

Premières explorations textométriques d'un corpus scolaire longitudinal (CP-CM1)

Claude Ponton, Claire Wolfarth et Catherine Brissaud

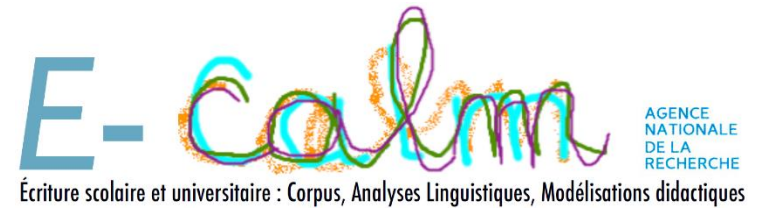
JLC2019, Grenoble, 28 nov. 2019

Plan

- Présentation du corpus Scoledit
- Exploration textométrique
 - selon le niveau, le genre et la composition sociale de la classe
- Perspectives

Le projet E-Calm

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

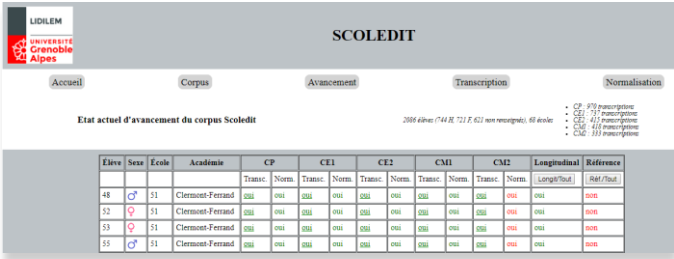


- caractériser ces écrits et les attentes des enseignants du point de vue de l'acquisition de l'orthographe et de la cohérence, dans des analyses sociologiquement contextualisées ;
- étudier les modalités d'écriture dans les avant-textes (plans, notes, brouillons) et les textes, notamment à travers l'influence réciproque des écrits remis et des interventions des enseignants sur les copies.
- structurer et mettre à disposition de la communauté scientifique un vaste corpus d'écrits d'élèves et d'étudiants permettant des analyses quantitatives et des traitements automatiques ;



Le projet Scoledit

- Objectifs
 - décrire les compétences en production d'écrits et leurs évolutions au primaire
 - proposer des pistes didactiques
- Moyens
 - Constituer un large corpus longitudinal de productions
 - Outiller l'analyse du corpus



The screenshot shows the Scoledit website interface. At the top, there is a header with the LIDILEM logo and the University of Grenoble Alpes. Below the header, there are navigation tabs: Accueil, Corpus, Avancement, Transcription, and Normalisation. The main content area displays the title "Etat actuel d'avancement du corpus Scoledit" and a table of student progress data. The table has columns for Élève, Sexe, Ecole, Académie, CP, CE1, CE2, CM1, CM2, Longitudinal, and Référence. The data rows show progress for four students (48, 52, 53, 55) across different levels (CP, CE1, CE2, CM1, CM2) and longitudinal data.

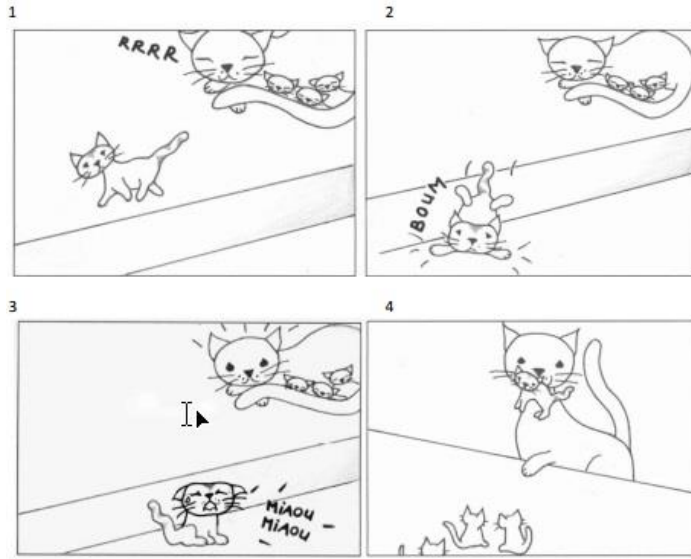
Élève	Sexe	Ecole	Académie	CP		CE1		CE2		CM1		CM2		Longitudinal	Référence
				Transc.	Norm.	Transc.	Norm.	Transc.	Norm.	Transc.	Norm.	Transc.	Norm.		
48	♂	S1	Clermont-Ferrand	000	000	000	000	000	000	000	000	000	000	000	000
52	♀	S1	Clermont-Ferrand	000	000	000	000	000	000	000	000	000	000	000	000
53	♀	S1	Clermont-Ferrand	000	000	000	000	000	000	000	000	000	000	000	000
55	♂	S1	Clermont-Ferrand	000	000	000	000	000	000	000	000	000	000	000	000

Le corpus Scoledit

- Un corpus longitudinal
- Recueil de productions écrites et de dictées
- Prolongement de la recherche « Lire-Ecrire au CP » (Goigoux)
- Suivi des mêmes élèves au primaire : 2014 (CP)-2018 (CM2)
- Consignes
 - spécifique au CP (recherche Lire-Ecrire)
 - même consigne CE1-CM2

Le corpus Scoledit

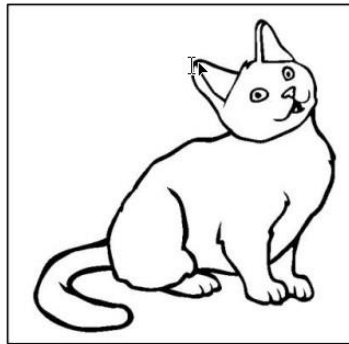
CP



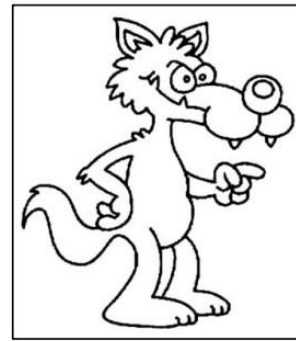
CE1-CM2



1



2

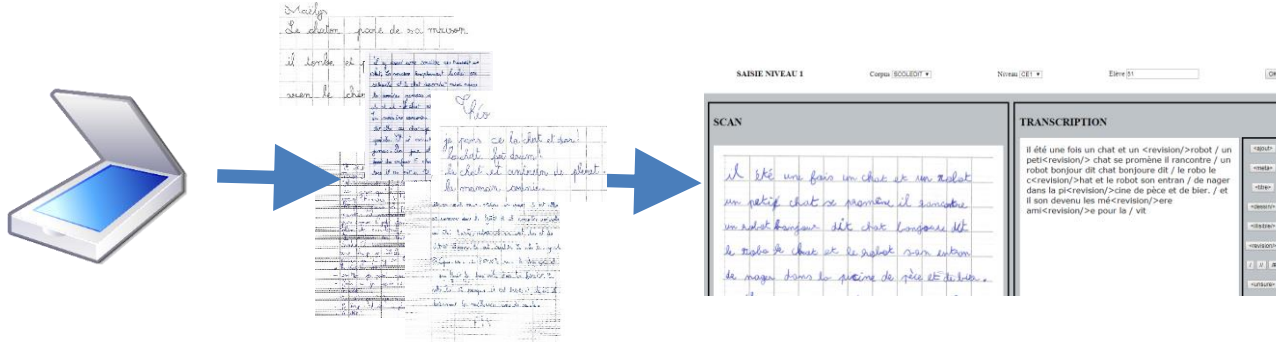


3



4

Le corpus Scoledit



Numérisation
Stockage, sauvegarde

Transcription
Manuelle
Guide + Outil + révision

La sorcière cherche
sons chat / dans toute
sa maison mes / trouve
pas sons chat. /# Alors
elle se mes a pleurer /
a <revision/>
<revision/> pleurer
qu'elle / tonbe par
terre. Tout a coup /
Elle entant la porte
souvrire / Elle dit «
c'est mon chat. » /
Sons chat sote dans
c'est / <revision/>
bras est il fut eurreu
/ jusca la fin des
<revision/>tans.

Un site de travail et de visualisation : <http://scoledit.org/scoledition>

The screenshot shows the SCOLEDIT website interface. At the top, there is a header with the LIDILEM logo and the text 'SCOLEDIT'. Below the header, there are navigation tabs: 'Accueil', 'Corpus', 'Avancement', 'Transcription', and 'Normalisation'. The 'Transcription' tab is selected. Below the tabs, there is a navigation bar with the following information: 'Corpus SCOLEDIT', 'Niveau CE2', 'Année 2016', 'Elève 398', 'Sexe homme', and 'École'. The main content area is divided into two panels: 'SCAN' and 'PRODUCTION'. The 'SCAN' panel shows a scanned image of a handwritten document. The 'PRODUCTION' panel shows the transcription of the document, with the text: 'Il était une fois un chat qui habitait une petite maison [x]¹ qui était un méchant sorcier. Il transformait les gens [x]² les animaux sur son passage. Mais son jeu préféré était de [x] fabriquer des potions. Il prenait humains et animaux pour les couper en rondelle ou quand il avait la fièvre de [x] prendre des crapauds ou des chauve-souris dans [x] les bois, il les transformait. Comme ça il pouvait faire les notions plus rapidement.'

Outiller l'analyse

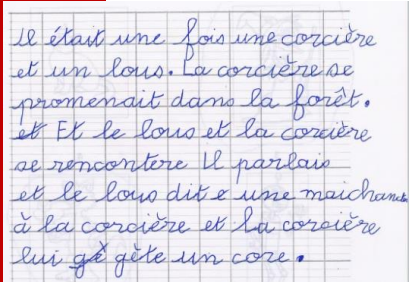
- Objectifs
 - exploitation du corpus
 - aide à la description linguistique
- Outils TAL non adaptés à ce type d'écrits
 - écarts à la norme
 - nécessité d'une version « normalisée »
- Approche par comparaison (Kraif & Ponton 2007)
 - Alignement production / normalisation
 - Comparaison à différents niveaux
 - catégorisation des erreurs : TALC 2018
 - morphologie verbale : *Repères 57*
 - segmentation : AIRDF 2019

Le corpus longitudinal

- Difficultés de suivi
 - absences, redoublements...
 - déplacement recueil
- Au total : 373 enfants suivis du CP au CM2
 - Données CM2 non prêtes
 - Une version transcrite + une normalisée
 - Métadonnées élèves et écoles
 - Format XML-TEI (projet E-Calm)

Extrait (CE1 1319)

Scan



Il était une fois une corcière
et un lous. La corcière se
promenait dans la forêt,
Et le lous et la corcière
se rencontère Il parlai
et le lous dit e une maichanct
à la corcière et la corcière
lui gè gète un core.

Transcription

Il était une fois une corcière
et un lous. La corcière se
promenait dans la forêt.
[x] Et le lous et la corcière
se rencontère Il parlais
et le loup dit une maichancter
à la corcière et la corcière
lui [x] gète un core.

```
<setting>
  <activity>Production</activity>
  <!-- Métadonnées PRODUCTION -->
  <!-- année de la production -->
  <date>2015</date>
  <!-- CP, CE1, CE2, CM1, CM2... à définir -->
  <name type="niveau">CE1</name>
  <!-- Métadonnées ECOLE -->
  <name type="ville">Non disponible</name>
  <name type="académie">Lyon</name>
  <name type="région">AURA</name>
  <!-- Métadonnées ELEVE -->
  <name type="élève">1319</name>
  <!-- homme : 1, femme : 2 -->
  <name type="sexe">1</name>
  <!-- langue parlée à la maison 1=français, 2=autre, 3=français+autre-->
  <name type="langue">1</name>
  <!-- Redoublement en CP (1) ou pas (0) -->
  <name type="RCP">0</name>
  <!-- Date de naissance jj/mm/aaaa (seulement mm/aaaa pour Scoledit) -->
  <name type="naissance">12/07</name>
  <!-- CSP parents -->
  <name type="csp_pere">3 - Professions libérales et assimilés ...</name>
  <name type="csp_mere">5 - Employés de la fonction publique ...</name>
```

Métadonnées

Normalisation

```
<text>
  <body>
    <head>NORM-EC-CE1-2015-14-D1-S1319-V1</head>
    <p>Il était une fois une sorcière et un loup. La sorcière se promenait dans la forêt. Et le loup et la sorcière se rencontrèrent. Ils parlaient et le loup dit une méchanceté à la sorcière et la sorcière lui jette un sort.</p>
  </body>
</text>
```

Métadonnées

- Niveau
- Genre
 - 159 garçons, 214 filles
- Composition sociale des écoles

Compo. sociale	EP	DF	MX	FV
Nombre d'élèves	46	70	225	32

Compo. sociale	Moins	Plus
Nombre d'élèves	116	257

Moins : DF + EP

Plus : MX + FV

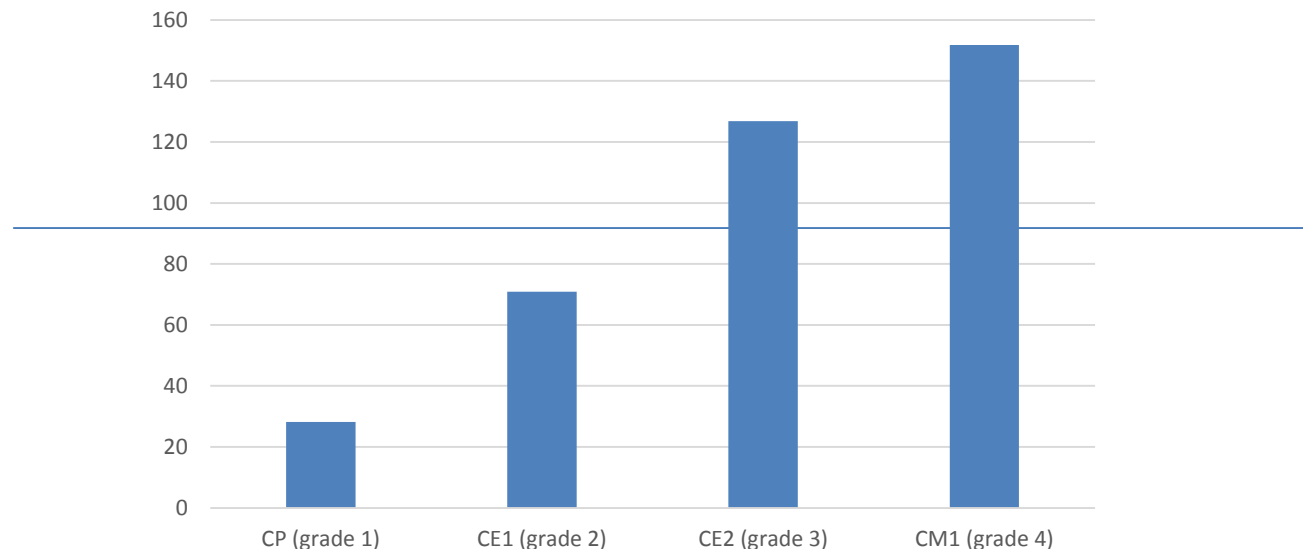
TXM

- Choix
 - gratuit et facilement accessible
 - connaissance en interne
 - prise en compte des métadonnées
 - suffisant pour nos objectifs
- Utilise Treetagger
 - tokenisation (découpage en token; word pour TXM)
 - lemmatisation
 - POS Tagging (catégories morphosyntaxiques)
 - F-Mesure sur notre corpus (Wolfarth 2019) : 93,9%
- les ponctuations sont considérées comme un mot
- unités polylexicales => n mots

Caractéristiques générales

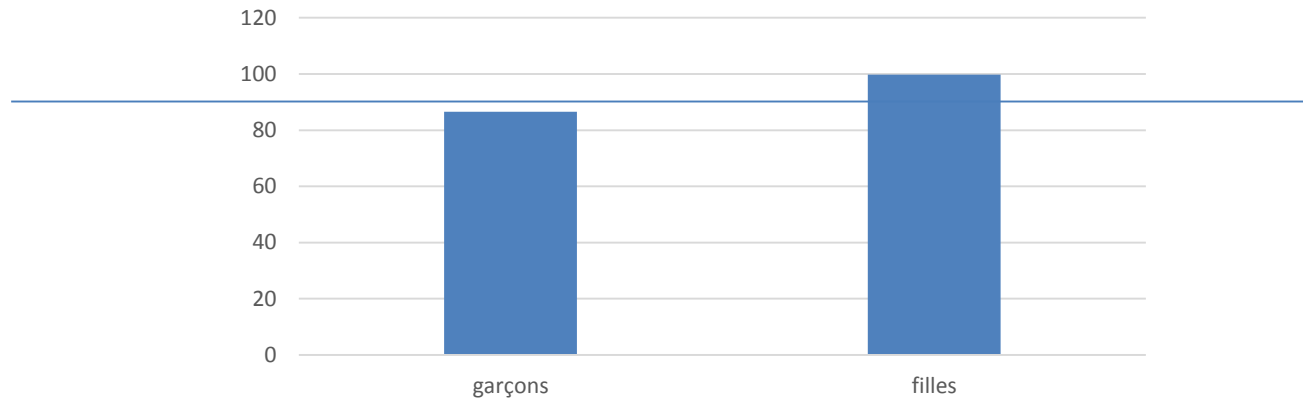
- 373 textes par niveau = 1.492 textes
- Total = 140.878 mots
- Longueur moyenne = 94,42 mots / texte.

Longueur moyenne selon le niveau

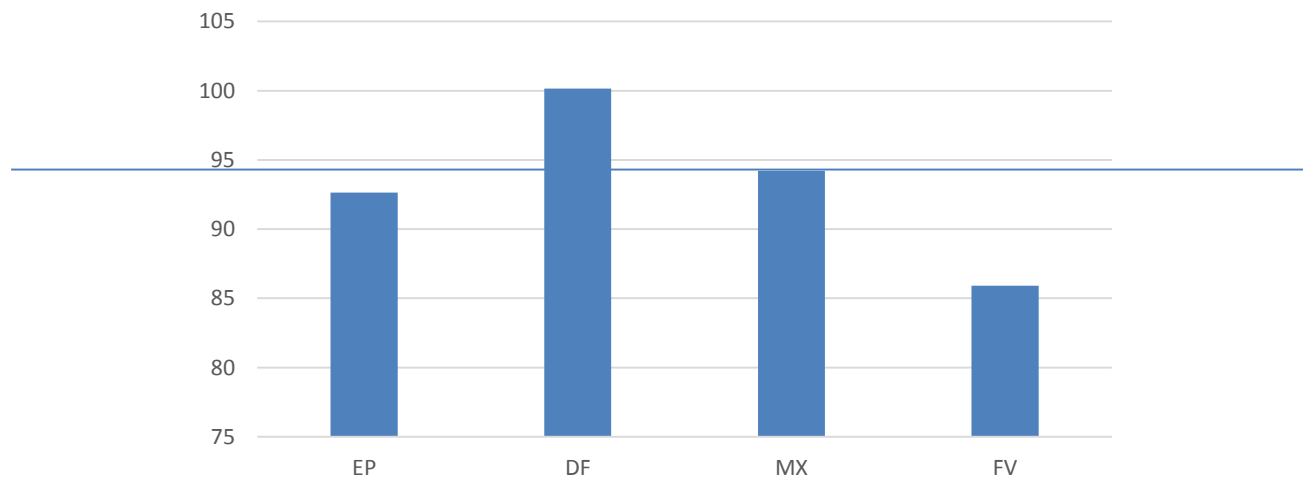


Longueur moyenne

Longueur moyenne selon le genre



Longueur moyenne selon la compo. sociale



Lexique (CP - lemmes)

frlemma	Fréquence
le	1128
chat	685
et	678
.	657
il	656
maman	360
son	356
se	345
petit	335
un	322
être	301
tomber	288
pleurer	155
avoir	147
miaou	143
faire	132
réveiller	121
de	114
qui	103
chaton	98
mal	96
marche	86
sur	82
o4	

ORTHO	U ▼1	SFI
le	88362.40	89.46
un	57649.29	87.61
de	29779.39	84.74
être	25611.10	84.08
il	24901.66	83.96
avoir	16599.45	82.20
et	16574.33	82.19
à	14660.67	81.66
son	11182.73	80.49
se	10539.88	80.23
dans	10264.52	80.11
je	9249.43	79.66
elle	8977.08	79.53
ce	8201.02	79.14
ne	7377.22	78.68
qui	6617.53	78.21
du	6517.12	78.14
en	6156.27	77.89
faire	6148.04	77.89
dire	5949.40	77.74
au	5886.91	77.70
pour	5832.38	77.66
petit	5785.90	77.62
tu	5467.58	77.38

MANULEX CP

Lexique (CE1 – CE2/CM1)

flemma	Fréquence
le	2210
un	1395
.	1355
il	1155
et	1121
être	778
chat	737
avoir	516
de	495
sorcier	474
elle	463
se	462
loup	385
son	358
qui	339
du	337
dans	304
,	278
à	275
mais	246
ne	246
fois	242
robot	237

ORTHO	U ↓ ₁	SFI
le	86501.89	89.37
un	36777.37	85.66
de	36032.89	85.57
être	24540.55	83.90
il	20824.75	83.19
et	18519.65	82.68
avoir	17206.18	82.36
à	14815.37	81.71
se	11911.98	80.76
je	10690.79	80.29
dans	9939.51	79.97
ne	9850.00	79.93
son	9275.78	79.67
qui	8019.05	79.04
en	7929.93	78.99
ce	7594.21	78.80
elle	7199.06	78.57
tu	7018.77	78.46
du	6991.62	78.45
au	6710.75	78.27
pas	6642.91	78.22
que	6524.87	78.15
faire	6365.23	78.04

MANULEX CE1

flemma	Fréquence
le	8324
.	4681
un	4405
il	3682
et	3472
être	2904
de	2280
chat	2162
avoir	2086
elle	1829
,	1800
sorcier	1676
se	1544
loup	1282
son	1160
que	1095
du	1084
à	1079
dans	1049
ne	980
ce	979
mais	969
qui	944
robot	933

ORTHO	U ↓ ₁	SFI
le	85034.66	89.30
de	43354.65	86.37
un	33043.08	85.19
être	22961.10	83.61
il	19320.12	82.86
et	18143.27	82.59
à	16868.45	82.27
avoir	16254.53	82.11
se	11404.82	80.57
ne	9685.98	79.86
en	9476.82	79.77
son	9343.79	79.71
je	9041.06	79.56
dans	8429.10	79.26
du	8184.73	79.13
qui	8106.92	79.09
que	7481.87	78.74
au	7298.22	78.63
ce	6807.55	78.33
le	6693.70	78.26
ce	6556.39	78.17
pas	6257.58	77.96
pour	5964.94	77.76
faire	5670.05	77.54

MANULEX CE2-CM2

Indice de spécificité

- Une des fonctionnalités de TXM
 - la probabilité d'apparition d'un élément dans une partie
 - Sur-spécificité (indice >2) : sur-représentation
 - Sous-spécificité (indice <-2) : sous-représentation
 - Banalité ($-2 < \text{indice} < 2$)
- Permet d'avoir une idée sur les termes plus ou moins spécifiques à une partie

Spécificité Lemmes / Niveau

Lemme	F	f_CP	Score_CP	f_CE1	Score_CE1	f_CE2	Score_CE2	f_CM1	Score_CM1
maman	426	360	1000	19	-19,1722	24	-44,0261	23	-58,5264
petit	816	335	153,4569	109	-5,7251	203	-7,5214	169	-28,9487
pleurer	207	155	125,457	14	-6,7256	20	-15,1779	18	-22,4678
chat	3584	685	110,8901	737	1,2102	1084	-5,2819	1078	-29,8568
son	1874	355	56,2896	358	-0,4862	537	-5,7835	623	-7,1715
chaton	193	98	55,7356	36	-0,3774	33	-6,6834	26	-14,4262
et	5271	678	39,9408	1121	3,1512	1607	-6,5631	1865	-8,2189
se	2351	345	30,3021	462	0,342	670	-7,4665	874	-1,5625
il	5493	655	29,2034	1155	2,6057	1712	-4,5174	1970	-6,6891
ramener	85	40	21,6324	11	-1,1185	16	-2,7249	18	-3,4749
marcher	128	45	18,8715	18	-1,1609	20	-5,4414	44	-0,8039
promener	227	61	17,4905	64	2,9807	49	-4,343	53	-6,4643
prendre	349	78	16,9709	52	-1,8326	98	-1,829	121	-1,3082

Lemmes les plus spécifiques au CP

Spécificité Lemmes / Niveau

CE1

Lemme	F	f_CP	Score_CP	f_CE1	Score_CE1	f_CE2	Score_CE2	f_CM1	Score_CM1
manger	593	12	-8,9493	193	13,3652	187	-0,8337	201	-2,311
un	6122	322	-14,3065	1395	10,2842	2050	-0,4038	2355	-0,8248
souris	113	3	-1,6542	52	9,7642	37	-0,3348	21	-5,659
foi fois	929	61	-0,978	242	6,1255	308	-0,4109	318	-2,936
lait	26	0	-0,9051	16	5,4996	8	-0,3298	2	-3,4203

CE2

Lemme	F	f_CP	Score_CP	f_CE1	Score_CE1	f_CE2	Score_CE2	f_CM1	Score_CM1
sorcier	2150	0	-75,6354	474	2,7186	888	13,3211	788	-2,0158
je	1016	9	-23,6748	141	-5,9211	416	6,1872	450	3,396
lui	852	18	-12,1461	138	-2,1773	343	4,5155	353	1,0615
«	486	1	-15,3357	55	-6,124	204	4,1132	226	3,2789
dire	1150	17	-20,9989	212	-0,7439	448	4,0652	473	1,0796

CM1

Lemme	F	f_CP	Score_CP	f_CE1	Score_CE1	f_CE2	Score_CE2	f_CM1	Score_CM1
-	507	0	-17,6909	20	-24,6232	193	1,692	294	17,2257
,	2159	81	-13,809	278	-16,122	806	3,8096	994	10,6716
de	2889	114	-16,4378	495	-3,3691	985	0,5073	1295	9,8974
!	684	25	-5,0508	85	-6,3309	240	0,6498	334	6,8502
elle	2352	60	-26,8148	463	0,3582	790	-0,3155	1039	6,6405

Spécificité Lemme selon le genre

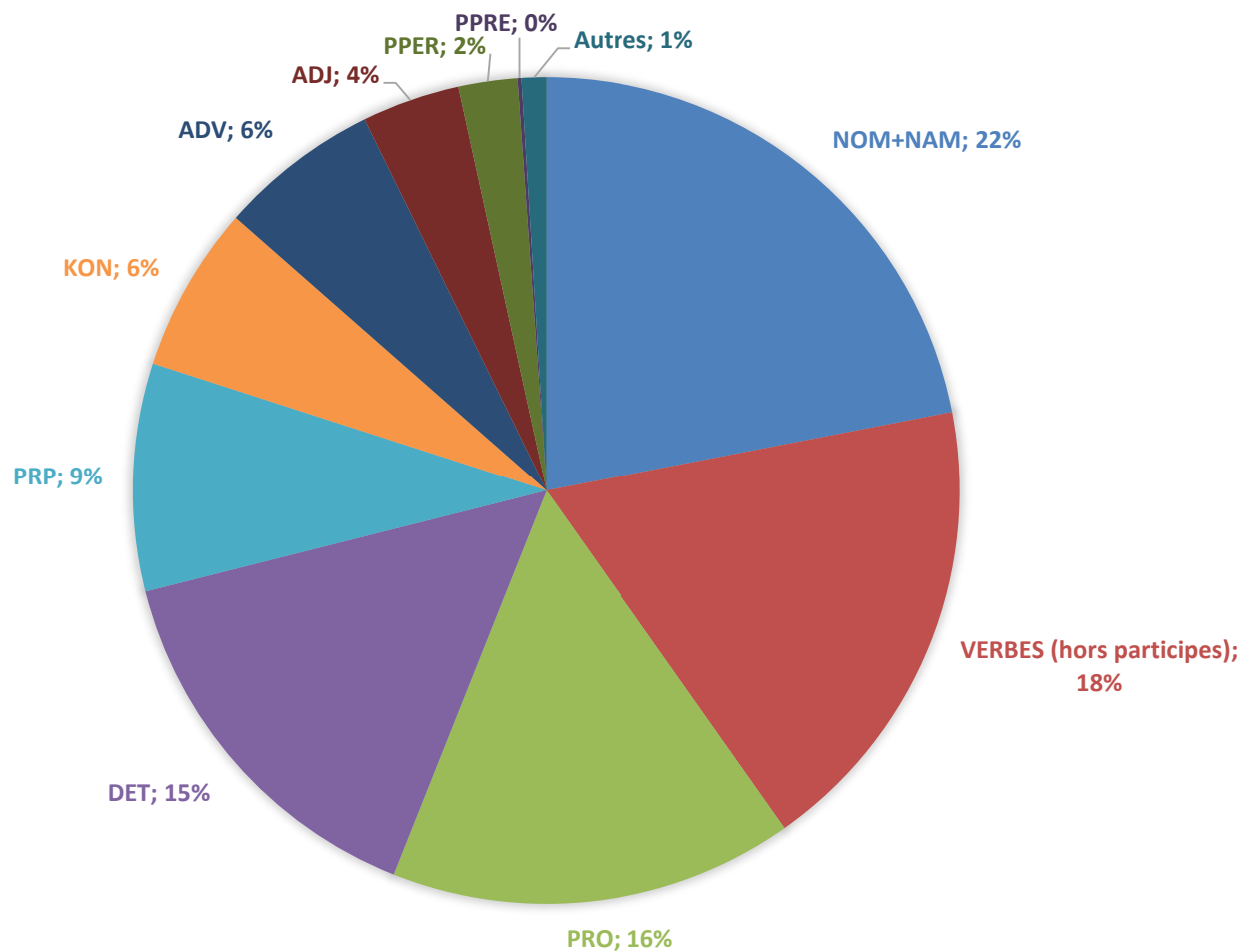
Lemme	F	f_Garçons	Score_Garço	f_Filles	Score_Filles
robot	1170	706	48,7302	464	-48,7302
loup	1667	880	29,6853	787	-29,6853
il	5493	2421	14,3929	3072	-14,3929
le	11662	4926	13,2696	6736	-13,2696
manger	593	281	4,6636	312	-4,6636
et	5271	2200	4,6205	3071	-4,6205
tuer	92	54	3,9678	38	-3,9678
@card@	338	166	3,9677	172	-3,9677
un	6122	2529	3,9674	3593	-3,9674
courir	130	71	3,6342	59	-3,6342

Garçons

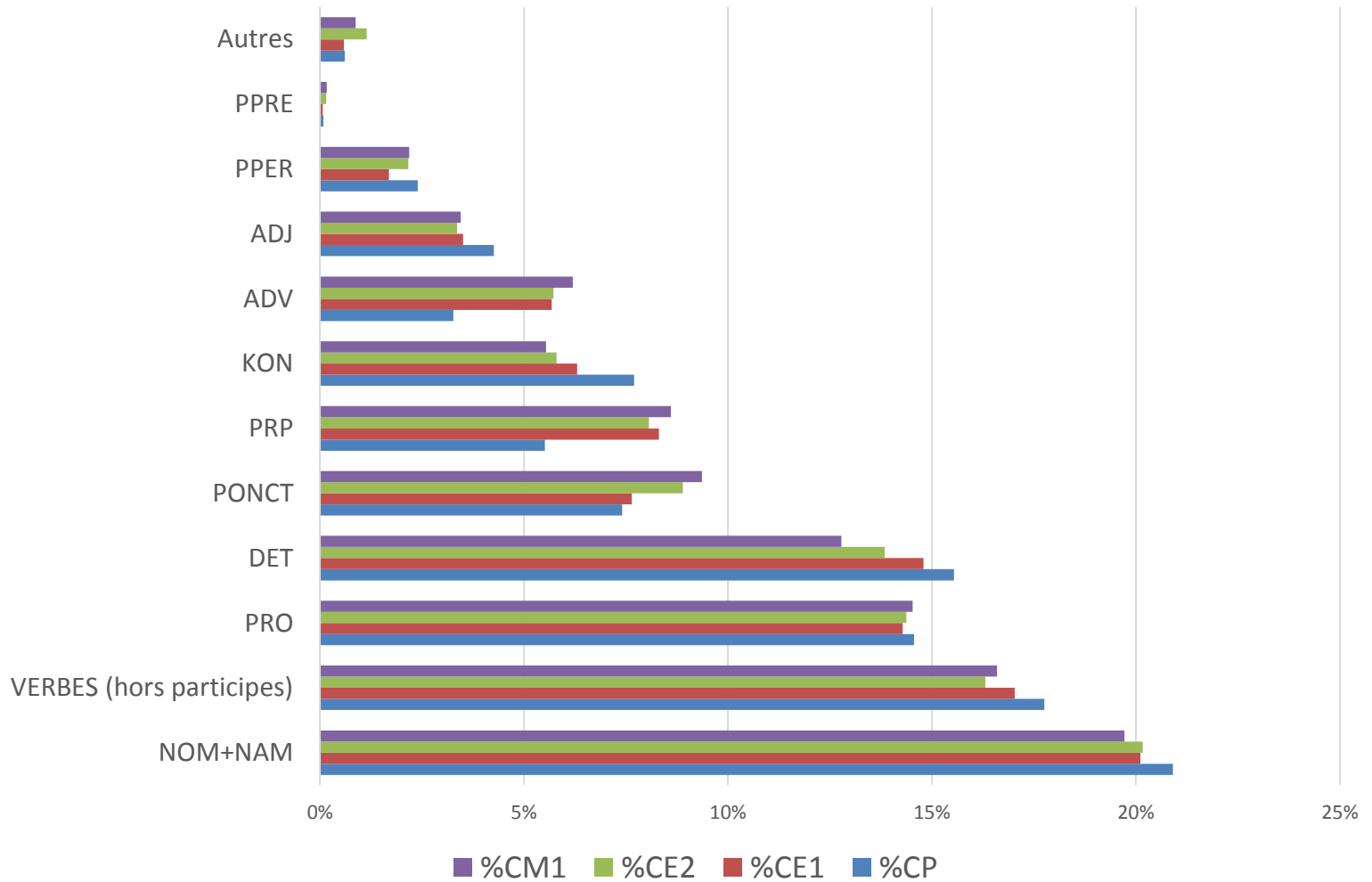
Lemme	F	f_Garçons	Score_Garço	f_Filles	Score_Filles
elle	2352	603	-43,1882	1749	43,1882
,	2159	688	-11,8413	1471	11,8413
sorcier	2150	721	-7,3032	1429	7,3032
mon	251	63	-5,6769	188	5,6769
avec	562	178	-3,795	384	3,795
au	565	180	-3,6571	385	3,6571
me	269	77	-3,6553	192	3,6553
magique	177	47	-3,4936	130	3,4936
son	1874	661	-3,4453	1213	3,4453
appeler	487	154	-3,4168	333	3,4168
chaton	193	53	-3,3143	140	3,3143

Filles

Répartition des catégories



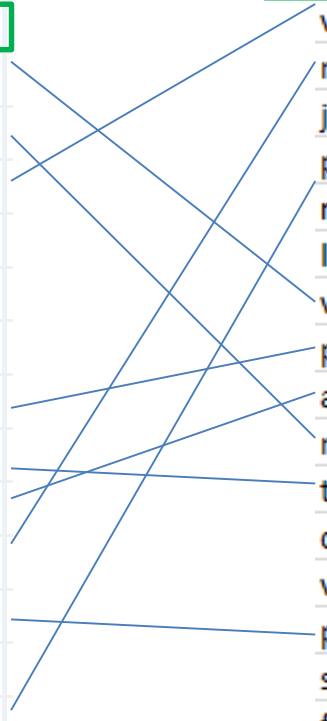
Répartition des catégories par niveaux



Lexique des verbes

frlemma	Fréquence
être	3981
avoir	2749
dire	1149
aller	1145
faire	1058
vouloir	569
manger	561
voir	492
appeler	487
tomber	414
vivre	371
prendre	349
trouver	298
arriver	286
mettre	279
partir	274
transformer	247
pouvoir	240
devenir	232

LEMMES	CP U
être	25611,1
avoir	16599,45
faire	6148,04
dire	5949,4
aller	5241,73
voir	2507,35
mettre	2133,68
jouer	2037,79
pouvoir	1864,06
regarder	1779,27
lire	1757,45
vouloir	1672,04
prendre	1656,89
arriver	1628,44
manger	1468,3
trouver	1349,06
donner	1230,13
venir	1159,32
partir	1086,88
savoir	1062,68
falloir	1009,21
passer	990,79



Manulex

Verbes / niveaux

Unités	Fréquence T 26635	CP t=2129	score	CE1 t=4966	score	CE2 t=8811	score	CM1 t=10729	score
tomber	414	288	219,0	27	-12,1	44	-26,2	55	-33,2
pleurer	205	154	124,2	14	-6,0	19	-15,1	18	-23,6
réveiller	187	121	84,9	13	-5,4	31	-6,6	22	-17,3
dormir	159	64	28,5	41	1,8	20	-8,7	34	-6,5
ramener	84	39	20,4	11	-0,9	16	-2,5	18	-3,7
marcher	128	46	18,3	18	-1,0	20	-5,2	44	-1,0
miauler	64	32	18,0	4	-2,4	19	-0,5	9	-5,3
promener	227	61	16,9	64	3,6	49	-4,1	53	-7,3

Unités	Fréquence T 26635	CP t=2129	score	CE1 t=4966	score	CE2 t=8811	score	CM1 t=10729	score
manger	561	12	-8,8	187	16,4	172	-0,9	190	-3,0
vouloir	569	21	-4,7	155	6,6	167	-1,5	226	-0,4
jouer	163	3	-3,1	54	5,2	55	0,3	51	-2,0
promener	227	61	16,9	64	3,6	49	-4,1	53	-7,3
boire	66	0	-2,4	24	3,3	16	-1,1	26	-0,3
gouter	4	0	-0,1	4	2,9	0	-0,7	0	-0,9
taper	7	0	-0,3	5	2,5	0	-1,2	2	-0,4

Unités	Fréquence T 26635	CP t=2129	score	CE1 t=4966	score	CE2 t=8811	score	CM1 t=10729	score
détruire	37	0	-1,3	1	-2,3	24	4,1	12	-0,7
jurer	7	0	-0,3	0	-0,6	7	3,4	0	-1,6
écraser	22	0	-0,8	2	-0,7	15	3,1	5	-1,2
voir	492	16	-5,0	99	0,7	195	2,9	182	-1,1
attraper	140	11	-0,3	35	1,4	63	2,7	31	-5,4
dévoré	14	0	-0,5	4	0,6	10	2,4	0	-3,1
terroriser	5	0	-0,2	0	-0,4	5	2,4	0	-1,1
aimer	143	0	-5,2	37	1,7	62	2,2	44	-1,9
rater	7	0	-0,3	0	-0,6	6	2,2	1	-0,8

Verbes / genre

Unités	Fréquence T 26635	Garçons t=10474	score	Filles t=16161	score
gagner	28	25	7,2	3	-7,2
détruire	37	30	6,6	7	-6,6
arrêter	102	63	5,4	39	-5,4
attaquer	56	39	5,4	17	-5,4
manger	561	268	4,6	293	-4,6
tuer	92	54	3,9	38	-3,9
courir	130	71	3,5	59	-3,5
sauver	48	30	3,0	18	-3,0
lâcher	7	7	2,8	0	-2,8
rater	7	7	2,8	0	-2,8
mourir	97	53	2,8	44	-2,8
jeter	95	52	2,8	43	-2,8

Unités	Fréquence T 26635	Garçons t=10474	score	Filles t=16161	score
travailler	23	1	-3,8	22	3,8
appeler	487	154	-3,6	333	3,6
occuper	19	1	-3,0	18	3,0
changer	22	2	-2,7	20	2,7
dire	1149	405	-2,7	744	2,7
partir	274	87	-2,3	187	2,3
parler	107	29	-2,3	78	2,3
traverser	15	1	-2,2	14	2,2
apprendre	22	3	-2,0	19	2,0
répondre	114	33	-1,9	81	1,9
coucher	38	8	-1,9	30	1,9

Remarques générales

- La variable composition sociale des classes semble peu discriminante
- Le lexique obtenu est comparable à Manulex si l'on omet les effets de consigne
- Le lexique est assez genré
- Les adjectifs sont peu mobilisés

Perspectives

- comparer ces résultats à d'autres corpus (E-Calm)
- gestion du lexique
 - gestion des champs lexicaux
 - les types de verbes (mouvement, parole...)
 - exploitation des lemmes « moins fréquents »
 - signal textes plus élaborés => persp. didactiques
- qualité des textes ?
 - richesse lexicale
- Variété des déterminants
- Utilisation des pronoms
- Utilisation de la subordination
- ...

Références

- Wolfarth C., Ponton C., Totereau C. (2017). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire. Dans C. Doquet, J. David & S. Fleury, Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement, *Corpus*, 16:2017, 185-214.
- Wolfarth C., Brissaud C., Ponton C. (2018). Transcrire et normer un corpus scolaire : pour quelles analyses ? *Dyptique*, 36, Presses Universitaires de Namur, pp.121-145
- Wolfarth C., Ponton C., Brissaud C. (2018). Gestion de la morphologie verbale en production d'écrits : que peut nous apprendre un corpus longitudinal ?. *Repères : Recherches en didactiques du français langue maternelle*. Analyses linguistiques des écrits d'élèves. (57). 209-226
- Wolfarth C., Ponton C., Brissaud C. (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
- 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