É:CALM Resource: a Resource for Studying Texts Produced by French Pupils and Students

Lydia-Mai Ho-Dac, Serge Fleury, Claude Ponton
CLLE – CNRS – University Toulouse Jean Jaurès
CLESTHIA – University of Paris 3 Sorbonne Nouvelle
LIDILEM – University of Grenoble Alpes
lydia-mai.ho-dac@univ-tlse2.fr, serge.fleury@sorbonne-nouvelle.fr, claude.ponton@univ-grenoble-alpes.fr

Abstract

The É:CALM resource is constructed from French student texts produced in a variety of usual contexts of teaching. The distinction of the É:CALM resource is to provide an ecological data set that gives a broad overview of texts written at elementary school, high school and university. This paper describes the whole data processing: encoding of the main graphical aspects of the handwritten primary sources according to the TEI-P5 norm; spelling standardizing; POS tagging and syntactic parsing evaluation.

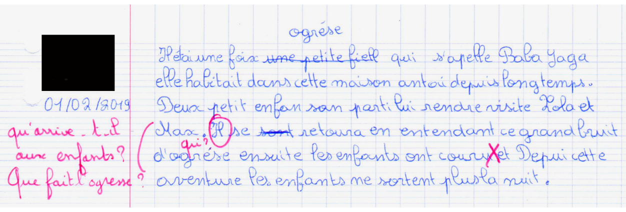
Keywords: education, handwritten encoding, spelling

1. Introduction

The É:CALM resource is constructed from French student texts produced in a variety of usual contexts of teaching. The key feature of the É:CALM resource is to provide an ecological digital data set that gives a broad overview of texts written at elementary school, high school and university (Doquet et al., 2017b).

The advantages of such a resource are multidisciplinary. From a scientific point of view, the É:CALM resource provides a valuable data set for the digital humanities and writing sciences, since it allows the unparalleled possibility to observe the acquisition of literacy and especially writing skills through all the education levels. From an education point of view, it could be used for teaching literacy by working with students on real-life texts and by focusing on attested misspellings and coherence issues. In addition, it could be used for identifying the main problems and doubts encountered by students at each grade. As for NLP and corpus linguistics, such a resource constitutes a good experimental field for evaluating and adapting models, methods and tools on the (manual or automatic) annotation of non-standard corpora. The decision to encode handwritten student texts according to the TEI-P5 norm ensures consistency, practicability, compatibility with a broad range of corpus tools and data exchange facilities (Burnard, 2007).

Figure 1 gives an example of a primary source that composed the É:CALM resource.



[redacted]

01/02/2019

qu'arrive-t-il aux enfants?
Que fait l'ogresse?

ogrése
Il étai une foix ~~une petite fiell~~ qui s'apelle Baba Yaga
elle habitait dans cette maison antoi depuis longtemps.
Deux petit enfan san parti lui rendre visite Lola et
Max. Il se ~~sont~~ retoura en entendant ce grand bruit
qui?
d'ogrése ensuite les enfants ont couru et Depui cette
oventure les enfants ne sortent plus la nuit.

Figure 1: Example of primary source collected from 4th grade pupils.

The main part of the resource is made up with such handwritten school works. As the figure shows, the collected manuscripts may contain student revisions (e.g. erasures) and teacher comments (here in red).

This paper describes the whole data processing established for digitizing and standardizing such manuscripts in order to provide a new resource for NLP, Corpus Linguistics and Education.

2. Related Works

The Lancaster Corpus of Children's Project Writing (Smith et al., 1998) is one of the first children's corpora transcribed and available online. It is the first in the field of school corpora, bringing together a large number of texts written by a group of students followed during three years. A decade later, in 2006, the Oxford Children's Corpus (Banerji et al., 2013) proposes more than 70,000 short texts written by English-speaking children aged 4-13 as part of public online writing competitions. In 2011, the University of Karlsruhe collected and digitized German-language spontaneously written texts from Grades 1-8 (Lavalley et al., 2015). The first French-language children corpus appeared in 2005 and includes 500 written texts from Grades 5-7 (Elalouf, 2005). Since 2010, various French-language children corpora projects have been launched (Garcia-Debanc and Bonnemaison, 2014; Doquet et al., 2017a; Boré and Elalouf, 2017; De Vogüé et al., 2017; Wolfarth et al., 2017). The É:CALM resource takes advantage of all these project in order to provide the broadest French-language children corpus. It will include more than 6,700 texts with a longitudinal coverage of writings from primary school to university and a wide variety of genres. It constitutes the first resource encoded in XML TEI-P5 and designing for corpus linguistics and corpus annotation.

3. Data Collection

All data composing the É:CALM resource are ecological written data produced by students in their daily school life under the supervision of their regular teacher.

The É:CALM resource takes part from the pooling of 4 pre-existing data sets collected according to different protocols.

The most relevant differences depend on (a) whether the schoolwork has been written as part of the usual activities or in reply to a dedicated instruction designed by the researchers; (b) the grade levels taking into account; and (c) whether there are teacher comments or not.

On the one hand, the *EcriScol*[1] (Doquet et al., 2017a) and the *Advanced Literacy*[2] (Jacques and Rinck, 2017) resources are made up with texts written by students at school for the first or at university for the latter. All these data were collected without predefined instructions. A large part of these texts contains teacher comments and the *EcriScol* resource also includes drafts and intermediate versions. The *Advanced Literacy* corpus is the only one composed with typewritten texts.

On the other hand, the *ResolCo*[3] (Garcia-Debanc et al., 2017) and the *Scoledit*[4] (Wolfarth et al., 2017) corpora are made up with texts produced in reply to a specific instruction.

The *ResolCo* resource is characterized by an instruction which has been designed for causing strategies in terms of discourse coherence and confronting the writer to cohesion problems such as anaphora, encapsulation, sequence of tenses, generic vs. specific mood (Garcia-Debanc, 2016). The *ResolCo* instruction consists in asking the students to write a narrative by inserting three predefined sentences in it. Each sentence contains, amongst others, anaphora and a specific tense.

The *Scoledit* resource constitutes an unparalleled longitudinal corpus of texts i.e. a corpus giving access to texts written by individuals throughout all their elementary grades. 373 pupils were included in the study. One narrative and several dictations per student per year have been collected between 2014 and 2018. This corpus allows studies focusing on the individual evolution of language skills during elementary school (Wolfarth et al., 2018).

These four Data Collections cover almost all the education levels, from the 1st grade to the master degree. Table 3. gives a quantitative overview of the current version of the É:CALM resource. The number of texts per educational level is approximately equivalent to the number of students per grade.

Education level	#texts	#words	Data Collection
Elem. School	2,375	656,010	[E][R][S]
Middle School	1,077	958,500	[E][R]
High School	86	129,000	[E]
University	789	2,575,000	[R][AL]

Table 1: Quantitative overview of the current version of the É:CALM resource with approximate number of words, with [AL] for the *Advanced Literacy* corpus, [E] for *Ecriscol*, [R] for *Resolco* and [S] for *Scoledit*

4. Data Digitization and Standardization

Once the handwritten primary sources are collected, a data processing starts in order to (1) encode each text into XML format with respect to the TEI-P5 norm and (2) proposed an aligned spelling standardized version. This meticulous data processing follows 6 successive steps:

1. Scanning, cropping and de-identifying texts (cf. Figure 1)
2. Encoding the metadata in the teiHeader
3. Digitizing text manually and encoding into XML format according to the TEI-P5 norm
4. Checking the transcription and the XML encoding
5. Manual spell checking via misspelling annotation (misspelled/spelled word alignment)
6. Checking the spell checking

This data processing takes around 3 hours per text: 30 min. for a first XML encoding, 25 min. for XML encoding checking, 30 min. for XML encoding finalization, 35 min. for misspelling annotation, 30 min. for misspelling annotation checking, 20 min. for misspelling annotation finalization.

4.1. Encoding MetaData according to TEI-P5

Each text is systematically associated with metadata about the classroom and the class work. Table 2 gives an overview of main metadata available for a large part of texts composing the resource (more details about these elements are given in the TEI-P5 guidelines[5]).

TEI-P5 tag	Description
`settingDesc`	Region and social characteristics of the educational institution (e.g. Priority Education Zones, rural vs. urban population)
`textDesc`	Information about the instruction given to the students and about preparedness and derivation i.e. if the text is a draft, a prepared work or a revised one
`particDesc`	- Students characteristics among which age, mother tongue, language disorder (e.g. dyslexia, apraxia) and teacher assessment - Teacher characteristics e.g. years of teaching

Table 2: Metadata available in the É:CALM resource

4.2. Transcription according to TEI-P5

The TEI-P5 is the most appropriate norm for encoding the body of the collected manuscripts. On the first hand, it favours the sharing between multi-disciplinary approaches and the perpetuation of the data set. On the other hand, handwritten manuscripts encoding was already fairly well described in the TEI-P5 guidelines, especially for the revisions[6].

[1] http://www.univ-paris3.fr/ecriscol

[2] "Littératie Avancées" in French, https://www.ortolang.fr/market/corpora/litteracieavancee

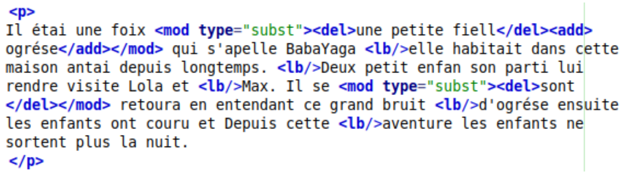
[3] http://redac.univ-tlse2.fr/corpus/resolco.html

[4] http://www.scoledit.org/scoledit

[5] https://tei-c.org/Guidelines/P5/

[6] See the chapter 10 Manuscript Description

The transcription process is totally manual with the help of a visualization via xslt transformation for checking. Figure 2 illustrates the result of the TEI-P5 encoding.



```
<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>
```

Figure 2: Extract of the TEI-P5 XML file relative to the primary source given in Figure 1

As illustrated in Figure 2, the `<mod>` TEI-P5 element is used for representing any kind of revision. Three kind of revisions are nowadays encoded: deletion, addition and substitution (e.g. simultaneous deletion and addition). Table 3 lists all the TEI-P5 elements used for encoding the graphical aspects that occur in the manuscripts.

TEI-P5 tag	Description
mod	Revision (containing a deletion and/or an addition) that may be associated with a participant (student or teacher)
del	Deleted text portion
add	Added text portion
gap	Unreadable characters
unclear	Text portions where the coder is not sure of his/her transcription
p	Paragraph (intentional line break)
lb	Line break (because of the margin)
pb	Page break
metamark	Global notes about the transcription and global comments written in the margins by the teacher as in Figure 1.

Table 3: Graphical aspects encoded in the É:CALM resource

4.3. Spell checking

The next step concerns the spell checking of the primary sources. As for the encoding, this step is totally manual with the help of the annotation tool GLOZZ[7] (Mathet and Widlöcher, 2009). The reasons for a manual spell checking is twofold: first, the extreme non-standard spelling in quite a lot of texts; and second, the necessity of having a very accurate error detection for further spelling analyses. The annotation tool GLOZZ was chosen for allowing multi-layer annotation: revisions, spelling errors and further annotations such as coreference and discourse relations (Asher et al., 2017).

Manual spell checking consists in delimiting all the misspelling text segments and indicating for each annotated unit the correct spelling. When more than one spelling is possible, several suggestions may be indicated with a ranking from the most to the less obvious, taking the meaning of the whole text. Examples of such multi-spelling occur when two verb tenses are probable or when there is no cue for choosing between correcting the number/gender of the subject or of the verb.

The spell checking mainly concerns spelling and morphology. Punctuation may also be annotated but only in two cases: when a final punctuation occurs without capitalization in the next word and vice-versa; and when there is a lack of comma between items in a list. No errors are annotated in case of problematic sequence of tenses.

Each misspelling unit is also associated with a feature indicating the certainty degree of the coder about the unit delimitation and the spelling suggestion, from totally sure to strongly unsure.

Once the misspellings are annotated, a standardized version of the text is automatically generated and checked by another coder with the help of automatic spellchecking. Data are now ready for applying usual Natural Language Processing Tools.

5. Data POS tagging and Syntactic Parsing

The first NLP tool used on the standardized data is the Talismane toolkit (Urieli, 2013) for POS tagging and syntactic parsing. Because the data remain non standards, even after spell checking, with for example very long sentences and some syntactic peculiarities, an evaluation of the POS tagging and Parsing accuracy was done. A Gold Standard Data Set (henceforth GSDS) was built containing 68 texts, 11,706 token (out of punctuation) and covering 5 education levels (grades 3rd, 4th, 6th, 9th and Master Degree).

Two coders validated the output provided by the best configuration of the Talismane toolkit (Urieli and Tanguy, 2013) by using the Brat annotation tool (Stenetorp et al., 2012) and the guidelines put in place for the French Tree Bank (Candito et al., 2009).

The Cohen's ***k***appa scores are fairly bad: ***k*** = 0.45 for POS tagging (i.e. wrong POStag Y/N) and ***k*** = 0.28 for Parsing (i.e. wrong governor Y/N).

These bad inter-annotator agreements entailed a long period during which the two coders adjudicated for finalizing the GSDS.

5.1. POS tagging and Parsing Evaluation

Fortunately, the scores obtained by Talismane on the GSDS are fairly good. Table 4 gives the number of correct POS tags, syntactic dependencies on the number of tokens in the GSDS (UAS – unlabelled attachment score); and number of correct labels on the number of correctly attached tokens (LAS – labelled attachment score). As it shows, the lowest accuracy concerns the LAS.

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

Table 4: Talismane global accuracy

Table 5 gives the precision and recall scores for each POS occurring more than 20 times in the GSDS. The lowest scores are observed on the usually problematic POS such

[7] http://glozz.free.fr/

as the confusion between Adjectives ($R = 0.88$) and the Past Participles ($P = 0.75$); and the Subordinating Conjunctions ($P = 0.73$ and $R = 0.82$).

POS	#tokens	P	R
Adjective	675	0.94	0.88
Adverb	837	0.92	0.93
Coordinating Conjunction	388	0.99	0.94
Clitic (object)	222	0.99	0.98
Clitic (reflexive)	248	0.97	1.00
Clitic (subject)	739	1.00	0.99
Subordinating Conjunction	156	0.73	0.82
Determiner	1818	0.99	0.99
Common Noun	2241	0.96	0.98
Proper Noun	294	0.95	0.96
Preposition	1311	0.96	0.99
Prep. + Det. (e.g. *du*)	119	0.84	1.00
Prep. + Pro. (e.g. *duquel*)	1514	1.00	1.00
Pronoun	170	0.95	0.92
Relative Pronoun	125	0.90	0.98
Indicative Verb	1522	0.98	0.99
Infinitive Verb	323	0.99	0.99
Past Participle	223	0.75	0.98
Present Participle	115	0.98	0.96

Table 5: POS tagging precision (P) and recall (R)

As for the syntactic parsing, the lowest scores concern the labeled attachment score (LAS) and especially the distinction between direct object and adjunct (cf. Table 6).

Verb Dependency	#occ.	P
Subject	1306	0.94
Direct object (of Verbs and Preposition)	824	0.80
Indirect object of Verbs	20	0.87
Adjunct	2663	0.79
Predicative adjective	73	0.85

Table 6: Verb dependencies precision (P)

According to these results, the POS tagging and Parsing processed by the Talismane toolkit are good enough for providing a consistent É:CALM Tree Bank.

6. First Analyses

Even if the resource is not yet complete, first analyses have been conducted for describing the evolution of the writing skills. The next sections provide the first results of preliminary studies that show the wealth of the É:CALM resource.

6.1. Do individuals write longer passages through the successive education grades ?

As mentioned above, the *Scoledit* protocol let us to follow the evolution of writing skills of individuals throughout their elementary grade according to a common instruction (i.e. tell the story of one or two fictional characters: a robot, a cat, a wolf and/or a witch). This data set contains 1,865 texts and 140,878 words. As Figure 3 shows, the average text size increases significantly through educational levels.

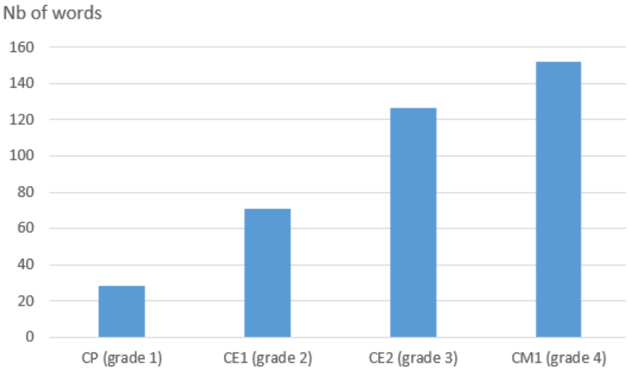


Figure 3: Evolution of text size throughout individuals' elementary grade in the *Scoledit* subpart

6.2. Revisions through the successive education grades

Revisions taken into account here consist in deletions, additions and substitutions made by the student during the writing process before any teacher comments. Their encoding provides insight into the students doubts and inquiries in contrast with student corrections made in response to marks on the page by a teacher. Table 7 gives the number of annotated revisions through the successive education grades in the currently encoded manuscripts[8]. The grades when pupils produce the most of revisions are the 4h, 5th and 9th grades.

grade	#mod	#texts	mod/text
all	23587	3034	8
1st (CP)	280	373	1
2nd (CE1)	2651	604	4
3rd (CE2)	5011	564	9
4th (CM1)	1703	208	8
5th (CM2)	9008	626	14
6th (6e)	1104	154	7
8th (4e)	204	47	4
9th (3e)	1075	103	10

Table 7: Number of instinctive revisions (`mod`) including deletions, insertions and substitutions

It is currently difficult to interpret these results without a further study that will inform us about the POS and the syntactic role of the text segments concerned with the revisions.

6.3. Misspellings through the successive education grades

Misspellings annotation permits to highlight the spelling issues that remain unsolved at each grade. Table 8 gives the

[8] Type-written texts from the *Advanced Literacy* part are not taking into account here.

number of annotated misspellings through the successive education grades in the *ResolCo* subpart.

grade	#texts	#tokens	% err/token
3rd (CE2)	31	3252	12.8
4th (CM1)	37	4823	13.7
5th (CM2)	42	8351	12.8
6th (6e)	45	7860	13.1
8th (4e)	15	4887	12.6
9th (3e)	27	8622	9.2
Master	12	5318	2.1

Table 8: Number of misspellings (`err`) in the *ResolCo* subpart, #tokens excludes punctuation.

Fortunately, the proportion of tokens with spelling error decreases with the grade. When looking at the error rate for each POS (Figure 4), it appears that Past Participles remain problematic, even at the Master degree with a top average of 57.5% of misspelled token at the 6th grade and still 8.2% at the Master degree.

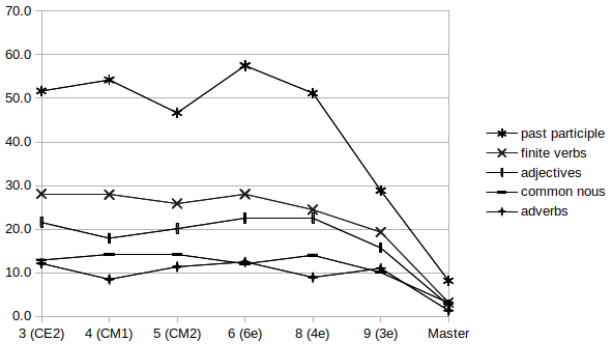


Figure 4: Proportion of misspellings per POS in the *ResolCo* subpart.

Past Participle (PP) misspelling is very frequent in French because PPs must be inflected to show gender and number but also and above all because a large part of PPs end with the same phoneme than Infinitives (e.g. [e]) but with a different grapheme: *é(e)(s)* for PPs and *er* for Infinitives. Example (1) gives an extract of a texts from the 6th grade where PP's ending misspellings are underlined with the correct spelling in brackets.

(1) *Pol et Marina sont éfréiller [effray**és**], ils rentrent chez Marina en courant, ils ce[se] sont cacher [cach**és**] dans la chambre*

As for Adjectives that require gender and number agreement in French and finite verbs that show a quite complex morphology with usually more than 20 inflections per verb, their spelling remains problematic at all elementary and high school grades but seems overcome at the Master degree.

7. Conclusion

This paper presents the É:CALM resource composed with French student texts produced at school and at university. The whole data process is fairly long and requires a precise and careful work for encoding the main graphical aspects of the handwritten primary sources and annotating the misspellings. The evaluation of the Talismane analyses shows that we could be confident in NLP tools for POS tagging and syntactic parsing, even with the non standard syntactic structures and punctuation usages that occur often in young student texts.

Once all the data processed, the resource will be made available for the community[9]. A large part of each subpart is already available but not in a standardized, homogeneous and structured data set. Meanwhile, several studies have started to exploit the É:CALM resource even in progress. A first group of studies focuses on revisions, misspellings and teacher comments categorization. A second one aims at annotating coherence in the spell checked texts, in order to handle the discourse organization acquisition through grade levels.

8. Acknowledgements

This huge amount of work could not have been done without the labor performed by the Linguistics students involved in the project: Giuditta Bucciaglia, Astrid Chemin, Silvia Federzoni, Mathilde Lala, Sara Mazziotti, Jade Moillic, Arnaud Moysan, Thi Mai Nhi Tran, Anastasiia Larionova, Alexis Robert, Andréane Roques, Laura Roudaut, Carla Sarraud, Claire Wolfarth.

9. Bibliographical References

Asher, N., Muller, P., Bras, M., Ho-Dac, L. M., Benamara, F., Afantenos, S., and Vieu, L. (2017). Annodis and related projects: Case studies on the annotation of discourse structure. In Nancy Ide et al., editors, Handbook of Linguistic Annotation, pages 1241–1264. Springer Netherlands, Dordrecht.

Banerji, N., Gupta, V., Kilgarriff, A., and Tugwell, D. (2013). Oxford children's corpus: A corpus of children's writing, reading and education. Corpus Linguistics, pages 315–317.

Boré, C. and Elalouf, M.-L. (2017). Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. Corpus, 16:31–64.

Burnard, L. (2007). New tricks from an old dog: An overview of tei p5. In Lou Burnard, et al., editors, Digital Historical Corpora- Architecture, Annotation, and Retrieval, number 06491 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

Candito, M., Crabbé, B., and Falco, M. (2009). Dépendances syntaxiques de surface pour le français. Rapport technique, Université Paris, 7.

De Vogüé, S., Espinoza, N., Garcia, B., Perini, M., and Marzena Watorek, F. (2017). Constitution d'un grand corpus d'écrits émergents et novices : Principes et méthodes. Corpus, 16:65–86.

Doquet, C., Enoiu, V., Fleury, S., and Maziotti, S. (2017a). Problèmes posés par la transcription et l'annotation d'écrits d'élèves. Corpus, 16:133–156.

[9] `http://e-calm.huma-num.fr/`

Doquet, C., David, J., and Fleury, S. (2017b). 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.

Elalouf, M.-L. (2005). Ecrire entre 10 et 14 ans : Un corpus, des analyses, des repères pour la formation. Canopé - CRDP de Versailles.

Garcia-Debanc, C. and Bonnemaison, K. (2014). La gestion de la cohèsion textuelle par des élèves de 11-12 ans : Réussites et difficultés. In Actes du 4e Congrès Mondial de Linguistique Française, pages 961–976.

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., and Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. Corpus, 16.

Garcia-Debanc, C. (2016). Une tâche problème pour analyser les compétences d'élèves de sixième en matière de cohésion textuelle. In Sarda L., et al., editors, Connexion et indexation. Ces liens qui tissent le texte, pages 263–278. ENS Editions, June.

Jacques, M.-P. and Rinck, F. (2017). Un corpus de littéracie avancée : résultat et point de départ. Corpus, 16.

Lavalley, R., Berkling, K., and Stücker, S. (2015). Preparing children's writing database for automated processing. In Proceedings of LTLT@SLaTE, pages 9–15.

Mathet, Y. and Widlöcher, A. (2009). La plate–forme GLOZZ : environnement d'annotation et d'exploration de corpus. In Actes de TALN 2009, Senlis, June. ATALA, LIPN.

Smith, N., McEnery, T., and Ivanic, R. (1998). Issues in transcribing a corpus of children's handwritten projets. Literacy and Linguistic Computing, 13:217–225.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 102–107. Association for Computational Linguistics.

Urieli, A. and Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013), pages 188–201, Les Sables d'Olonne, France.

Urieli, A. (2013). Analyse syntaxique robuste du français : concilier méthodes syntaxiques et connaissances linguistiques dans l'outil Talismane. Ph.D. thesis, Université de Toulouse - Jean Jaurès.

Wolfarth, C., Ponton, C., and Totereau, C. (2017). Apports du tal à la constitution et à l'exploitation d'un corpus scolaire. Corpus, 16.

Wolfarth, C., Ponton, C., and Brissaud, C. (2018). Gestion de la morphologie verbale en production d'écrits : que peut nous apprendre un corpus longitudinal ? Repères, 57.