The É:Calm Resource: Transcription, Encoding and Annotation of Handwritten Manuscripts produced by French Pupils and Students

Claire Doquet, Silvia Federzoni, Serge Fleury, Lydia-Mai Ho-Dac, Sara Mazziotti, Arnaud Moysan, Claude Ponton

The É:Calm Resource: Transcription, Encoding and Annotation of Handwritten Manuscripts produced by French Pupils and Students

Claire Doquet[1] Silvia Federzoni[2] Serge Fleury[1] Lydia-Mai Ho-Dac[2] Sara Mazziotti[1] Arnaud Moysan[1] Claude Ponton[3]

[1] CLESTHIA, Université Paris 3 Sorbonne Nouvelle [2]CLLE, CNRS, Université Toulouse Jean Jaurès [3]LIDILEM, Université Grenoble Alpes

Contact: claire.doquet@sorbonne-nouvelle.fr, lydia-mai.ho-dac@univ-tlse2.fr, claude.ponton@univ-grenoble-alpes.fr

Objectives

The **É:Calm** resource is made up with French scholar and student manuscripts produced in a variety of usual contexts of teaching. The specificity of the É:CALM resource is to provide an ecological digital dataset that gives a broad overview of texts written at school, highschool and university [1].

Handwritten primary sources

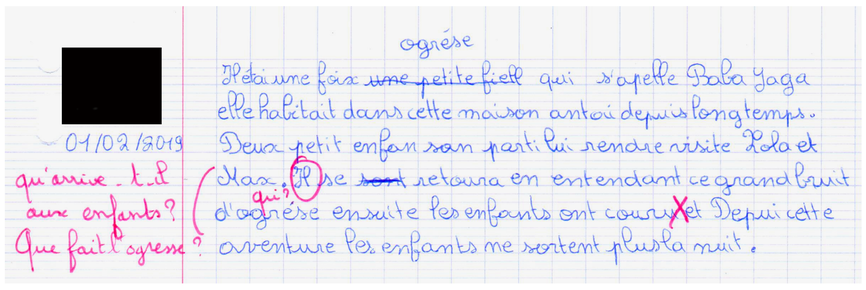

Figure 1: Example of primary source collected from 4th grade pupils. Comments in pink were written by the teacher. The original copy has been scanned, cropped and de-identified

The É:Calm data processing

1. Collecting real data produced by pupils and students as part of usual writing activities or in reply to instructions designed by the researchers but always supervised by the regular teacher
2. Collecting metadata about the classroom and the class work (`<settingDesc>`)
3. Scanning, cropping and de-identifying texts
4. Digitalizing text manually and encoding into XML format acc. to the TEI-P5 norm the main graphical aspects such as layout (`<p>`, `<lb>`) and revisions i.e. deleted and added text portions (`<mod>`) and unreadable or unclear segments (`<gap>`, `<unclear>`)
5. Checking the transcription and the XML encoding
6. Manual spell checking via misspelling annotation (misspelled/spelled word alignment)
7. Checking the spell checking
8. POS-tagging and automatic parsing (Talismane[2])
9. Giving access to the data for research and teaching

Partnership (ANR project)

The **É:Calm** resource results from the pooling of pre-existing corpora developed by 3 French research groups in Linguistics with the help of one in Educational Studies

CLESTHIA (Paris) – the **EcriScol** corpus made up with elementary school texts presenting successive versions including teacher comments

CLLE (Toulouse) – the **ResolCo** corpus made up with texts representing all the educational levels (from primary school to university) and written acc. to a common instruction

Lidilem (Grenoble) – the **ScolEdit** corpus giving access to texts written by individuals throughout their elementary grades; the **Littéracie Avancée** corpus made up with texts written by students (Higher education)

Circeft (Créteil), research in Educational Studies

```
<?xml version='1.0' encoding='UTF-8'?>
<TEI>
<teiHeader>
    <fileDesc>
    <encodingDesc>
    <profileDesc>
      <langUsage>
        <settingDesc>
         <setting>
            <name type="city">Sevran (93270)</name>
            <name type="region">Île-de-France</name>
            <date>2019-02-01</date>
            <locale>École élémentaire Emile Zola</locale>
            <activity>Rédaction</activity>
         </setting>
        </settingDesc>
    </profileDesc>
</teiHeader>
<text>
<front>
  <p>Nom de l'élève</p>
  <p>01/02/2019</p>
</front>
<body>
  <p>
      Il étai une foix <mod type="subst"><del>une petite fiell</del><add>
      ogrése</add></mod> qui s'apelle BabaYaga <lb/>elle habitait dans cette
      maison antai depuis longtemps. <lb/>Deux petit enfan son parti lui
      rendre visite Lola et <lb/>Max. Il se <mod type="subst"><del>sont
      </del></mod> retoura en entendant ce grand bruit <lb/>d'ogrése ensuite
      les enfants ont couru et Depuis cette <lb/>aventure les enfants ne
      sortent plus la nuit.
  </p>
</body>
<back>
    <metamark who="teacher">Qu'arrive-t-il aux enfants ?</metamark>
    <metamark who="teacher">Que fait l'ogresse ?</metamark>
    <metamark who="transcriber">The teacher inserted comments in the text.
    Those comments have not been transcribed.</metamark>
</back>
</text>
</TEI>
```

Available data

Table 1: Quantitative overview of the current version of the E:Calm resource (*estimated nb. of words)

Education level	#texts	#words*
Elem. School	2375	656010
Middle School	1077	958500
High School	86	129000
University	648	2599250

Gold Standard Dependency SubCorpus: manual checking of the POStagging and Parsing made by the Talismane toolkit [2]:

- 68 texts, 11 706 token out of punctuation
- 5 education levels (grades 3, 4, 6, 9 and Master Degree)
- 2 coders: Cohen's ***k***appa score before adjudication
  - $k = 0.45$ for POStagging (i.e. wrong POStag Y/N)
  - $k = 0.28$ for Parsing (i.e. wrong governor Y/N)

Time consuming: 1:30' per text (30' for XML encoding, 25' for XML checking, 35' for misspelling annotation) + 35' for checking the Talismane output

First Results

Revisions and Misspellings Annotation

Table 2: Number of annotated revisions (**mod**) and misspellings (**err**)

grade	#mod	#texts	mod/text	#err	#texts	err/text
	23587	3034	8	24256	1521	16
1 (CP)	280	373	1		na	
2 (CE1)	2651	604	4	3211	231	14
3 (CE2)	5011	564	9	4381	168	26
4 (CM1)	1703	208	8	745	68	11
5 (CM2)	9008	626	14	9145	517	18
6 (6e)	1104	154	7	3333	98	34
8 (4e)	204	47	4		na	
9 (3e)	1075	103	10	1324	84	16
10 (2de)	1681	67	25	1308	67	20

- At grade 6, a strange misspelling spike probably related to text length

POStagging and Parsing Evaluation

Table 3: Talismane accuracy for POStagging and Parsing i.e. nb. of correct POStags, syntactic attachment (UAS – unlabelled attachment score) on the nb. of tokens in the Gold Standard; and nb. of correct labels (LAS – labelled attachment score) on the nb. of correctly attached tokens

	#tokens	accuracy
POS	11 706	96.2
UAS	11 706	97.5
LAS	11 262	90.7

- cf. Tran, T.M.N (2019) *Évaluation de l'analyseur syntaxique Talismane et évolution des phrases complexes du corpus RESOLCO* Master Dissertation, Univ. Bordeaux Montaigne.

Next steps: corpus annotation and analyses

- Co-reference and discourse relations management through the successive education grades
- Misspellings typology and analysis through the successive education grades with a focus on verbs
- Teacher comments typology
- Correlations between these 3 annotation levels

Acknowledgments, References

This huge amount of work could not have been done without the labor performing by the Linguistics students involved in the project: Mathilde Lala (initiation of the whole process), Thi Mai Nhi Tran and Anastasiia Larionova (POStagging and Parsing evaluation), Alexis Robert and Carla Sarraud (XML encoding and Gold constitution), Jade Moillic and Astrid Chemin (XML encoding and spell checking), Andréane Roques (XML encoding)

[1] C. Doquet, J. David, and S. F. (Eds).
Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement.
In Corpus [Online], volume 16 (Special Issue). OpenEdition, 2017.

[2] A. Urieli.
Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit.
PhD thesis, University of Toulouse Jean Jaurès, France, 2013.

E-Calm Écriture scolaire et universitaire : Corpus, Analyses Linguistiques, Modélisations didactiques

AGENCE NATIONALE DE LA RECHERCHE

PROJET FINANCÉ PAR L'ANR – PROJECT FUNDED BY THE ANR