LIDILEM

UNIVERSITÉ Grenoble Alpes

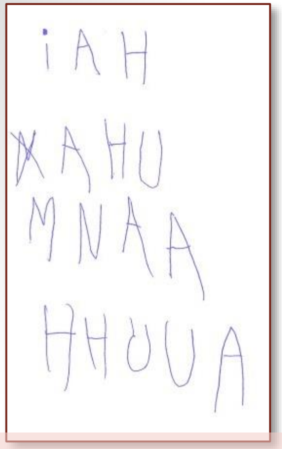
Which Method to Develop a Natural Language Processing Tool to automatically analyse First Language Learner Corpora?

Claire Wolfarth, Claude Ponton, Catherine Brissaud

Lidilem

Université Grenoble Alpes

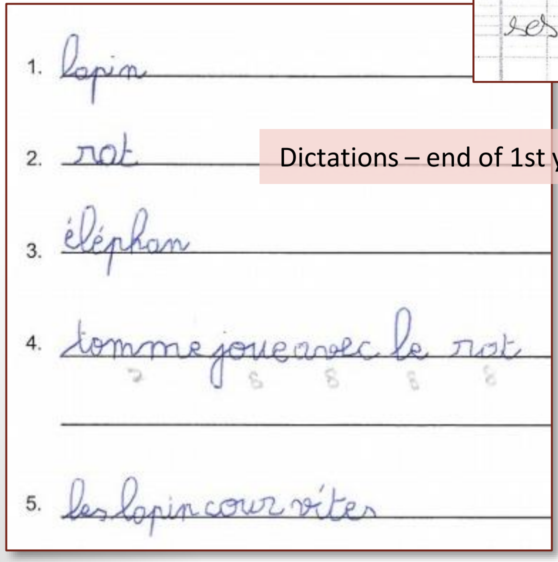
13 Teaching and Language Corpora Conference
18 – 21 July 2018, F. Education, University of Cambridge



iAH
XAHU
MNAA
HHUUA

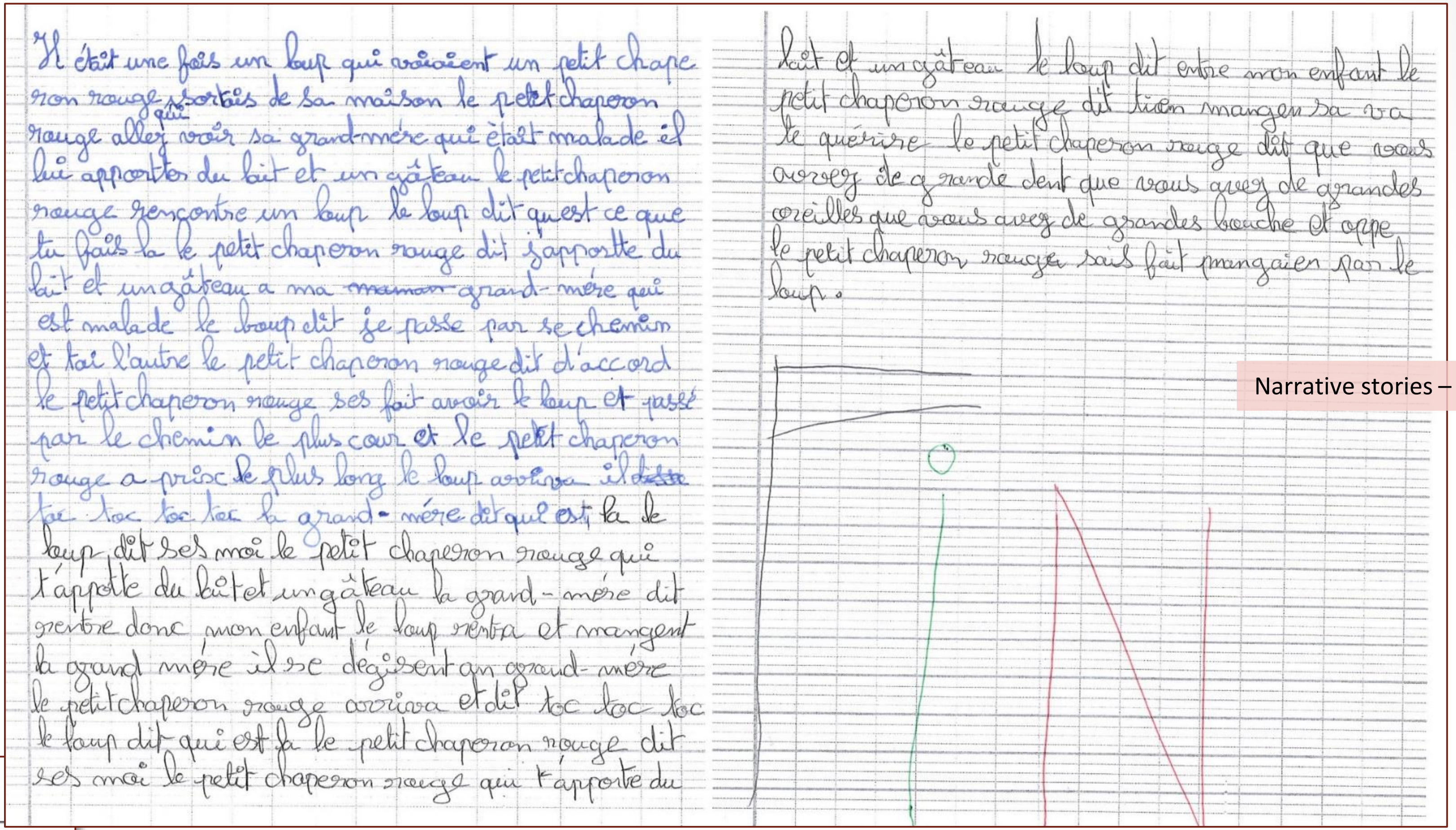
Dictations – begin of 1st year

LLAP
RA
ÉL
TERA



1. lapin
2. rot
3. éléphan
4. tomme joue avec le rot
5. les lapin cour vites

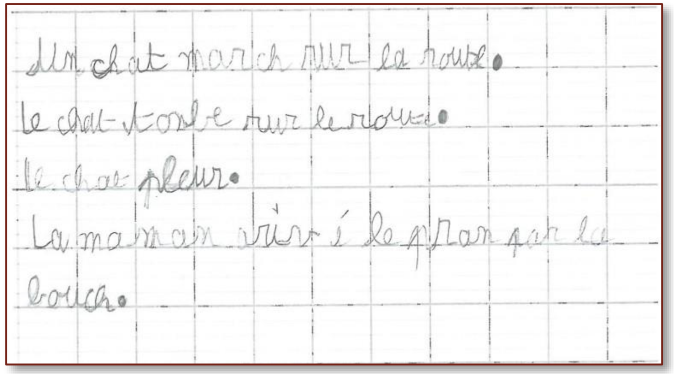
Dictations – end of 1st year



Il était une fois un loup qui avaient un petit chaperon rouge qui sortaies de sa maison le petit chaperon rouge allez voir sa grand-mère qui était malade el lui apporter du lait et un gâteau le petit chaperon rouge rencontre un loup le loup dit quest ce que tu fais la le petit chaperon rouge dit j'apportte du lait et un gâteau a ma ~~maman~~ grand-mère qui est malade le loup dit je passe par se chemin et toi l'autre le petit chaperon rouge dit d'accord le petit chaperon rouge ses fait avoir le loup et passé par le chemin le plus cour et le petit chaperon rouge a prise le plus long le loup arriva il ~~dette~~ toc toc toc toc la grand-mére dit qui est la le loup dit ses moi le petit chaperon rouge qui t'appotte du lait et un gâteau la grand-mère dit rentre donc mon enfant le loup rentra et mangent la grand mére il se dégisent an grand-mére le petit chaperon rouge arriva et dit toc toc toc le loup dit qui est la le petit chaperon rouge dit ses moi le petit chaperon rouge qui t'apporte du lait et un gâteau le loup dit entre mon enfant le petit chaperon rouge dit tien manger sa va le quérire le petit chaperon rouge dit que vous avez de grande dent que vous avez de grandes oreilles que vous avez de grandes bouche et oppe le petit chaperon rouge sais fait mangaien par le loup.

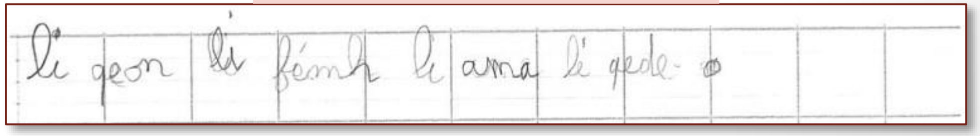
Narrative stories – end of 3rd year

Narrative stories – end of 1st year



Un chat march sur la route.
le chat tombe sur le route.
le chat pleure.
La maman vien é le pran par la bouch.

Narrative stories – end of 1st year



li qeon li fémh li ama li qede.

Inventory of school corpora

- 2nd language learner corpora
- Emergence of large school corpus projects
  - Longitudinal
    - Juel, 1988 (English)
    - Lancaster Corpus of Children's Project Writing, 1998 (English)
    - Corpus ÉMA, Boré & Elalouf, 2017 (French)
  - Non longitudinal
    - Elalouf, 2005 (French)
    - Project Ecriscol, David et Doquet, 2016 (French)
- Length, accessibility, processes

I –*Scoledit:* a French longitudinal school Corpus

Project funded by French research national Agency (E-CALM project)

Scoledit corpus

Issues

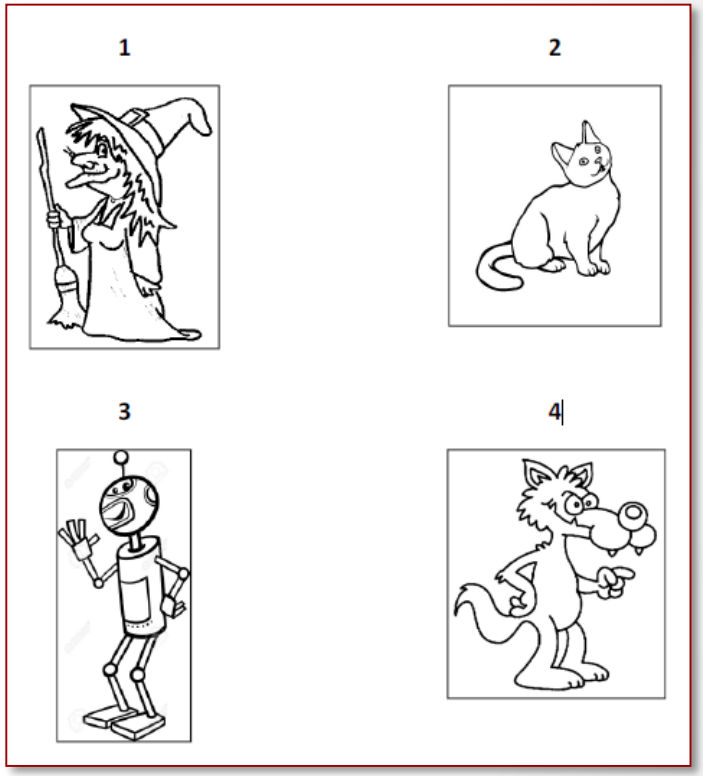
- In linguistics
  - Synchronous and diachronic description of the characteristics of learners' writings from 6 to 11 years old
  - Better knowledge of writing acquisition phenomena (evolution of syntax acquisition, spelling, punctuation, etc.)
- In automatic language processing (NLP)
  - Type of texts still little studied
- In didactics
  - Development of sequences and devices based on this knowledge

Scoledit corpus

- A longitudinal corpus of narrative writings and dictations
  - 5 years (primary school)
  - 40 schools: 800 – 1200 pupils per year
  - narrative writings + dictations

http://scoledit.org/scoledit/

Grade	Number of writings	Number of words
1st year (2013-2014)	977	16 000
2nd year (2014-2015)	876	67 000
3rd year (2015-2016)	1190	153 500
4th year (2016-2017)	1102	180 000
5th year (2017-2018)	1000	/
Total	**5244**	**416 300**

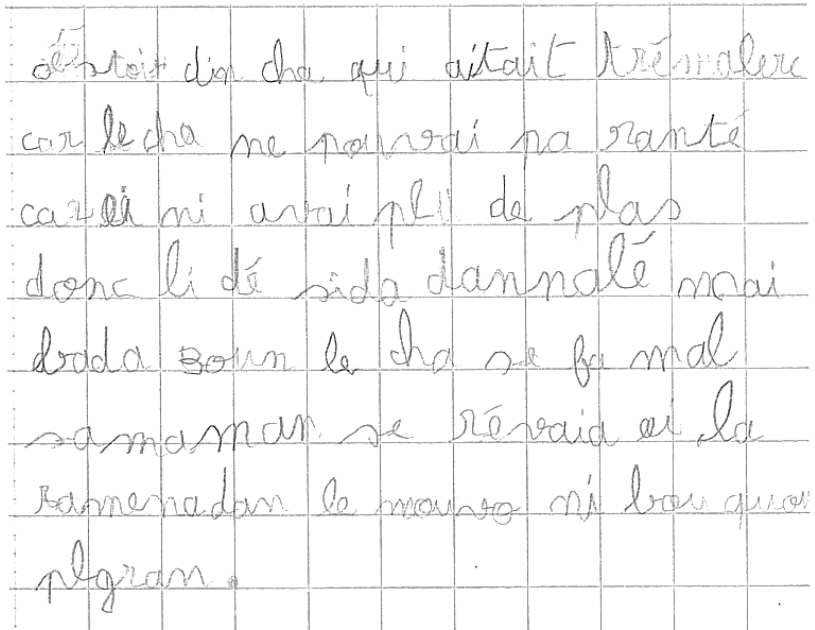


Children's writing specificities

Children 297, at the end of the first year

« sé stoir din cha qui aitait tré malere car le cha ne pouvai pa ranté car li ni avai plu de plas donc li dé sida dannalé mai bada Boum le cha se fa mal sa maman se révaia ai la ramena dan le nouvo ni bou quou plgran. »

C'est l'histoire d'un chat qui était très malheureux car le chat ne pouvait pas rentrer car il n'y avait plus de place donc il décida de s'en aller mais badaboum le chat se fait mal <segmentation/> sa maman se réveilla et le ramena dans le nouveau nid beaucoup plus grand.



Children's writing specificities

Children 297, at the end of the third year

« Il était une fois une petite chatte qui saplai Boubou, on l'avait apelée Boubou par cequ elle adorait faire des blages. Mais un jour, cette petite farsez a u l'idée de dir une blage mechate aux policier. Mais plus quel etait toute petite elle ne savai pas quel pouvé aller en prison. mai quand elle vu les abi du monsieur tombée elle vu que seté un robeau qui lui di « C est carnavale aujourd'hui » vien je vais te dégiser. Et il ver la fête tout la journée, et depui se jour le chat et le robo ne se quiter plus "FIN" »

Il était une fois une petite chatte qui s'appelait Boubou, on l'avait appelée Boubou parce qu'elle adorait faire des blagues. Mais un jour, cette petite farceuse a eu l'idée de dire une blague méchante aux policiers. Mais puisqu'elle était toute petite elle ne savait pas qu'elle pouvait aller en prison. Mais quand elle vit les habits du monsieur tomber elle vit que c'était un robot qui lui dit "C'est carnaval aujourd'hui" viens je vais te déguiser. Et il fait la fête toute la journée, et depuis ce jour le chat et le robot ne se quittèrent plus "FIN"

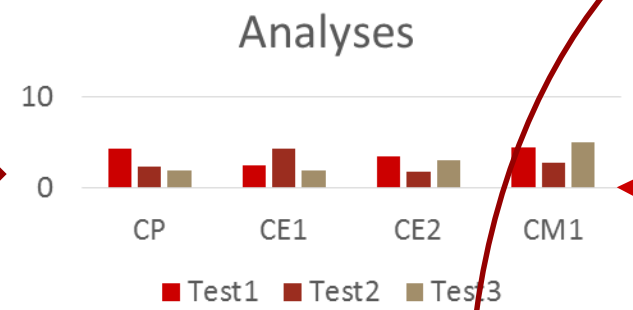
II – Analysis of school corpora : Alignment/Comparison approach

Processing chain

Data collection (protocol)

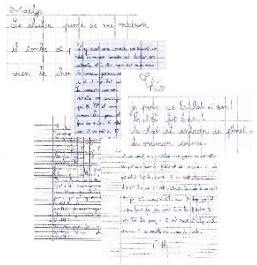
???

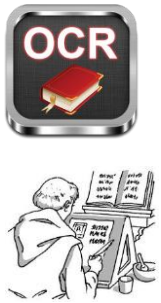
Question of volume and resource sharing



Alignment / comparison modules

Digitalisation
Scanning,
Storage, backup,
loss of data





Transcription
Choices (loss of data),
A description language,
Interpretation, errors

```
La sorciére 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.
```

Annotation
Data enrichment,
Some errors

```
<BE-ILV-008-005>
Christian Bex en
direct du
<F><MAJ><NOP>
#Vendée$ vendée
</NOP></MAJ></F>
globe challenge, 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)

Annotation methods

- Classical approach

```
- <item type="forme" pos="19">
  - <f>
      s(remplacement)(b)(del)é(/del)(/b)(a)e(/a)(/remplacement)rai
    </f>
    <c>CAT</c>
    <l>serai</l>
  </item>
- <item type="delim" pos="20">
    <f> </f>
    <c>DELIM</c>
    <l> </l>
  </item>
- <item type="forme" pos="21">
  - <f>
      (remplacement)(b)(del)inventore(/del)(/b)(a)inventerr(/a)
    </f>
    <c>CAT</c>
    <l>inventeur</l>
  </item>
```

```
<BE-ILV-008-005>
Christian Bex en direct du <F><MAJ><NOP> #Vendée$
vendée </NOP></MAJ></F> globe challenge, une aventure
périlleuse <G><CLA><POR> #qui$ qu' </POR></CLA></G>est
<G><CLA><ADE> #une$ la </ADE></CLA></G> course à la
voile.
```

```
Je pense qu'aller à l'école en vélo et bien et en effet £
ça ne <polue>_<pollue> pas. Le vélo <et>_<est> très bon pour la santé. §
En premier lieu ça permet de mieux <conaitre>_<connaître> la £
nature. De plus on peut sauter par dessus les <obstables>_<obstacles> §
Mais quand il n'y a pas de <trotoir>_<trottoir> ou de piste £
<cyclabes>_<cyclable> le vélo peut être dangereux. §
Finalement, je pense donc que le vélo est une £
bonne idée. §
```

Annotation methods

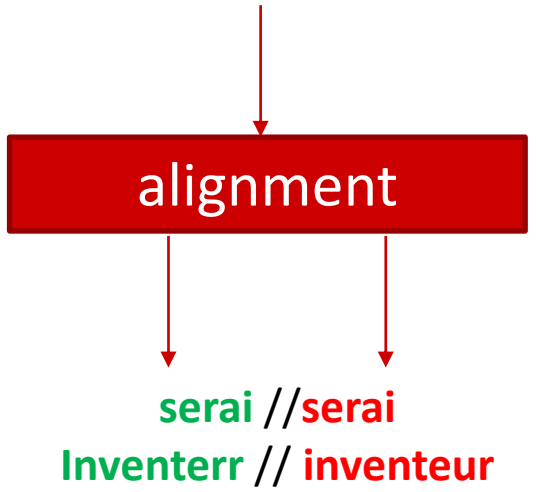
- **Classical approach**

```
<item type="forme" pos="19">
 <f>
  s(remplacement)(b)(del)é(/del)(/b)(a)e(/a)(/remplacement)rai
 </f>
 <c>CAT</c>
 <l>serai</l>
</item>
<item type="delim" pos="20">
  <f> </f>
 <c>DELIM</c>
 <l> </l>
</item>
<item type="forme" pos="21">
 <f>
  (remplacement)(b)(del)inventore(/del)(/b)(a)inventerr(/a)(/remplacement)
 </f>
 <c>CAT</c>
 <l>inventeur</l>
</item>
```

- **Scoledit approach**

Transcription: **serai inventerr**

Normalisation: **serai inventeur**



Scoledit approach: normalisation

Children 651 transcription: " *sé stoir din cha qui aitait tré malere … "*

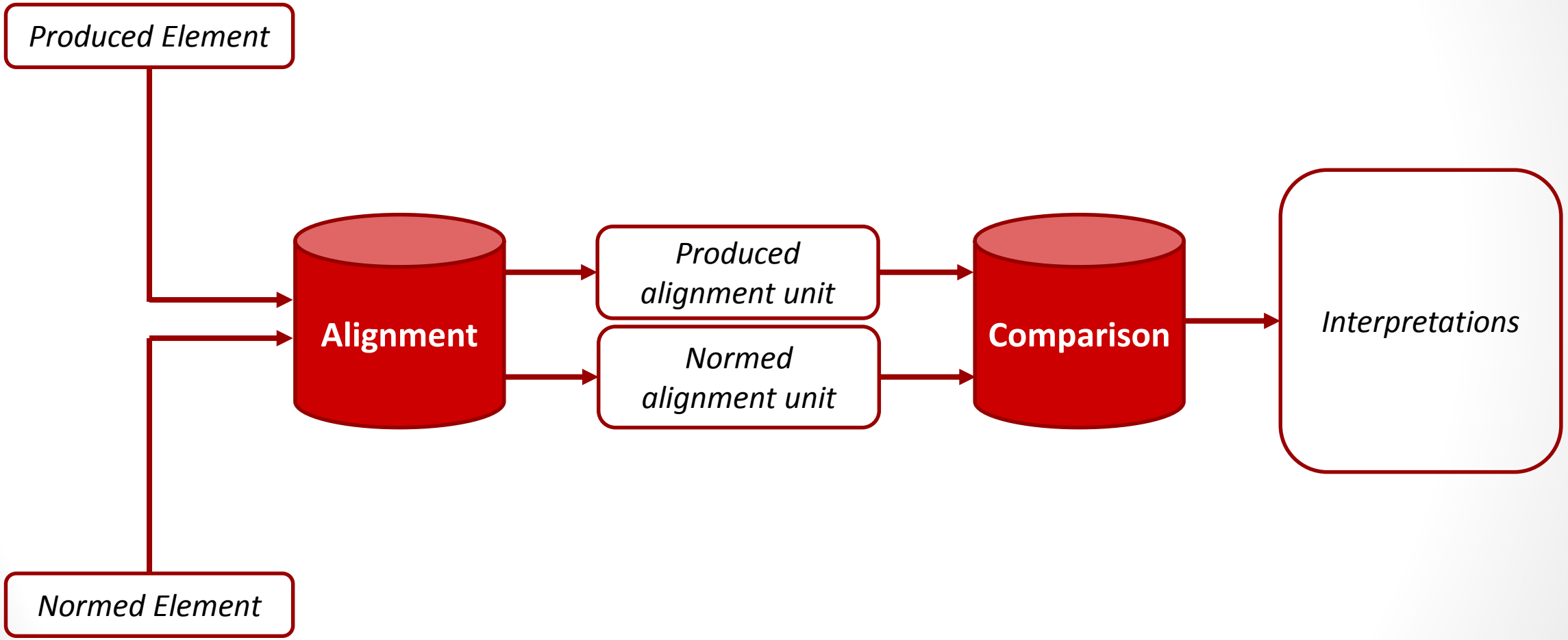
Children 651 normalisation: " C'est l'histoire d'un chat qui était très malheureux… "

- Advantages
  - No heavy formalism
  - Simpler than annotation
  - Reusability of procedures: e.g. dictation, rewriting exercises...
  - More freedom in the exploitation?
- Disadvantages
  - Automatic alignment => increased risk of error

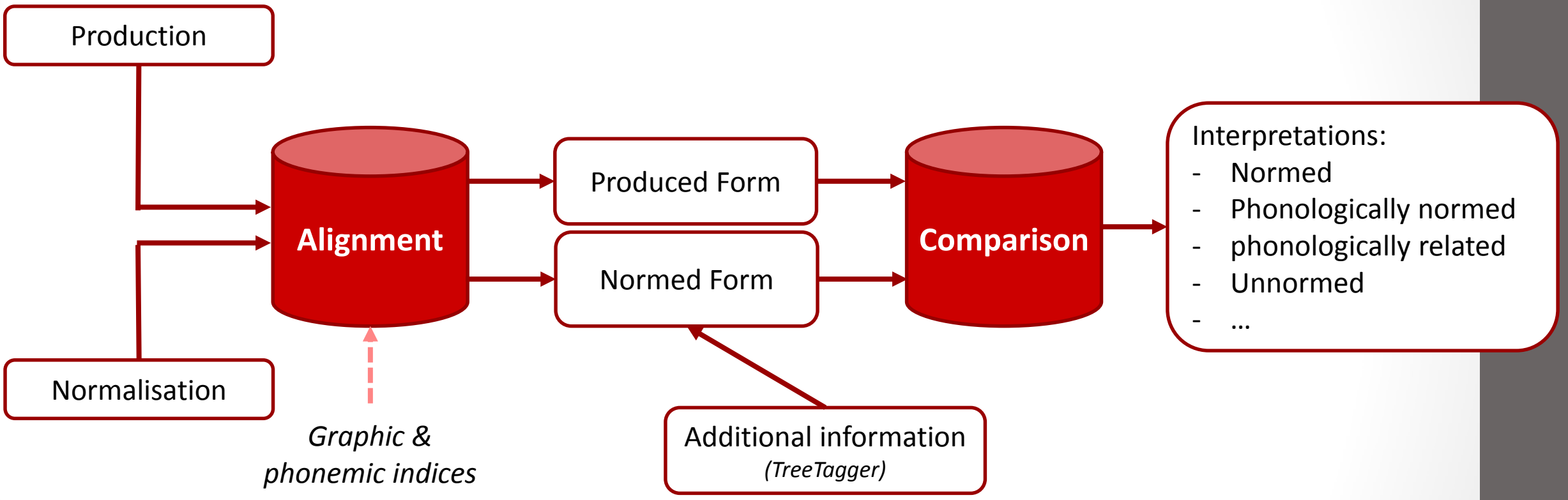
Related works

- Dictations
  - Santiago-Oriola (1998)
  - Beaufort et Roekhaut (2011)
- Translation
  - Desmet, Hérogue, 2005

Alignment/Comparison approach



Alignment/Comparison approach



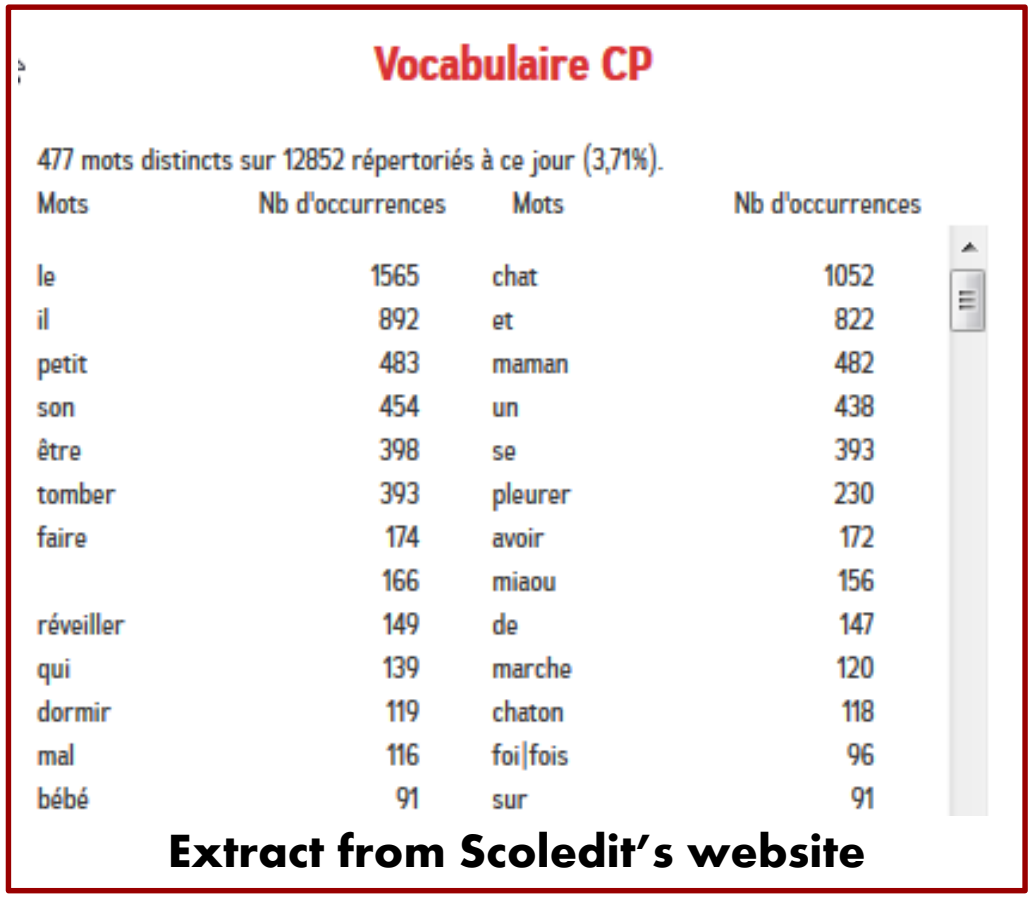
Transcription : *din cha* ⟶ / dɛ̃ʃa /

Normalisation : d'un chat ⟶ / dœ̃ʃa /

d	d‘	PRP	de	phonologically normed
in	un	DET:ART	un	phonologically related
cha	chat	NOM	chat	phonologically normed

Under-segmented
Under-segmented
Well-segmented

Alignment/Comparison approach



Vocabulaire CP

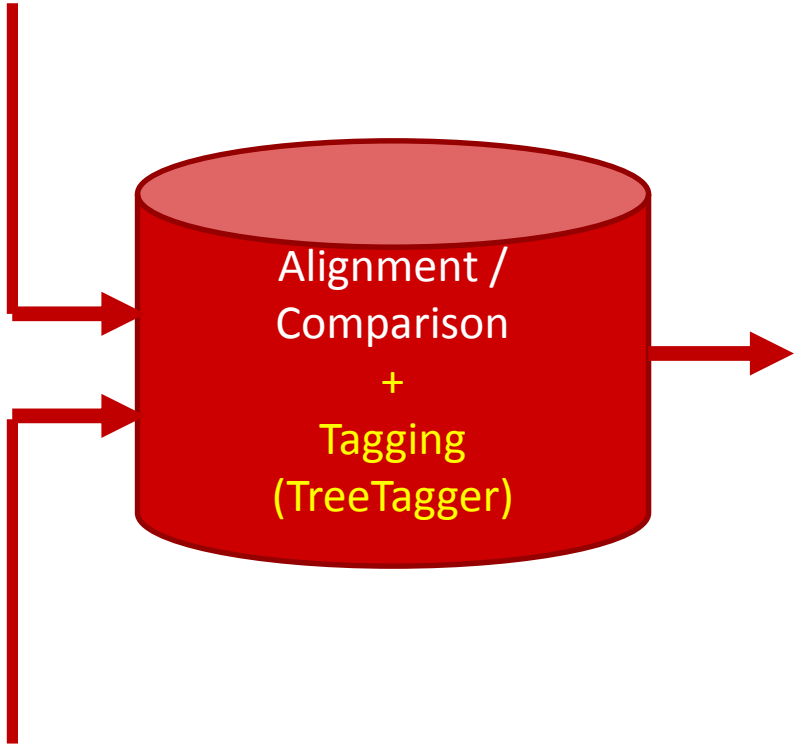
477 mots distincts sur 12852 répertoriés à ce jour (3,71%).

Mots	Nb d'occurrences	Mots	Nb d'occurrences
le	1565	chat	1052
il	892	et	822
petit	483	maman	482
son	454	un	438
être	398	se	393
tomber	393	pleurer	230
faire	174	avoir	172
	166	miaou	156
réveiller	149	de	147
qui	139	marche	120
dormir	119	chaton	118
mal	116	foi|fois	96
bébé	91	sur	91

Extract from Scoledit's website

Alignment/Comparison approach

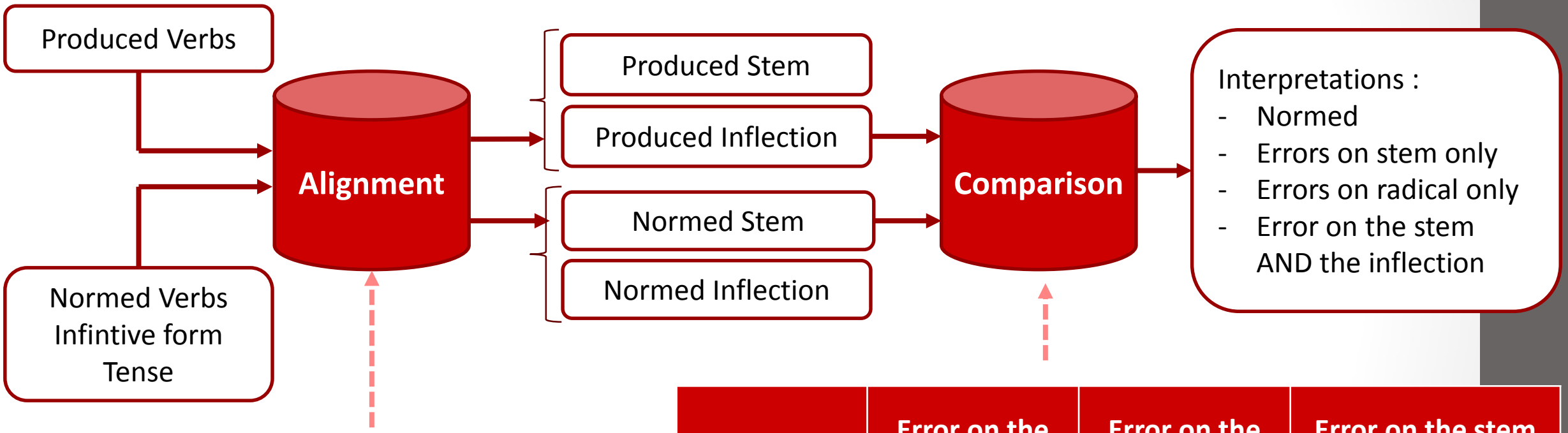
Children 651 transcription: " *sé stoir din cha qui aitait tré malere car le cha ne pouvai pa ranté … "*



Transcribed forms	*Normalised forms*	*Parts of speech*	*Lemmas*
…	…	…	…
qui	qui	*PRO:REL*	*qui*
aitait	était	*VER:impf*	*être*
tré	très	*ADV*	*très*
malere	malheureux	*ADJ*	*malheure*
car	car	*KON*	*car*
le	le	*DET:ART*	*le*
cha	chat	*NOM*	*chat*
ne	ne	*ADV*	*ne*
pouvai	pouvait	*VER:impf*	*pouvoir*
pas	pas	*ADV*	*pas*
ranté	rentrer	*VER:infi*	*rentrer*
	…		

Children 651 normalisation: " C'est l'histoire d'un chat qui était très malheureux car le chat ne pouvait pas rentrer… "

Alignment/Comparison approach



Linguistic theory (Martinet, 1979 ; Meleuc et Fauchart, 1999; Blanche-Benveniste, 2002 ; Pellat, 2009)

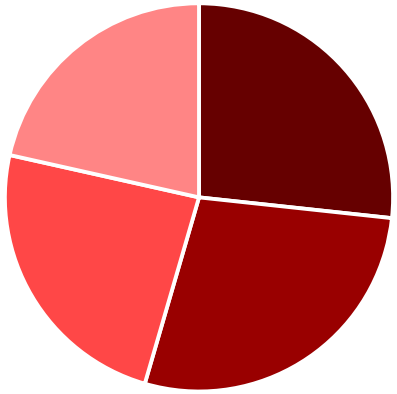
Normed	Error on the stem only	Error on the inflection only	Error on the stem AND the inflection
di + t // di + t	fai+re // fé+re	av+ait // av+ai	miaul+e // miol + Ø

Alignment/Comparison approach

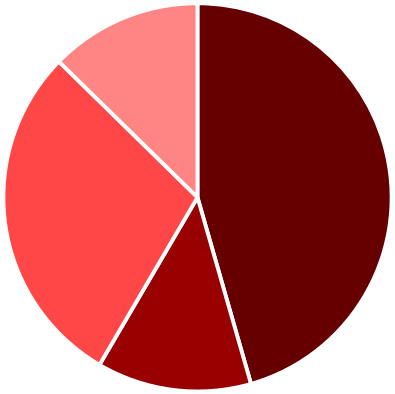
statut	forme	baseForme	desiForme	prod	baseProd	desiProd
normé	suivant	suiv	ant	suivant	suiv	ant
phonologi	fais	fai	s	fait	fai	t
non normé	était	ét	ait	étét	ét	ét
équivalen	faire	fai	re	fére	fé	re
normé	sais	sai	s	sais	sai	s
normé	dit	di	t	dit	di	t
normé	était	ét	ait	était	ét	ait
normé	est	es	t	est	es	t

Analysis of verbal morphography

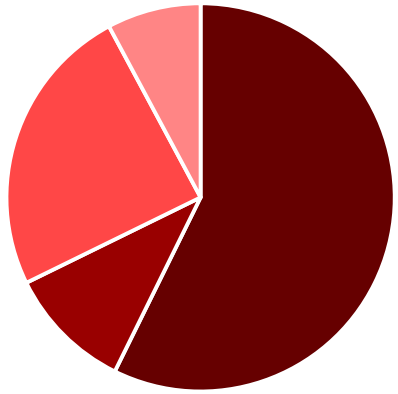
1st year



2nd year



3rd year



- Normed
- stem only
- termination only
- stem and termination

- Normed
- stem only
- termination only
- stem and termination

- Normed
- stem only
- termination only
- stem and termination

- Brissaud C., Totereau C., Ponton C., Wolfarth C. (to be published in 2018). Usage d'un corpus longitudinal, le cas de la morphologie verbale. *Revue Repères* : collecter, interpreter, enseigner l'écriture. Analyses linguistiques des écrits d'élèves.

III – Perspectives

- Method development
  - Treatment chain : evaluation des différents modules
  - quality of the data : Checking of the transcriptions and normalisations
- New modules
  - graphemes - phonemes correspondences
  - Agreement in noun groups
- E-Calm Project
  - Textual cohesion and coherence
  - Inflectional morphology and derivational morphology
- Purpose : describe linguistic uses in these corpora and their evolutions

Bibliography

- Beaufort R., Roekhaut S. (2011). Automation of dictation exercises. *Computational Linguistics in the Netherlands Journal*, (1), 1-20.
- Boré C., Elalouf M.-L. (2017). Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles. In Doquet C., David J., Fleury S., *Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement*, Corpus, 16 | déc. 2016, 185-214.
- David J., Doquet C. (2016). Les écrits d'élèves: un corpus de référence pour le français contemporain. Actes de *SHS Web of Conferences* (Vol. 27, p. 11001). EDP Sciences.
- Desmet P., Héroguel A. (2005). Les enjeux de la création d'un environnement d'apprentissage électronique axé sur la compréhension orale à l'aide du système auteur IDIOMA-TIC. Actes d *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, 8(1), 281–303.
- Elalouf M.-L. (dir) (2005). *Écrire entre 10 et 14 ans. Un corpus, des analyses, des repères pour la formation*, SCérén, CRDP de Versailles.
- Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. Journal of educational Psychology, 80(4), 437.
- Santiago Oriola C. (1998). Système vocal interactif pour l'apprentissage des langues. La synthèse de la parole au service de la dictée. Thèse. Toulouse III.
- Smith, N., & McEnery, T. (1998). Issues in Transcribing a Corpus of Children's Hanwritten Projects. Literary and linguistic computing, 13(4), 217-225.