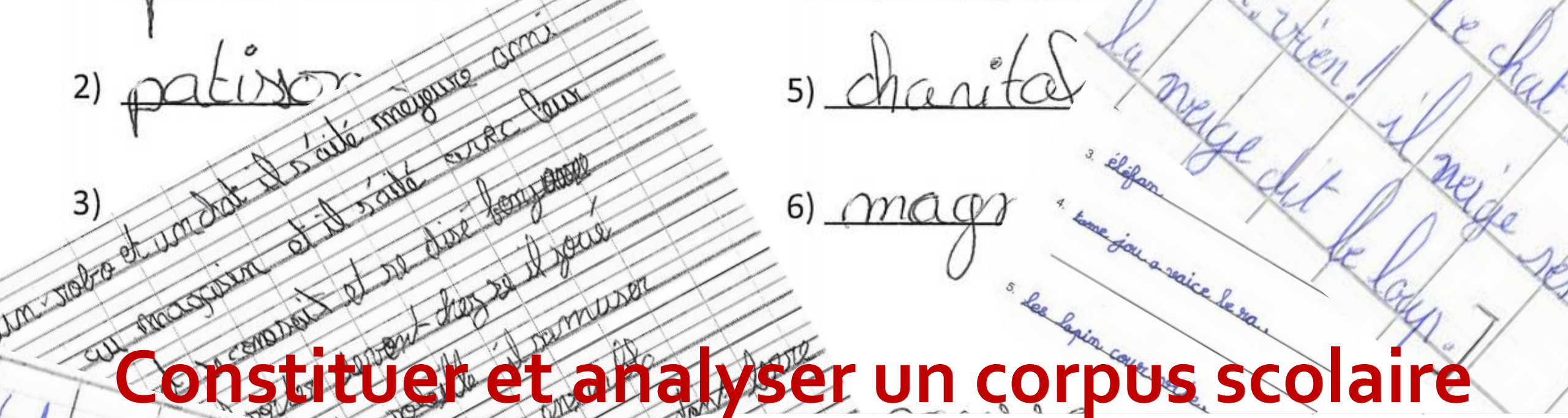


Constituer et analyser un corpus scolaire

L'approche Scoledit

Claude Ponton

Ecole de Printemps PEtALE 3 – 12 et 13 juin 2018

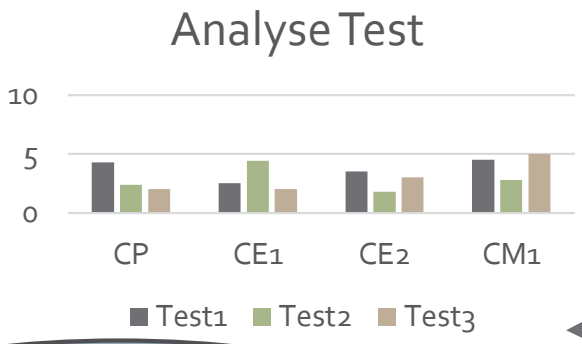
LIDILEM
UNIVERSITÉ Grenoble Alpes

Des copies à l'analyse…

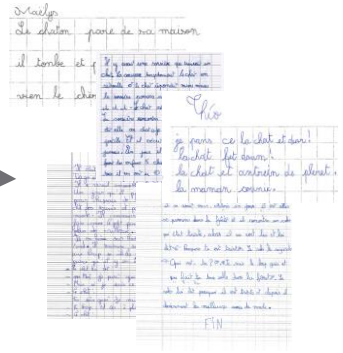
Recueil des données (protocole)

???

Question de la volumétrie et du partage des ressources



Traitements



Numérisation
Stockage, sauvegarde
perte de données

Transcription
Des choix (perte de données),
Un langage de description,
De l'interprétation, des erreurs

Didactique

Psycholinguistique

Linguistique

....

TAL

Informatique

```
                                    -008-005>
                                    ax en direct
                                    OP>
                                    ée
                                    /F> globe
                                    llenge, une aventure
périlleuse <G><CLA><POR>
#qui$ qu'
</POR></CLA></G>est
<G><CLA><ADE> #une$ la
</ADE></CLA></G> course
à la voile.
```

Extrait corpus Frida
(Granger, 2001)

tation
sement,
des erreurs

Aspects juridiques

Souvent parent pauvre…

Gros travail :

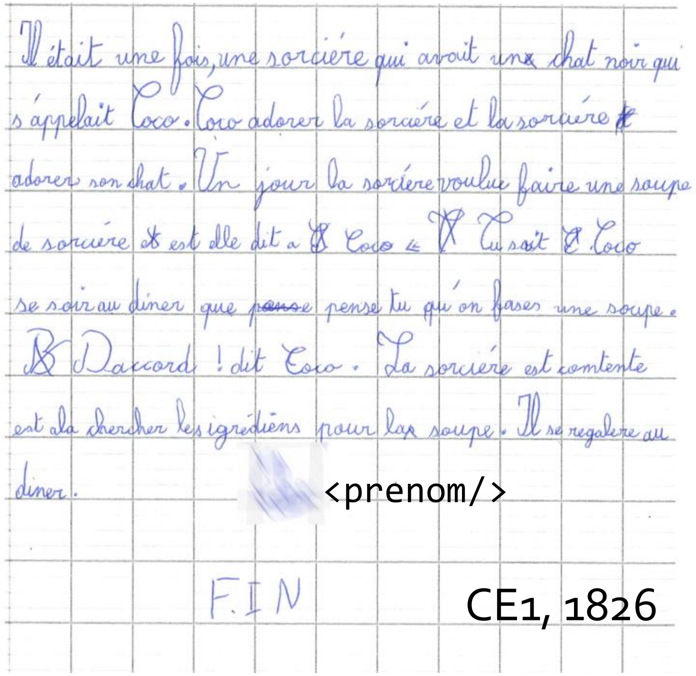
- autorisations parentales,
- anonymisation, floutage, destruction données…
- Déclaration CNIL

Enjeu éthique :

- respect vie privée, transparence…

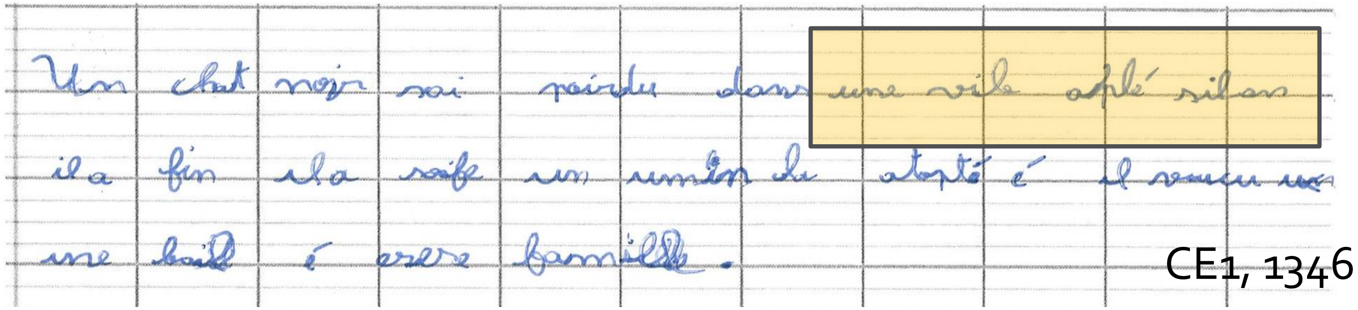
Enjeu pour la recherche et sa diffusion :

- disposer de données « utilisables » et partageables


Il était une fois, une sorcière qui avait une chat noir qui s'appelait Coco. Coco adorer la sorcière et la sorcière adorer son chat. Un jour la sorcière voulue faire une soupe de sorcière et elle dit : « Coco tu sait Coco se soir au dîner que pense pense tu qu'on fases une soupe. » « D'accord ! dit Coco. La sorcière est contente est ala chercher les igrédiens pour la soupe. Il se regalere au dîner.
<prenom/>
FIN

CE1, 1826

???


Un chat noir soi poirdu dans une vile aplé silon
il a fin ila saife un umin de atopté é il ...
une baile é erere famille.

CE1, 1346

Conférence Rachel Panckhurst
Colloque Cedil18, Grenoble
1er ou 2ème téléchargement de leur corpus SMS = Gendarmerie

L’équipe Scoledit

Catherine Brissaud

Corinne Totereau

Claire Wolfarth

Claude Ponton

- Des stagiaires
  - Web
  - XML/TEI
  - Statistiques
- Des vacataires
  - Préparation des données ,
  - Scans,
  - Transcriptions,
  - Normalisations

Contexte de la recherche

http://otus.u-grenoble3.fr/scoledit/
http://scoledit.org/scoledit/ (> juillet)

- Recherche IFÉ : Lire-écrire au CP
  - Recueil de dictées et de productions écrites en CP
  - [2014] 2507 productions pour 131 classes de CP
- [2014-2015] : Mémoire puis thèse de Claire Wolfarth : place du TAL dans l'analyse d'écrits scolaires ?
- Début du projet Scoledit : constitution d'un corpus longitudinal CP-CM2
  - [2015] Recueil en CE1, [2016] CE2…
- Projets liés en cours
  - E:Calm : ANR
  - Corpuscol : IRS/IDEX UGA
  - Projet Démarre SHS
  - Projet Ortolang/Corli

Scoledit

E-Calm
AGENCE NATIONALE DE LA RECHERCHE
Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques

http://e-calm.huma-num.fr/
http://otus.u-grenoble3.fr/corpuscol/
http://scoledit.org/corpuscol/ (> juillet)

Scoledit : un corpus en cours de constitution

CP	CE1	CE2	CM1	CM2
Dictées (Sept. 2013) 1131 dictées	Dictées (Juin 2015) 630 dictées	Productions de textes (Mai - Juin 2016) 1060 textes	Dictées (Mai - Juin 2017) 500-800 dictées	Dictées (Mai - Juin 2018) 500-800 dictées attendues
Dictées (Juin 2014) 954 dictées	Productions de textes (Juin 2015) 864 textes		Productions de textes (Mai - Juin 2017) 800-1100 textes	Productions de textes (Mai - Juin 2018) 800-1100 textes attendus
Productions de textes (Juin 2014) 965 textes				

~4500 textes

récupérés

transcrits

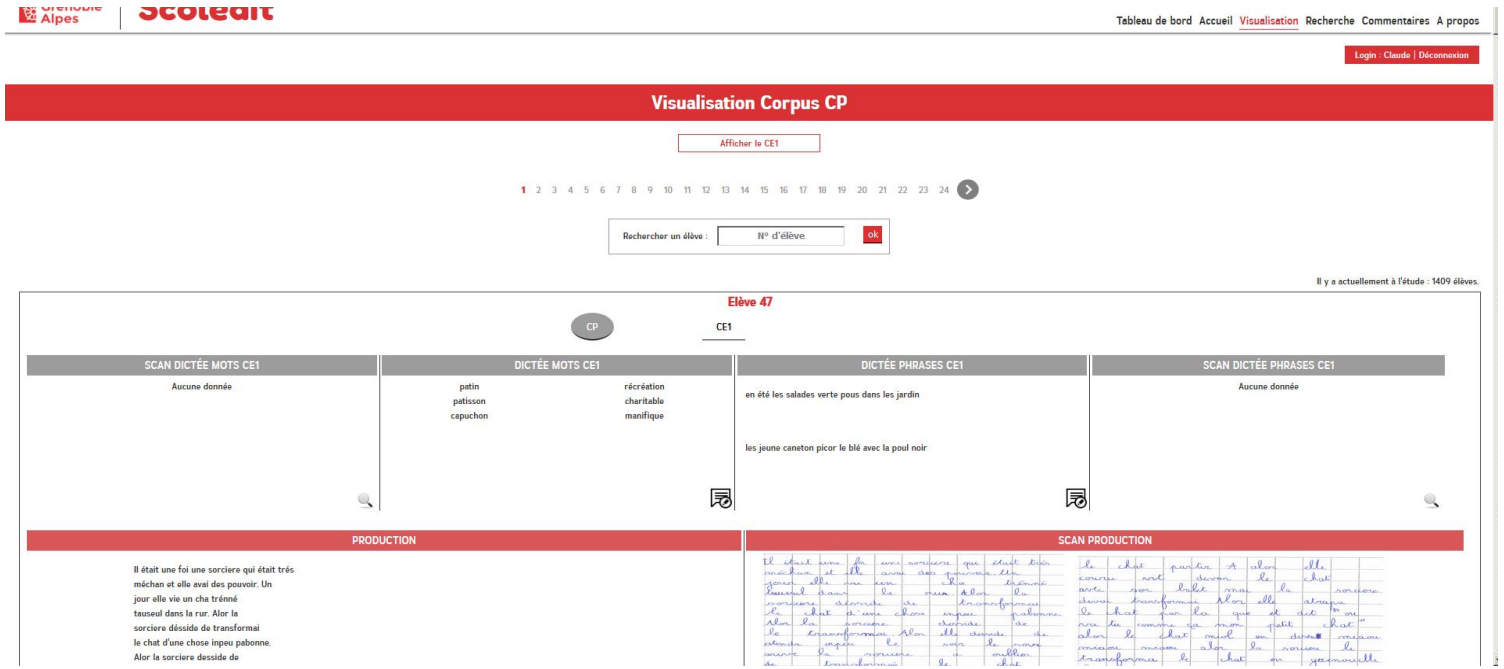
En cours

CORPUS LONGITUDINAL : ~370 élèves

Etat d'avancement

Scoledit : un site

- Maintenir le lien entre le projet et les enseignants impliqués
  - + présentation poster cette année
- Donner accès au corpus (enseignants, chercheurs, autres)
- Améliorer la qualité des données (système de commentaires)
- Proposer un outillage d'exploration
- Essayer de mesurer l'usage de l'outil (système de traces)



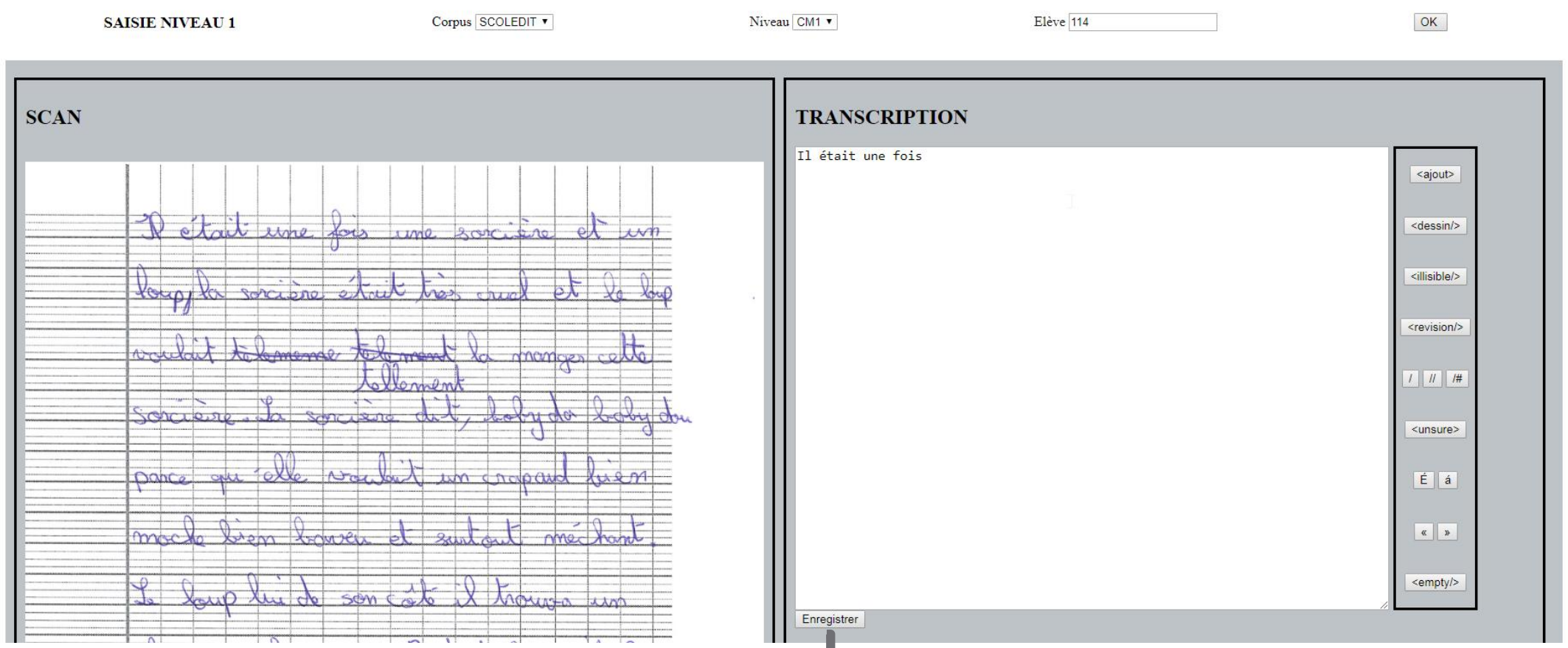
http://otus.u-grenoble3.fr/scoledit
http://scoledit.org/scoledit/ (à partir de juillet 2018)

Transcription Scoledit

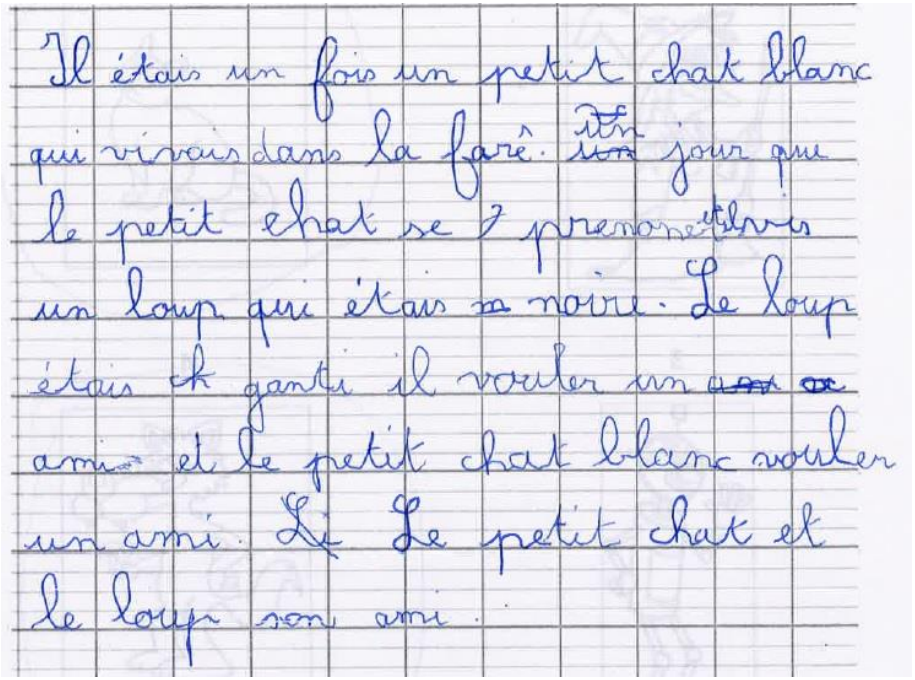
```
Il étais un fois un petit chat blanc / qui
vivais dans la forê.
<revision/><ajout>Un</ajout> jour que / le
petit chat se <revision/> prenoner <ajout>et
il</ajout> vis / un loup qui étais <revision/>
noire. Le loup / étais <revision/> ganti il
vouler un <revision/> <revision/> / ami et le
petit chat blanc vouler / un ami. <revision/>
Le petit chat et / le loup son ami.
```

- Objectifs
  - Description linguistique
  - Visualisation du corpus
- Productions Scoledit
  - Un seul jet avec d'éventuelles traces de révision de l'enfant
  - Pas d'intervention enseignante
- Transcription « semi-diplomatique »
  - Conservation et distinction des fins de ligne « physiques » *vs* volontaires
  - Peu ou pas de marques génétiques (balise <revision/>, <ajout>)
- Guide de transcription
  - En concertation au niveau national
  - Actuellement , stabilisation et bon accord inter-transcripteurs indépendamment de l'opération de déchiffrage
- Format
  - Un formalisme simple défini empiriquement en fonction des objectifs + travail de convergence au niveau national
  - Vers un format commun en XML-TEI (ANR E-Calm + financement Ortolang/Corli)
- Qualité des données
  - Une passe de transcription (vacataires) + une passe de révision (chercheurs)
  - Visualisation parallèle scan/transcription : système de commentaires en cas d'erreurs
  - Outil de transcription simple

Outil de transcription



Transcription *vs* visualisation



CE1, 212

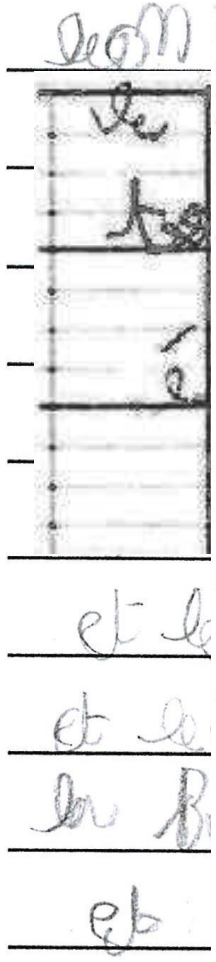
```
Il étais un fois un petit chat blanc / qui
vivais dans la forê.
<revision/><ajout>Un</ajout> jour que / le
petit chat se <revision/> prenoner <ajout>et
il</ajout> vis / un loup qui étais <revision/>
noire. Le loup / étais <revision/> ganti il
vouler un <revision/> <revision/> / ami et le
petit chat blanc vouler / un ami. <revision/>
Le petit chat et / le loup son ami.
```

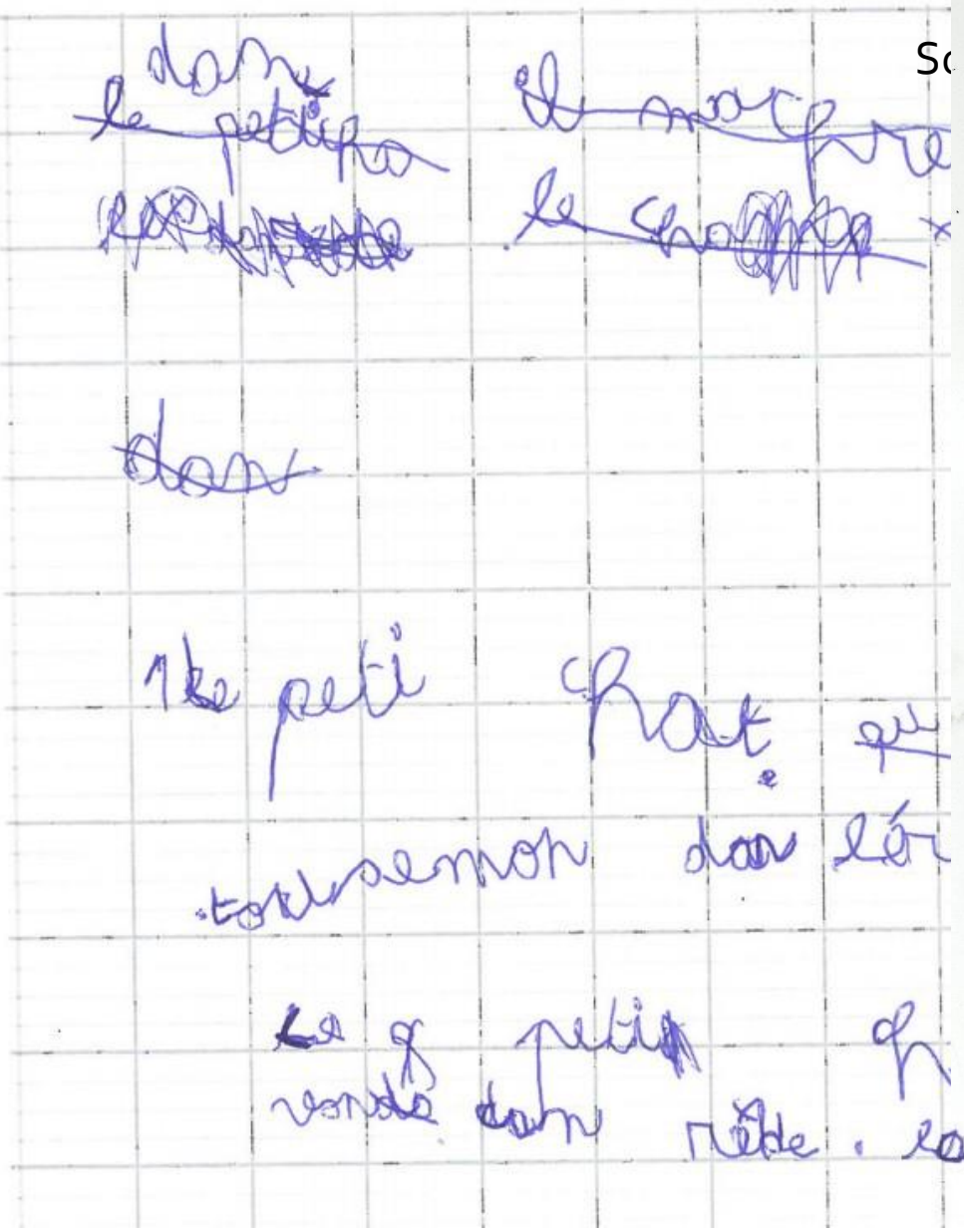
Il étais un fois un petit chat blanc
qui vivais dans la forê. Un jour que
le petit chat se prenoner et il vis
un loup qui étais noire. Le loup
étais ganti il vouler un
ami et le petit chat blanc vouler
un ami. Le petit chat et
le loup son ami.

E15

Ecriscol, EC-CM2-2016-PNT-D3-E15-V1-001

Raconte l'his





Taxte 1

Mardi 31 janvier

Rédaction

Que feras-tu quand tu seras adulte? Raconte par écrit une de Tes journées.

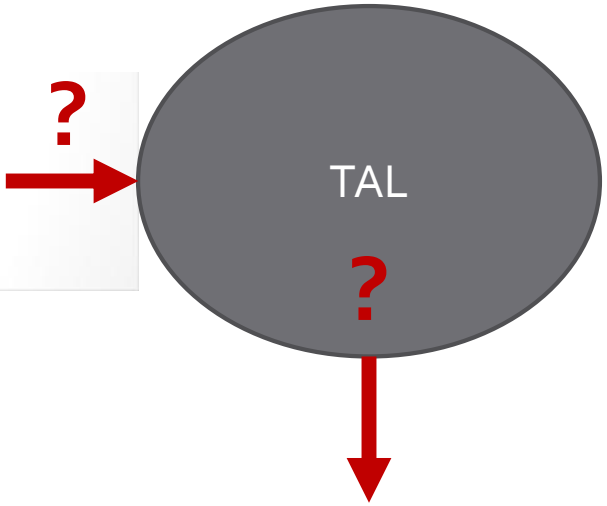
Aide :
- motivation
- l'esprit d'équipe
- entraînement/ échauffement/ mi-temps/
- fêter la victoire

Quand je serai adulte, je serai rugbymen (=D) serons rugbyman car les chocs ~~sont~~ puissants, mes crochets seront agiles, les plaquages seront intenses et mes essais seront magnifiques (On Ont ? Qui ?) L'équipe sera ~~tous~~ motivés pour le match Bayonne contre la section (nay) paloise (Nay). Il y ~~aura~~ aura l'esprit d'équipe, on s'entraînera, on s'échauffera ~~et~~ après, il y aura la

Approche TAL

- **Transcription**

le chat éfati gé é tonbe éaprê / révêlle sa maman é cés frêre é sa / maman le ramên avêc sec frére / é an site il dorme avêc séc frêre



Analyses : lexique, erreurs, syntaxe, cohérence/cohésion du discours….

Approche

- Approche empirique et « moins-disante » (Kraif & Ponton, 2007)
- Le TAL comme une assistance à la constitution et à l'exploitation de ces corpus
- S'appuyer sur le contexte de production
- Travailler par comparaison entre production et une forme d'attendu (« normalisation »)
  = forme d'annotation déportée (*stand-off annotation*)

Contexte

Pas ou peu de travaux en TAL

Textes éloignés de la norme

Difficulté à réutiliser les outils TAL classiques

Annotations intégrées

Elève 651 : "…le pe ticha ete tonbé et sa MaMan et se révéie"

Elève 651 : "… <?>pe ticha<?> <seg>été</seg> <

Elève 651 : "… pe ticha été <lex cat="VER

Elève 651 : "… pe ticha été <err type="?"

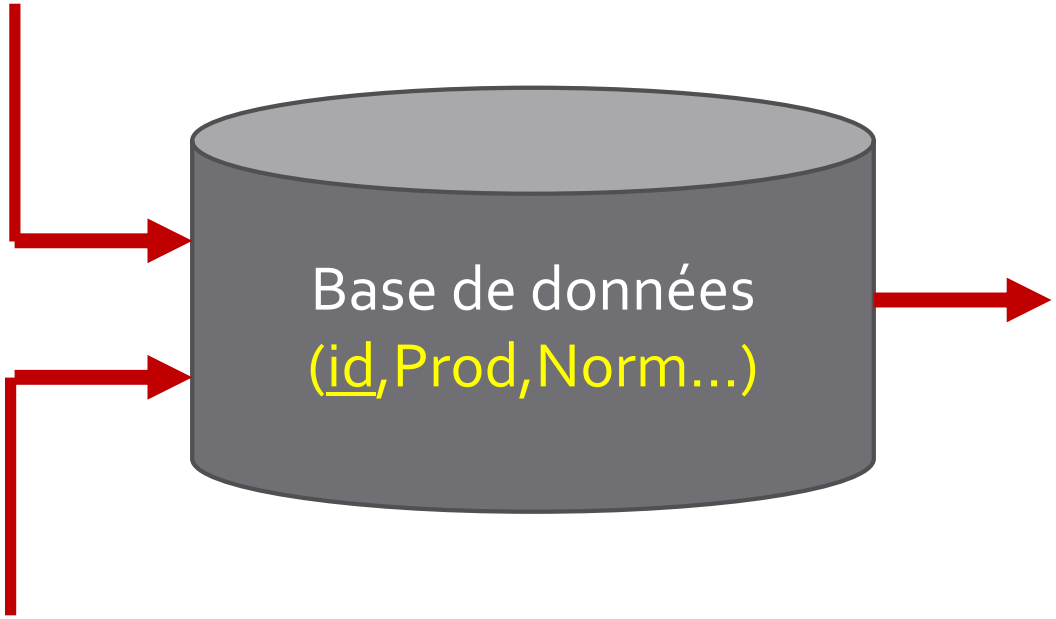
Elève 651 : "… <err>pe ticha été</err> <err cor

Elève 651 : …

Lourdeur du travail
Complexité du schéma
Prédétermination des catégories, des attributs…

Approche par « normalisation »

Prod Elève 651 : "…le pe ticha ete tonbé et sa MaMan et se révéie"



Norm Elève 651 : "…le petit chat était tombé et sa maman elle se réveille"

Normalisation ?

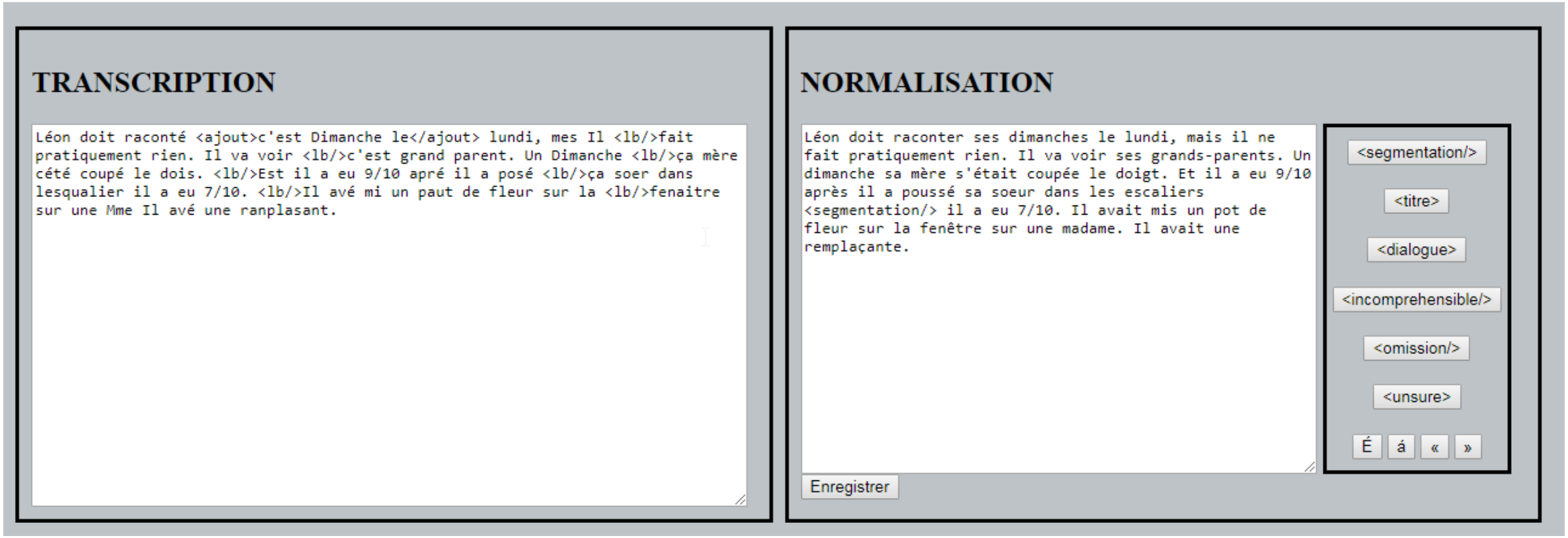
- Avantages
  - Pas de formalisme lourd
  - Plus simple qu'annotation
  - Plus de liberté dans l'exploitation ?
  - Réutilisabilité des procédures : ex. dictées, exercices de réécriture…

- Inconvénient
  - Alignement automatique => risque d'erreur augmenté

Guide de normalisation

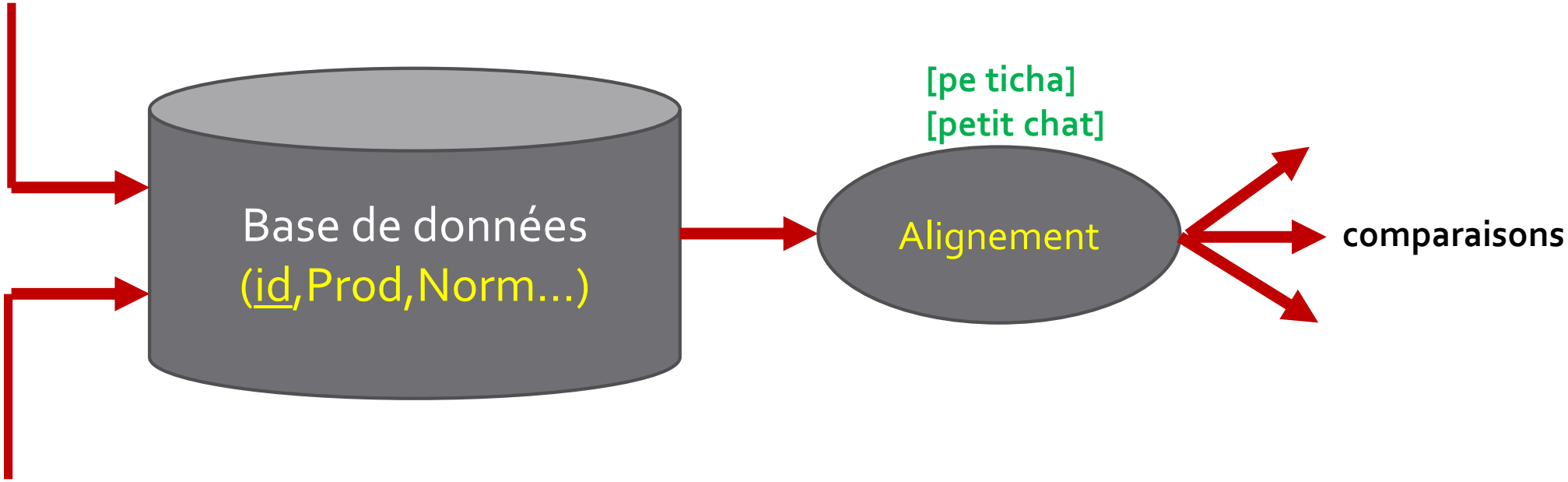
- Des choix de normalisation dépendront les analyses ultérieures
- Longues discussions… travail à peu près stabilisé
- **Quelques éléments**
  - Rester au plus près de la production de l'enfant
  - Rétablir la segmentation et l'orthographe
  - Rétablir les accords en genre et en nombre
  - Rétablissement de la négation et de certaines marques de ponctuation non ambigües
  - Conservation des temps verbaux
  - Marquage d'éléments de segmentation et de structure : <segmentation>, <titre>, <dialogue>

Outil de normalisation



Approche par alignement/comparaison

Prod Elève 651 : "…le pe ticha ete tonbé et sa MaMan et se révéie"



Norm Elève 651 : "…le petit chat était tombé et sa maman elle se réveille"

Alignement production-normalisation

$CP_{,49}$ $CP_{,49}$

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)

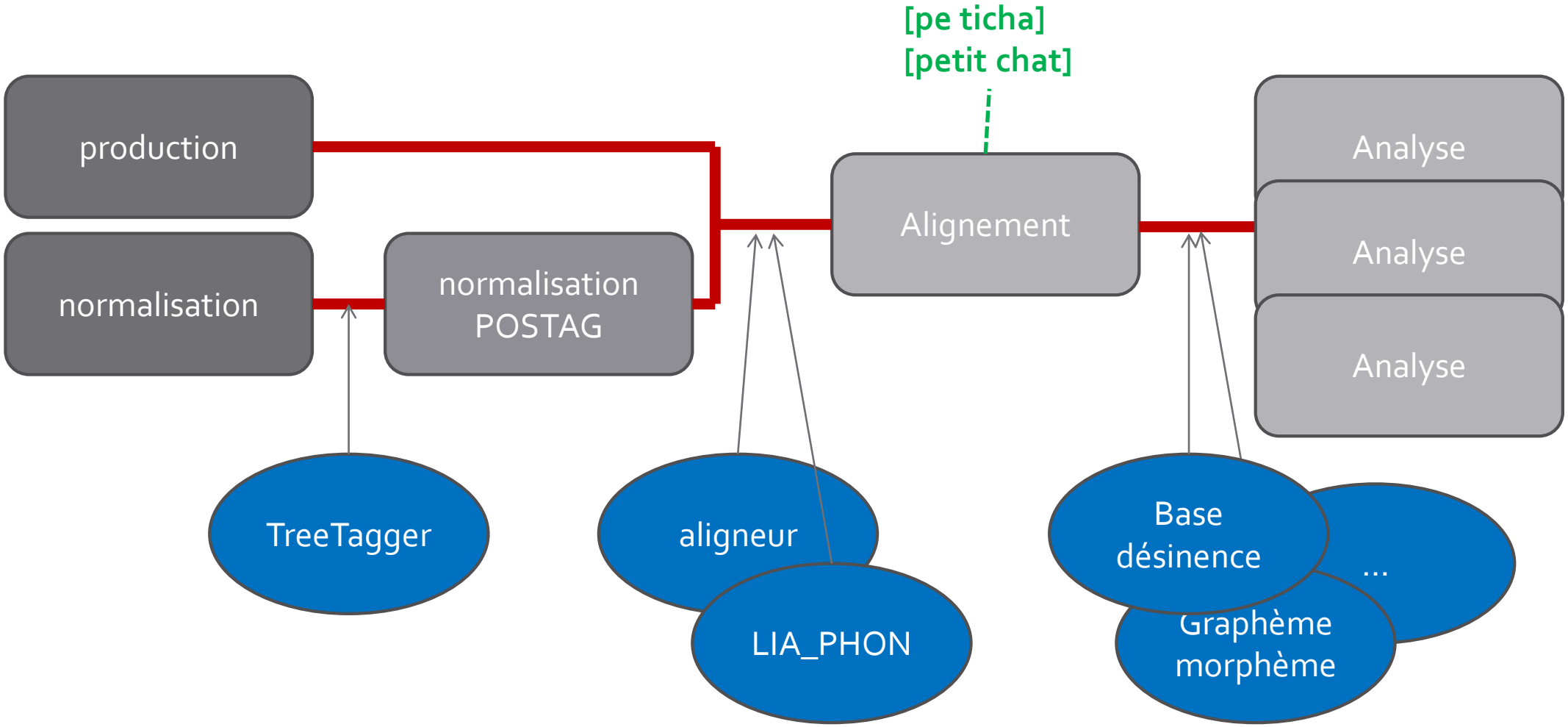
1. Comparaison graphique stricte
   nouvo ⇔ *nouveau*
2. Comparaison phonologique
   /nuvo/ ⇔ */nuvo/*
3. Comparaison archiphonologique

voulé / voulait : */vule/* ⇔ */vulɛ/*

Exemple de sortie de l'aligneur

A	B	C	D	E	F	G	H
N° élève	Niveau	Forme produite	Forme normée	Catégorie	Lemme	Statut de l'erreur	Segmentation
52	CP	chat	chats	NOM	chat	phonologie normée	normée
52	CP	abide	habite	VER:pres	habiter	non normé	normée
52	CP	désan	descend	VER:pres	descendre	non normé	normée
52	CP	tonbe	tombe	VER:pres	tomber	phonologie normée	normée
52	CP	a	a	VER:pres	avoir	normé	normée
52	CP	dit	dit	VER:pres	dire	normé	normée
52	CP	fé	fait	VER:pres	faire	équivalence archiphonologique	normée
53	CE1	histoire	histoire	NOM	histoire	normé	normée
53	CE1	chat	chat	NOM	chat	normé	normée
53	CE1	loup	loup	NOM	loup	normé	normée
53	CE1	loup	loup	NOM	loup	normé	normée
53	CE1	maison	maison	NOM	maison	normé	normée
53	CE1	chat	chat	NOM	chat	normé	normée
53	CE1	chat	chat	NOM	chat	normé	normée
53	CE1	maison	maison	NOM	maison	normé	normée
53	CE1	toc	toc	NOM	toc	normé	normée
53	CE1	Bonjour	Bonjour	NOM	bonjour	normé	normée
53	CE1	loup	loup	NOM	loup	normé	normée
53	CE1	porte	porte	NOM	porte	normé	normée
53	CE1	Bonjour	Bonjour	NOM	bonjour	normé	normée
53	CE1	chat	chat	NOM	chat	normé	normée
53	CE1	maison	maison	NOM	maison	normé	normée
53	CE1	loup	loup	NOM	loup	normé	normée
53	CE1	tée	thé	NOM	thé	phonologie normée	normée

Chaine de traitement



Alignement Base-désinence

- Objectif
  - Description plus fine des erreurs sur les verbes
  - Niveau base et désinence
- Tiroirs verbaux étudiés
  - Infinitif, présent de l'indicatif, imparfait, passé simple et participe passé
- Modèle linguistique
  - Entrée selon les désinences
  - 2 modèles
    - Infinitif en /e/ et infinitif en /R/ à l'oral
    - Martinet, 1979 ; Meleuc et Fauchart, 1999; Blanche-Benveniste, 2002 ; Pellat, 2009

Temps	Types de verbe	P1	P2	P3	P4	P5	P6
Présent	Verbes en -er	-e	-es	-e	-ons	-ez	-ent
	Autre verbes	-s	-s	-t (ou Ø)			
Imparfait	Tous les verbes	-ais	-ais	-ait	-ions	-iez	-aient
Passé simple	Verbes en -er	-ai	-as	-a	-^mes	-^tes	-rent
	Autre verbes	-(v)s	-(v)s	- (v)t			

Alignement Base-désinence

- 4 cas de comparaison
  - Normé
    - di + t // di + t
  - Erreur sur la base seule
    - fai+re // fé+re
  - Erreur sur la désinence seule
    - av+ait // av+ai
  - Erreur sur la base ET la désinence
    - miaul+e // miol + Ø

eleve	niv	lemme	cat	statut	forme	baseForme	desiForme	prod	baseProd	desiProd
83	CE2	suivre	ppre	normé	suivant	suiv	ant	suivant	suiv	ant
868	CE1	faire	pres	phonologi	fais	fai	s	fait	fai	t
1147	CE1	être	impf	non normé	était	ét	ait	étét	ét	ét
2037	CE1	faire	infi	équivalen	faire	fai	re	fére	fé	re
2442	CE2	savoir	pres	normé	sais	sai	s	sais	sai	s
2450	CE1	dire	pres	normé	dit	di	t	dit	di	t
2531	CE2	être	impf	normé	était	ét	ait	était	ét	ait
2899	CE1	être	pres	normé	est	es	t	est	es	t
2971	CE1	avoir	impf	phonologi	avait	av	ait	avai	av	ai
2986	CE1	vouloir	pres	phonologi	veux	veu	x	ve	ve	µ

Exemple de sortie de l'aligneur

Analyses

- Une première publication sur la morphographie verbale
  - Brissaud C., Totereau C., Ponton C., Wolfarth C. (à paraître en 2018). Usage d'un corpus longitudinal, le cas de la morphologie verbale. *Revue Repères* : COLLECTER, INTERPRÉTER, ENSEIGNER L'ÉCRITURE. Analyses linguistiques des écrits d'élèves
  - Données traitées manuellement sous Excel…
- Traitements statistiques
  - Stagiaire Master 2 Traitement des données
    - Organiser les données
    - Modèles statistiques
    - Interface d'interrogation : démo du prototype

Évaluation de la chaine de traitement

- « Des erreurs à tous les niveaux »…
  - Transcription/normalisation : humain
  - Alignement : LIA_Phon, aligneur
  - Étiquetage : treetagger
  - Comparaison : les différents modules…
- Nécessité d'évaluer cette chaîne pour quantifier le % d'erreur
  - Aligneur(s) : A priori bons résultats en précision, moins bons en rappel
- Travail en cours = temps long…
  - Constitution d'un corpus de référence (étiquetage manuel)
  - Évaluation des différents éléments
  - Tests de non-régression
  - …
- Permettra d'améliorer la qualité des traitements et donc des sorties

Perspectives

- Fin de constitution du corpus
  - CM2 (DémarreSHS) – en cours
  - Transcription et normalisation (Lidilem, DémarreSHS)
- Mise à disposition du corpus et des outils
  - Refonte du site : CP-CM2, intégration de l'alignement (concordancier) [fin 2018]
  - Corpus (transcription) en téléchargement (site Scoledit, ANR…)
- Amélioration de la qualité des données
  - Révision des transcriptions et des normalisations
- Évaluation de la chaine de traitement
  - Amélioration de la qualité (rappel/précision) des traitements
- Développement d'un module d'alignement graphème/phonème
- Autres modules…
- Outil d'exploration statistique des données

Projets en cours

- E:CALM (financement ANR) : 2018-2020 [Clesthia, Lidilem, Escol, CLLE]
  - **É**criture scolaire et universitaire **: C**orpus, **A**nalyses **Li**nguistiques, **M**odélisations didactiques
  - Vaste corpus d'écrits du primaire à l'université
  - Caractérisation des écrits
  - Étude des modalités de l'écriture dans les avant-textes et les textes (influence réciproque des écrits remis et des commentaires des enseignants)
  - Perspectives didactiques
- Corpuscol (financement IRS IDEX UGA) : 2017-2018
  - Constitution et diffusion d'un corpus commun longitudinal à partir des corpus Scoledit (Lidilem) et Longit (LARAC)
  - Mise au point d'une grille d'évaluation des productions scolaires
- Ortolang/Corli : décembre 2017-février 2018
  - Définition d'un format commun au standard XML/TEI des corpus scolaires des laboratoires CLESTHIA et LIDILEM
  - Développement de modules de transfert

Références

- Wolfarth C., Brissaud C., Ponton C. (à paraître en 2018 dans la revue *Dyptique*) Transcrire et normer un corpus scolaire, pour quelles analyses ? *Actes du colloque international, Enseignement et apprentissage de l'écriture de la maternelle à l'université et dans la formation tout au long de la vie*, Bordeaux, 19-21 octobre 2016
- Brissaud C., Totereau C., Ponton C., Wolfarth C. (à paraître en 2018). Usage d'un corpus longitudinal, le cas de la morphologie verbale. *Revue Repères* : COLLECTER, INTERPRÉTER, ENSEIGNER L'ÉCRITURE. Analyses linguistiques des écrits d'élèves
- Wolfarth C., Ponton C., Totereau C. (2018). Which Method to Develop a Natural Language Processing Tool to Automatic Analyse First Language Learner Corpora? In 13th Teaching and Learning Corpora Conference . 18-21 july 2018. Cambridge
- Wolfarth C., Ponton C., Totereau C. (2017). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire. Dans Doquet C., David J. et Fleury S., Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement, *Corpus*, 16 | 2017, 185-214.
- Wolfarth C. (2017). Aligner production et normalisation : une première approche pour l'étude d'écrits scolaires. Dans *Vol. RECITAL*, p. 56-69. Présenté à RECITAL 2017, Orléans.
- Wolfarth C., Ponton C., Brissaud C. (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.