpage de pèrdat ?

Approche

- Quelques exemples
  - désuinfécter/désinfecter, régnala/régala, trencforma/transforma...
  - disa/dit, changé/changer, meta/mit....
  - curus/courut, abiter/habitaient, lansair/lancèrent....

1. Recherche de la désinence attendue
2. Recherche de la base attendue
3. Pas de découpage : erreur base ET désinence

Sorties d'AliDés

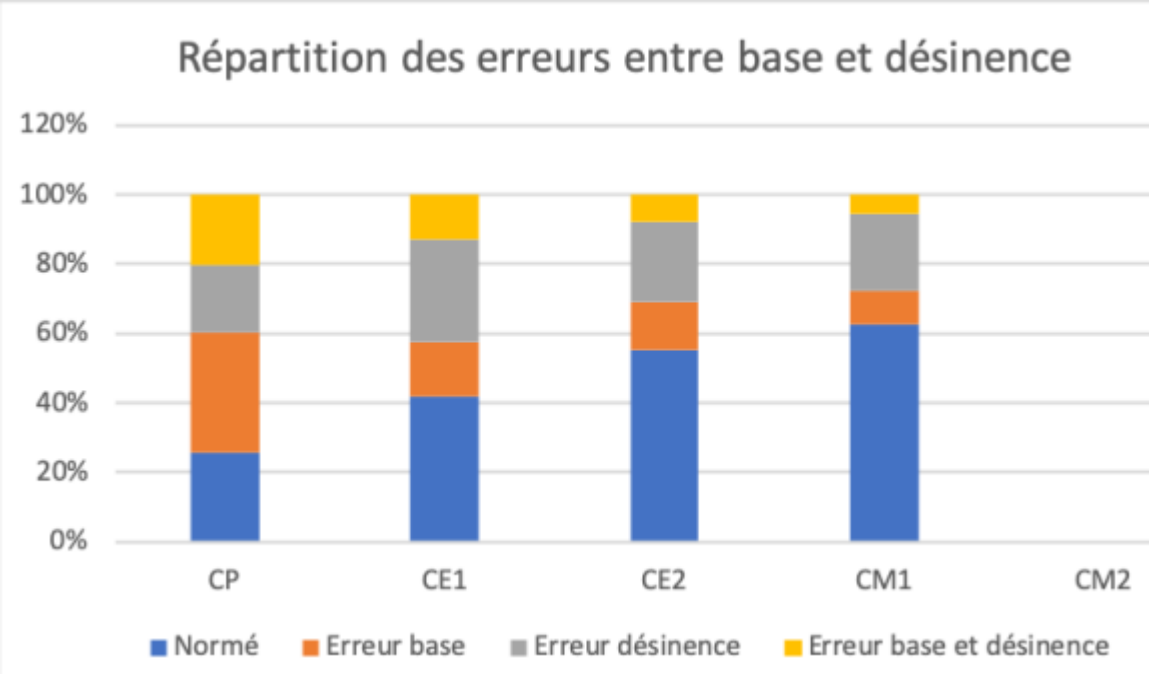
eleve	niv	lemme	cat	pers	statut	forme	baseForm	desiForm	prod	baseProd	desiProd	Err B seule	Err D seul	Err B ET D
1959	CE2	sortir	pres	P3	01-Normé	sort	sor	t	sort	sor	t	0	0	0
1364	CE2	être	pres	P3	01-Normé	est	est		est	est		0	0	0
1830	CE2	dire	pres	P3	07-Non norm	dit	d	it	disa	d	isa	0	1	0
2531	CE2	décider	impf	P6	02-Phono	décidaient	décid	aient	dessidaient	dessid	aient	1	0	0
576	CE2	être	impf	P6	03-Archi	étaient	ét	aient	été	été	µ	0	0	1
2446	CE2	aller	pres	P3	01-Normé	va	va		va	va		0	0	0
3026	CE2	chercher	impf	P3	01-Normé	cherchait	cherch	ait	cherchait	cherch	ait	0	0	0
949	CE2	emmener	pres	P1	02-Phono	emmène	emmèn	e	anmène	anmèn	e	1	0	0
2972	CE2	dire	pres	P3	02-Phono	dit	d	it	di	d	i	0	1	0
945	CE2	être	impf	P3	01-Normé	était	ét	ait	était	ét	ait	0	0	0
1665	CE2	cuire	infi	#	02-Phono	cuire	cui	re	cuir	cui	r	0	1	0
2906	CE2	dire	pres	P3	01-Normé	dit	d	it	dit	d	it	0	0	0
1665	CE2	avoir	impf	P3	02-Phono	avait	av	ait	avai	av	ai	0	1	0
1986	CE2	passer	impf	P3	02-Phono	passait	pass	ait	passai	pass	ai	0	1	0
1525	CE2	être	infi	#	01-Normé	être	êt	re	être	êt	re	0	0	0

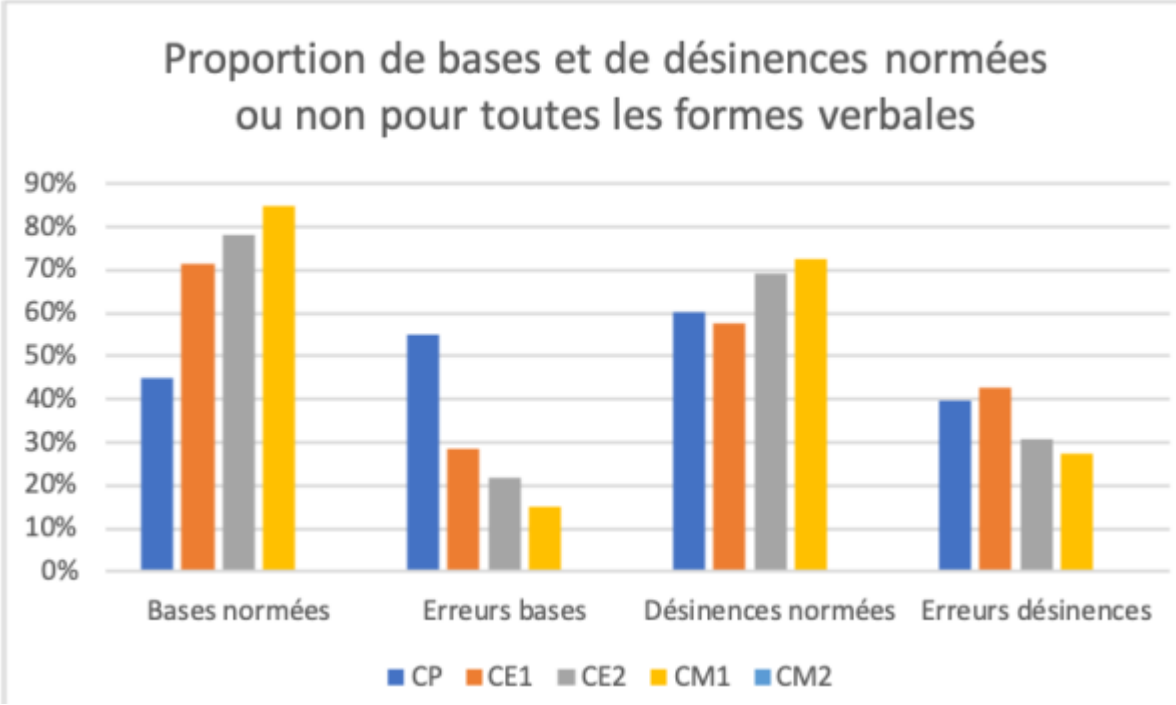
[Gaubil, 2020]

découpage base+désinence forme normée et forme produite

MORPHOLOGIE VERBALE

Quelques résultats



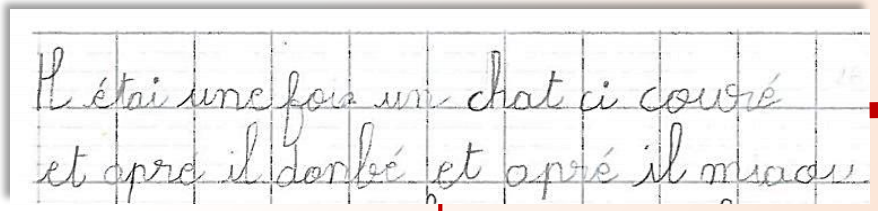


Traitement des adjectifs

- Module AliAdj [Gaubil 2020]
- Même approche méthodologique que pour les verbes
  - découpage : base + flexion genre + flexion nombre
- Objectif
  - comparer les formes adjectivales normées et produites au niveau
    - de la base
    - des flexions de genre et des flexions de nombre

Bilan actuel

Des corpus



Il étai une fois un chat ci couré *et apré il donbé et apré il miaou* […]

Des outils TAL

Des analyses linguistiques

première année	
était	
était	23,8%
été	14,3%
ete	14,3%
étai	14,3%
est	9,5%
aitai	4,8%
et	4,8%
et et	4,8%
est tai	4,8%
étés	4,8%

deuxième année	
était	
été	42,9%
était	38,1%
etait	4,8%
etais	4,8%
aitais	4,8%
êtés	4,8%

troisième année	
était	
était	71,8%
etait	7,6%
été	4,1%
étai	3,5%
étais	2,4%
ete	1,2%
avés	1,2%
étaits	0,6%
tes	0,6%
…	…

Des recommandations didactiques …

- Des **corpus** numérisés, annotés et disponibles
- Des **guides** de transcription et de normalisation (publiés dans E-CALM)

- Des **outils** et une méthodologie disponibles et réutilisables

- Des **analyses linguistiques** publiées ou en cours
  - Morphographie verbale
  - Segmentation en mots
  - Caractérisation des erreurs
  - Ponctuation

Scolinter

Université Grenoble Alpes

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

http://scoledit.org/scolinter/

- Reprise de la méthodologie Scoledit pour
  - constituer deux corpus d'écrits scolaires : italien (Milan) et espagnol (Alméria)
  - Corpus longitudinaux : CP (grade 1) au CM2 (grade 5)
  - mêmes consignes que Scoledit
  - mêmes guides de transcription et de normalisation
- Adaptation des modules AliScol, AliDés et AliAdj en cours
- Travail en cours sur la segmentation au CP en France, Italie et Espagne

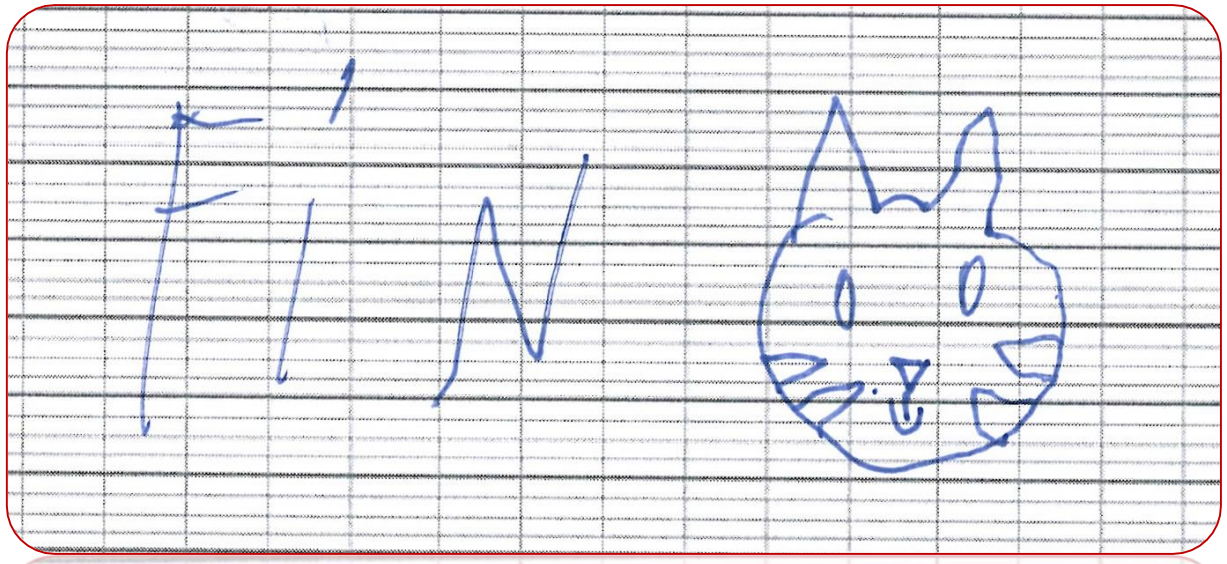
Langue	Éleve	Classe	Niveau	Transcription	Normalisation
	48	6307	1	le gro cha et drmi avec ses petis chatans et lotr cha il se promen. et le cha et tonbe sur le tapi. le cha pler et lotr cha et reveie	Le gros chat est endormi avec ses petits chatons et l'autre chat il se promène. Et le chat est tombé sur le tapis. Le chat pleure et l'autre chat est réveillé.
	1	0	1	IL GATINO CAMINA E CADE POI PIANGE POI LA MAMMA GATA LO PRENDE	Il gattino cammina e cade poi piange poi la mamma gatta lo prende.
	71	0	1	El gatito estába caminando y se callo pobre gatito lo rescato su madre	El gatito estaba caminando y se cayó [,] pobre gatito [•] lo rescató su madre.

Perspectives

- Extension des corpus
  - CM2 en cours
  - Collège : 6eme, 3eme (corpus Ecriscol et Resolco)
  - Lycée : 2ndes (Ecriscol)
  - Université : L et M (Ecriscol et Littéracie avancée)
- Affiner le traitement des verbes
  - perd+it *vs* pèrdat => pèrd + at
    - erreur radical et désinence mais
    - ~3ème pers. passé simple : temps et personne ok
- Finaliser et évaluer les modules
- Proposer une interrogation en ligne des résultats AliScol, AliDés et AliAdj
- Descriptions linguistiques et préconisations didactiques
  - Cohérence des temps
  - Traitement des lettres dérivatives
  - Ponctuation

Références

- Wolfarth, Claire, Ponton, Claude, Brissaud, Catherine. 2016. Du TAL dans les écrits scolaires : premières approches. Dans I. Smilauer & J. Kostov (éd.), Actes de la conférence conjointe *JEP-TALN-RECITAL 2016.* Vol. 9 : ELTAL. 30-36.
- Wolfarth, Claire, Ponton, Claude, Totereau, Corinne. 2017. Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire. *Corpus* 16: 185-214.
- Wolfarth, Claire, Brissaud, Catherine, Ponton, Claude. 2018. Transcrire et normer un corpus scolaire: pour quelles analyses ? *Repenser l'écriture et son évaluation au primaire et au secondaire* [Diptyque 36], *Brissaud, Catherine, Dreyfus, Martine & Kervyn, Bernadette* (eds), 121–146. Namur: PUN.
- Claire Wolfarth, Claude Ponton, Catherine Brissaud. 2018. *Which Method to Develop a Natural Language processing Tool to automatically analyze First Language Learner Corpora?*. In 3th Teaching and Language Corpora Conference (TALC2018) 18-21 July, 2018, Faculty of Education, University of Cambridge
- Wolfarth, Claire, Ponton, Claude, Brissaud, Catherine. 2018. Gestion de la morphographie verbale en production d'écrits: que peut nous apprendre un corpus longitudinal ? *Repères. Recherches en didactique du français langue maternelle* 57: 209–226.
- Ponton Claude, Wolfarth Claire, Brissaud Catherine. (2019. Premières explorations textométriques d'un corpus scolaire longitudinal (CP-CM1). JLC2019. 26-28 novembre 2019. Grenoble
- Brissaud C., Wolfarth C., Ponton C. 2019. *Analyse d'un grand corpus d'écrits d'élèves (CP-CM1) : des contraintes linguistiques à la didactisation de l'orthographe*. Colloque AIRDF. 27-29 août 2019. Lyon
- Wolfarth C., 2019. Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal (Doctoral dissertation, Université Grenoble Alpes).
- Ho-Dac M., Fleury S., Ponton C. 2020. "É:Calm Resource: a Resource for Studying Texts Produced by French Pupils and Students". *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. pp 4327–4332 Marseille. 11–16 May 2020.
- Gaubil Rachel. 2020. "Outiller la description de la morphologie adjectivale du primaire à l'université". Mémoire de master en Sciences du Langages – Parcours Industries de la Langue. Université Grenoble Alpes.



merci pour votre Attention !