Du TAL dans les écrits scolaires : PREMIÈRES APPROCHES

Claire Wolfarth, Claude Ponton, Catherine Brissaud

Lidilem

Université Grenoble Alpes

Le projet Scoledit

Contexte

- Travaux en écriture moins fréquents qu'en lecture (Boscolo, 2008)
- Développement des recherches sur l'écriture
- Ces recherches s'appuient sur des corpus souvent restreints, spécifiques ou difficilement accessibles (Elalouf, 2005 ; Auriac-Slusarczyk, Gunnarsson, 2014)
- *Enquête « Lire – Écrire CP » de l'Institut Français de l'Éducation (2013-2015)*

Objectif

- Constitution d'un corpus scolaire longitudinal en poursuivant le recueil du corpus jusqu'en CM2 (2018)

Enjeux

- Pour la linguistique :
  - Description en synchronie et en diachronie des caractéristiques des écrits d'apprenants de 6 à 12 ans
  - Meilleure connaissance des phénomènes d'acquisition de l'écriture
- Pour la didactique:
  - Développement de séquences à partir de ces connaissances

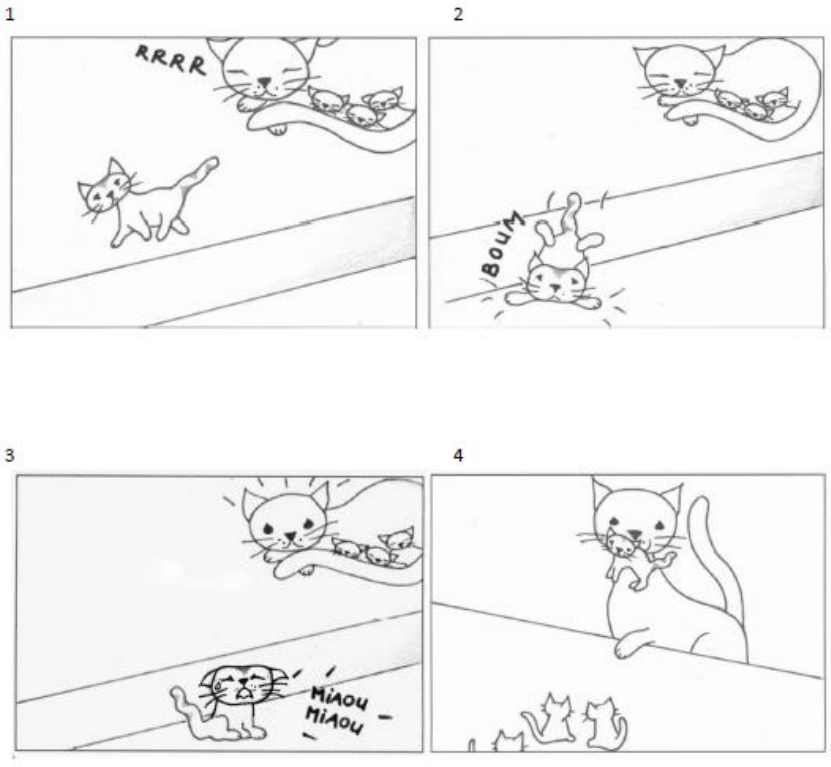
Corpus : Structure

<table>
<tr><th>CP</th><th>CE1</th><th>CE2</th><th>CM1</th><th>CM2</th></tr>
<tr><td>Dictées (Sept. 2013)
1151 dictées</td><td rowspan="2">Dictées (Juin 2015)
700-900 dictées</td><td rowspan="3">Productions de textes (Mai - Juin 2016)
900-1000 textes</td><td rowspan="3">Productions de textes (Mai - Juin 2017)
800-1100 textes attendus</td><td rowspan="3">Productions de textes (Mai - Juin 2018)
800-1100 textes attendus</td></tr>
<tr><td>Dictées (Juin 2014)
975 dictées</td></tr>
<tr><td>Productions de textes (Juin 2014)
965 textes</td><td>Productions de textes (Juin 2015)
700-900 textes</td></tr>
</table>

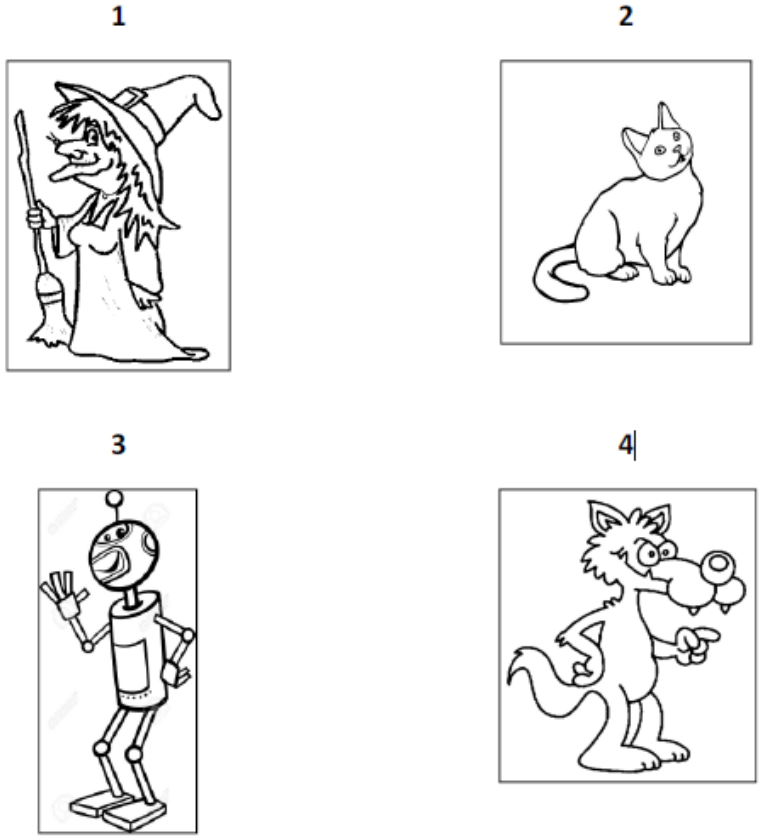
Corpus : Méthodologie

I – Corpus
II – Travaux TAL
III – Perspectives

Recueil CP (15 min)



Recueil CE1 (20 min) et CE2 (30 min)

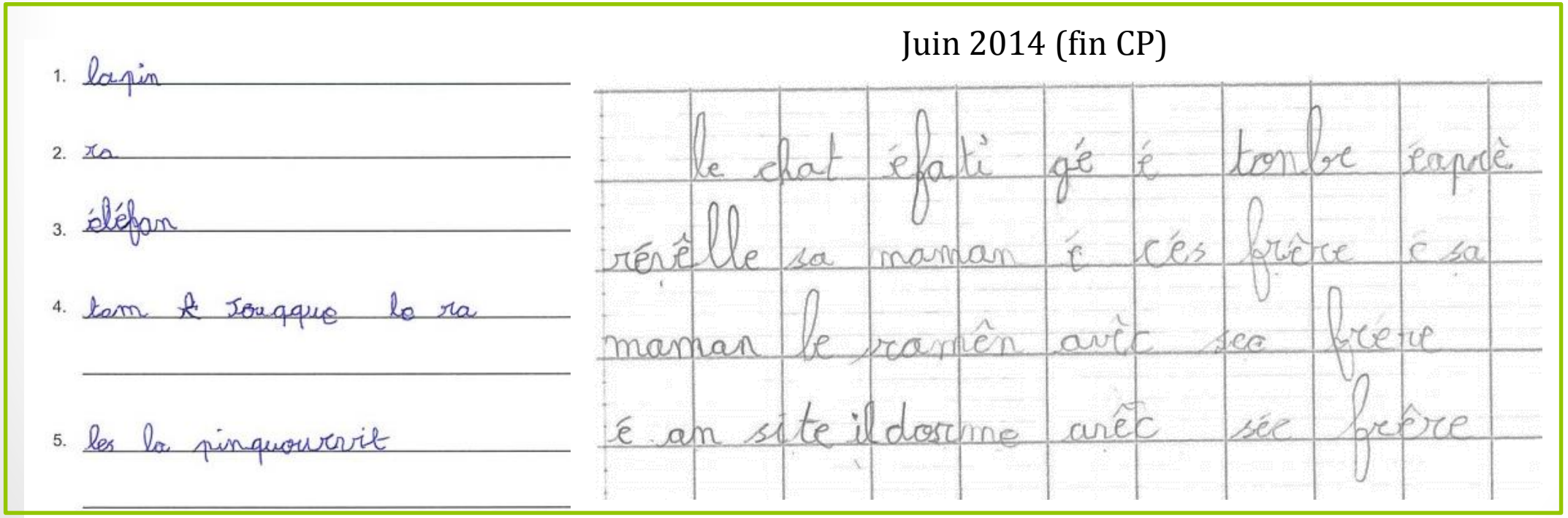


Corpus : Données recueillies (CP)

(ex. élève 1558)

Septembre 2013
(début CP)

Juin 2014 (fin CP)



Corpus : Données recueillies (CE1)

(ex. élève 1558)

Juin 2015 (fin CE1)

1/ ~~patimp~~.
2/ ~~patison~~.
3/ capuchons.
4/ récréations.
5/ charitable.
6/ manifice

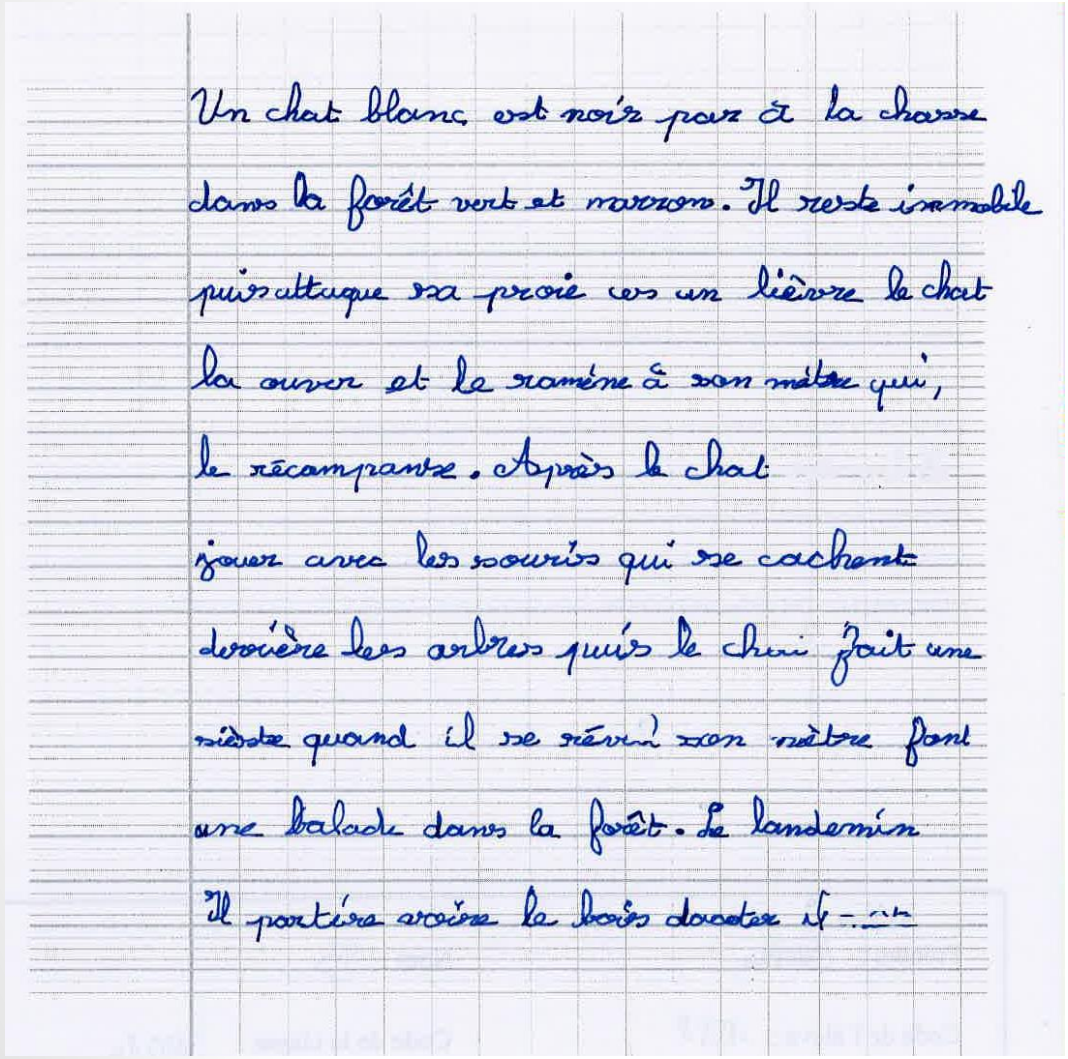
~~Les~~ En été, les salade verte pouse dans les jardins. Les jaunescanetons picore le blé avec la poule noire.

Le loup se ~~promen~~ promene dans les bois camp un chat ~~se~~ sor dis buson et lui di quefaitu je me promene pourquoi on peus se promene ensemble set une bonne idet nes pas pourcoi pas suivre la riviere il nia pas de chaseur sert sur sinan il matir desu et je cour baucoup maime que je me cache dans un buisson et il me chere et tire par tout et je cour a nouveau

il fau faire atantion il sont nombre et mete des pillege de partot et on fait trais atantion il met des corde de par tout.

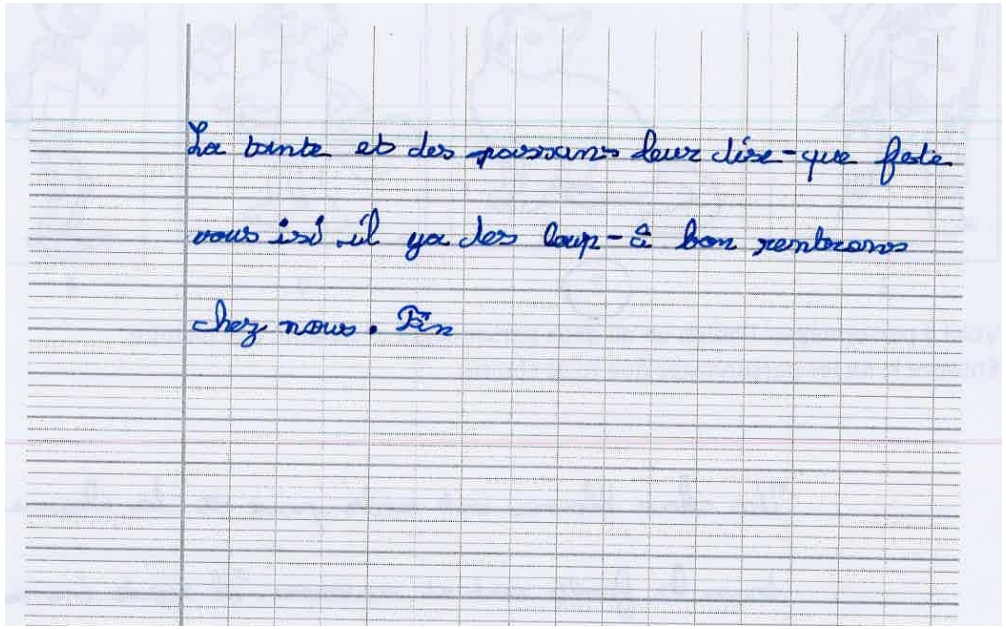
Corpus : Données recueillies (CE2)

(ex. élève 1558)



Un chat blanc est noir part à la chasse dans la forêt vert et marron. Il reste immobile puis attaque sa proie ces un lièvre le chat la avoir et le ramène à son maître qui, le récampanse. Après le chat jouer avec les souris qui se cachent derrière les arbres puis le chat fait une siesta quand il se réveil son maître font une balade dans la forêt. Le lendemain Il partira croire le bois dacater il y [illegible]

Juin 2016 (fin CE2)



La tante et des passants leur dire-que faite vous ici il ya des loup-à bon rentrons chez nous. Fin

UNIVERSITÉ Grenoble Alpes

Scoledit (version CP)

Accueil Visualisation Commentaires A propos

Login : ClaireWolfarth | Déconnexion

Visualisation

Recherche par :

Forme

Afficher tous les élèves

1558

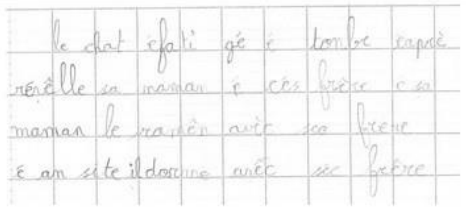
ok

Résultat de votre recherche :

Il y a actuellement à l'étude : 1169 élèves.

Elève 1558

SCAN DICTÉE SEPTEMBRE	DICTÉE SEPTEMBRE	DICTÉE JUIN	SCAN DICTÉE JUIN
	[illegible] AP RA É TA	[illegible] lapin éléphan ra tom jouqque le ra les la pinquourvit	

PRODUCTION	SCAN PRODUCTION
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	

Scoledit

Contact

Réalisation : Alan Gauthier

Spécificités du corpus : un défi pour le TAL

- Corpus manuscrit (transcription nécessaire)
- Corpus très « fautif »
- Longitudinal (évolutif)
- Taille du corpus
- Champ d'étude très récent en TAL

Ex. élève 49 (CP)

TAL et corpus scolaires

- Hypothèses :
  - TAL comme aide à l'exploitation du corpus
  - TAL nourri par :
    - Le contexte de production
    - Les travaux dans différents domaines connexes
- Méthodologie :
  - Approche empirique
  - Appui sur les niveaux les plus éprouvés

TAL : État de l'art

- Correction automatique de textes pour experts (Kukich, 1992)
- Normalisation de corpus peu normés
  - SMS, Tweets (Kobus *et al.*, 2008, Beaufort *et al.*, 2010, Baranes, 2012)
- TAL et apprentissage des langues
  - Freetext (Granger, Vandeventer, Hamel, 2001)
  - Exxelant (Antoniadis, Ponton, Zampa, 2010)

TAL : Particularités des productions de CP

- Écriture avec appui de l'oral

« tonb » (*tombe*)

- Segmentation en mots pas toujours maitrisée

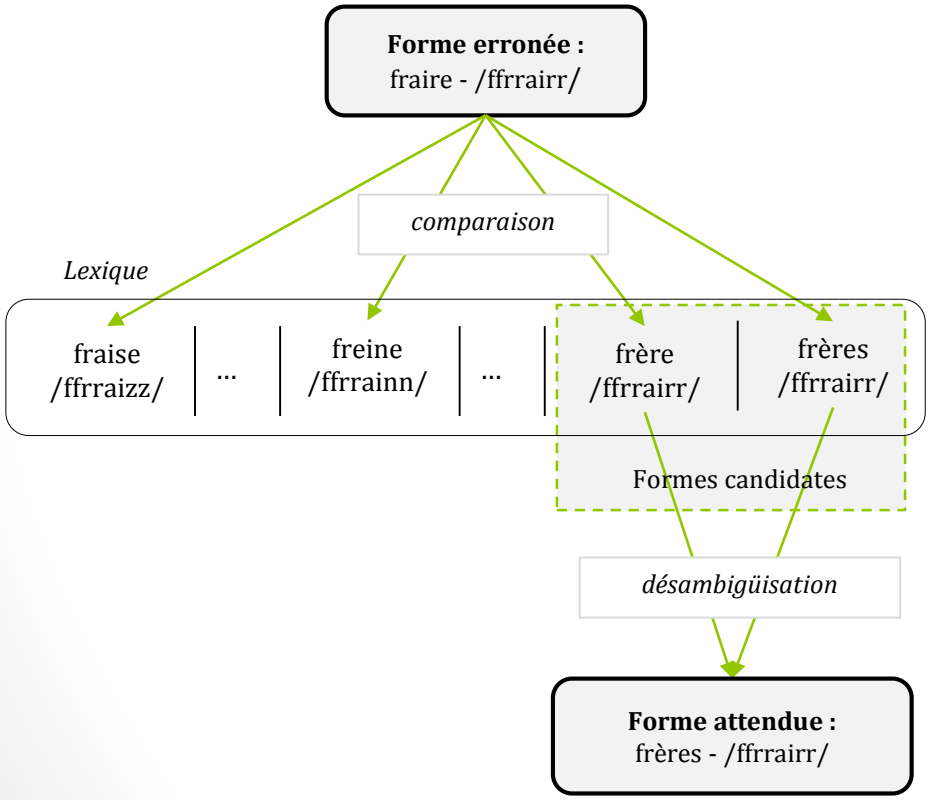
« Le petit chat sanva pan dan cesa maman dore. [...] » (*Le petit chat s'en va pendant que sa maman dort*.)

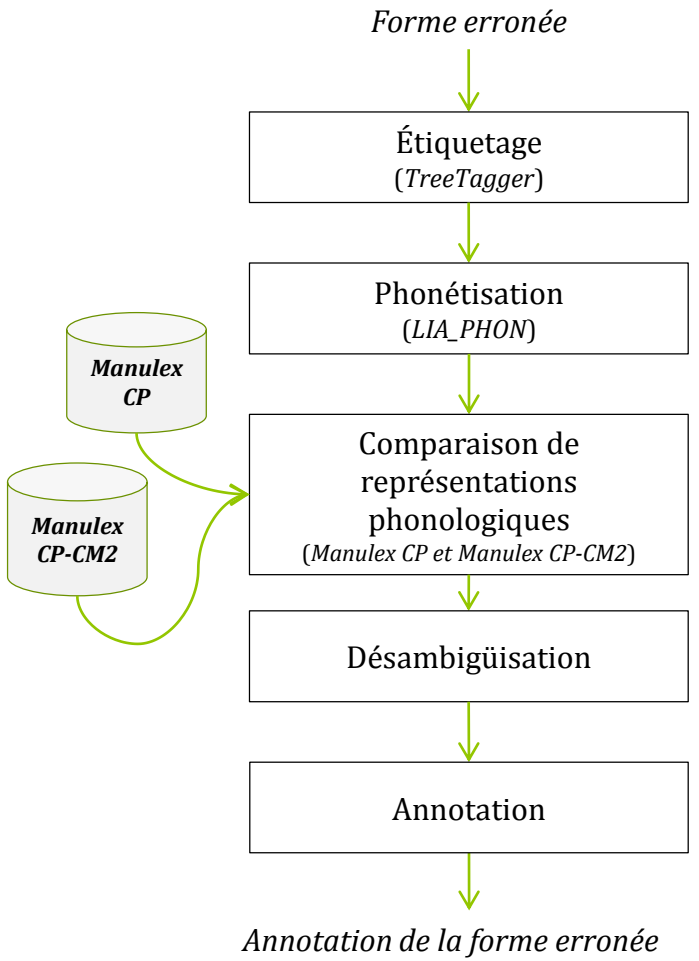
- Structures syntaxiques simples ou relativement simples

« le petit chas par et il sé fes male et mètn il plere sa maman se fach » (*Le petit chat part et il s'est fait mal et maintenant il pleure. Sa maman se fâche.)*

TAL : Identification des formes normées par comparaison phonologique

Procédure d'identification de la forme attendue



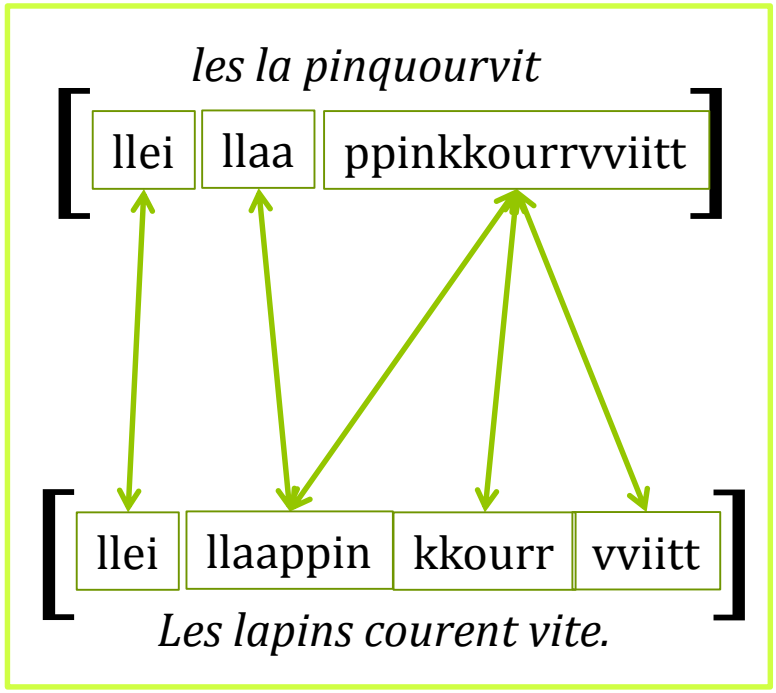


Processus d'annotation d'une forme erronée

TAL : Annotation par alignement avec un « corrigé »

Exemple

- « les la pinquourvit » (*les lapins courent vite*)



Identification des formes attendues par alignement

Perspectives

- Poursuite du recueil en CM1 et en CM2
- Schéma d'annotation
- Alignement entre les productions et les « formes attendues »
- Réflexion sur le statut de la « forme attendue »
- Publication du corpus sur un site web à destination des enseignants et des chercheurs impliqués
  - Instaurer un dialogue entre les deux communautés
  - Améliorer la qualité des données
  - Développer des fonctionnalités adaptées
- Publication sur Ortolang

Merci pour votre attention

Références

ANTONIADIS G., PONTON C., ZAMPA V. (2010). Exxelant et Mirto – Deux exemples d'environnement d'ALAO intégrant des outils TAL. *Multilinguisme et traitement des langues naturelles*. Montréal, Canada : PUQ.

AURIAC-SLUSARCZYK E., GUNNARSON-LARGY C. (2014). *Écriture et réécritures chez les élèves: un seul corpus, divers genres discursifs et méthodologies d'analyse*. Academia.

BARANES M. (2012). Vers la correction automatique de textes bruités: Architecture générale et détermination de la langue d'un mot inconnu. Actes de *RECITAL'2012 - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, 95-108.

BEAUFORT R., ROEKHAUT S., COUGNON L.-A. & FAIRON, C. (2010). Une approche hybride traduction/correction pour la normalisation des SMS, Actes de la 17e conférence sur le traitement automatique des langues naturelles (TALN'10), Montréal.

BÉCHET F. (2001). LIA_PHON - Un système complet de phonétisation de textes, *Traitement Automatique des Langues (T.A.L.)* 42(1), 47-67.

BOSCOLO, P. (2008). Writing in primary school. *Handbook of research on writing: History, society, school, individual, text*, 293-309.

DENIS P., SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation* 4(46), 721-736.

ELALOUF M.-L. (2011). Constitution de corpus scolaires et universitaires : vers un changement d'échelle ? Pratiques, 149-150.

ELALOUF M.-L. (dir) (2005). *Écrire entre 10 et 14 ans. Un corpus, des analyses, des repères pour la formation*, SCérén, CRDP de Versailles.

GRANGER S., VANDEVENTER A., HAMEL M.-J. (2001). Analyse des corpus d'apprenants pour l'ELAO basée sur le TAL. *Traitement automatique des langues* 42(2), 609-621.

KRAIF O., PONTON C. (2007). Du bruit, du silence et des ambiguïtés: que faire du TAL pour l'apprentissage des langues. Actes de *TALN* (Toulouse).

KUKICH K. (1992). *Techniques for Automatically Correcting Words in Text*. ACM Computing Surveys 24 (4), 377–439.

ORTÉGA É., LÉTÉ B. (2010). « eManulex: Electronic version of Manulex and Manulex-infra databases », http://www.manulex.org

Wolfarth, C., Ponton, C., Totereau, C. (2016). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire. Corpus.