Tracking the main characters in narratives written by pupils with referential continuity ties annotation

L.M. Ho-Dac, J. Rebeyrolle

(co)Referential expressions as discourse "markers" (ĐMs) that contribute to discourse organisation by (indicating how to) connecting discourse units with referential continuity ties











Project [Doquet et al., 2017]









Describe writing skills evolution through all the education grades (elementary school, high school and university)

- Make available a large resource of ecological handwritten school works for linguistics (corpus) analyses and NLP
- First studies on
 - Spelling and verbal morphology
 - Coherence and cohesion ties [Halliday and Hasan, 1976], esp. Refential continuities i.e. referential cohesive ties that create continuity, texture acc. to reference
- Provide a pedagogical tool for teaching writing skills by using real-life productions of other students and emphasize problems and doubts encountered by students: http://e-calm.huma-num.fr/



Project [Doquet et al., 2017]





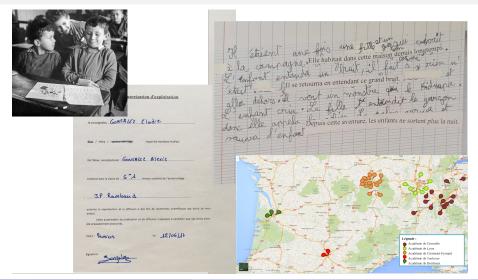




Describe writing skills evolution through all the education grades (elementary school, high school and university)

- Make available a large resource of ecological handwritten school works for linguistics (corpus) analyses and NLP: The E-Calm resource
- First studies on
 - Spelling and verbal morphology
 - Coherence and cohesion ties [Halliday and Hasan, 1976], esp. Refential continuities i.e. referential cohesive ties that create continuity, texture acc. to reference
- Provide a pedagogical tool for teaching writing skills by using real-life productions of other students and emphasize problems and doubts encountered by students: http://e-calm.huma-num.fr/

Collecting (meta)data and authorisation for CC BY-NC-ND 3.0



Encoding (TEI), Transcription, Spellchecking

Tle de 2017-04-25 école publique en zone mixte hors REP. EC-CM2-2016-SGLEA-D1-R24-V1 France Entre 300 et 400 élèves Le monstre Dans le nord de la France se trouvait un paysan |dont sa grand-mêre qui avait une maison blanche. Elle habitait dans cette maison depuis longtemps. Un Das le nort de la Brance en tromost en resport qui avoiture iour, le paysan vue un buisson bouger (pendant une nuit, il ne voulait en savoir motion blanks Elle habitait dans cette maison depuis longtemps. Un plus, alors, il sortit dehors avec un couteau let une torche. Il regarda dans le buisson four the regions were up brigan brought at at worldist on service et il vi du sang et des pas derrière la cloture, il sorti dehors et suivi les pas. Il sarretai dans un buisson à l'interieur de la forêt fine, alone, it sorted before and in onternal negacity done to income all it wisher the same at deep rate deriver it a circum the time to the same at the same at a circum the same at a circum the same to the same at a circum the same to the same at a circum the same at Il regarda dans le buisson et vue un ours mort. Il reparti alors. Tout à coup un grand bruit le fit sursoter. Il se retourna en entendant ce grand bruit. Des nos. Il squarti dono un enigor à l'interient te la sort -dont ce bruit et il vue un monstre dont ses doit griffes dégoutxx deont, il xx vert gluant et la xx : dans mais xx inodaxxt Il regards for be brigan at me un over most. Il reportison inimaginable. le cri en vovant le monstre il le tuer. Your a come on grand bruth get weresting. Il se retourna en entendant ce grand b Depuis cette aventure, les enfants ne sortent plus la nuit. -dank in bout at it was on monets hat an exit origins diagonation be part, il deput went glacent et prestries dang mois even unadout interreginable. It in en voyant le monte moit il le tre . Depuis cette aventure, les enfants ne sortent plus la nuit. - Il est difficile de comprendre l'écriture de l'élève. Si vous vovez une erreur dans la transcription, merci d'envover un mail à

http://redac.univ-tlse2.fr/corpus/resolco/copies/html normalise/

EC-CM2-2016-SGLEA-D1-R24-V1 N.html

- Ligne 1 et 3 : les ajouts ont été effectués au dessus de la ligne à

- Ligne 4 et 10 : les ajouts ont été effectués au dessous de la ligne à

l'aide d'une sorte de flèche.

l'aide d'une sorte de flèche.

hodac@univ-tlse2.fr

E-CALM ressource quantitative overview

http://e-calm.huma-num.fr

Grade	#Texts	#Words	#Words/Textes	Corpus
Primary	3,133	330,214	105	[E][R][S]
Middle school	555	115,866	209	[E][R]
High school & Univ.	607	671,056	1,106	[E][R][LA]
Total	4,295	1,117,136	260	[E][R][S][LA]

- Diversity of instructions (teachers usual pratices) EcriScol [E]¹ and the Advanced Literacy [LA]²
- Texts produced in reply to a specific instruction scientifically designed :
 - Scoledit [S]³
 - ResolCo [R]⁴

^{1.} http://www.univ-paris3.fr/ecriscol

^{2.} https://www.ortolang.fr/market/corpora/litteracieavancee

http://www.scoledit.org/scoledit

^{4.} http://redac.univ-tlse2.fr/corpus/resolco.html

E-CALM ressource quantitative overview

http://e-calm.huma-num.fr

Grade	#Texts	#Words	#Words/Textes	Corpus
Primary	3,133	330,214	105	[E][R][S]
Middle school	555	115,866	209	[E][R]
High school & Univ.	607	671,056	1,106	[E][R][LA]
Total	4,295	1,117,136	260	[E][R][S][LA]

- Diversity of instructions (teachers usual pratices) EcriScol [E]¹ and the Advanced Literacy [LA]²
- Texts produced in reply to a specific instruction scientifically designed :
 - Scoledit [S]³
 - ResolCo [R]⁴

^{1.} http://www.univ-paris3.fr/ecriscol

^{2.} https://www.ortolang.fr/market/corpora/litteracieavancee

^{3.} http://www.scoledit.org/scoledit

^{4.} http://redac.univ-tlse2.fr/corpus/resolco.html

The $\operatorname{SCOLEDIT}$ Corpus



Scoledit

A typical writing schoolwork task









Here are 4 characters. Choose one or two characters and tell a story. Circle the character(s) you have chosen.

A simple task that asks to introduce characters

that could become more complex acc. to

- the number of selected characters
- the genre of the selected characters

The RESOLCO Corpus Corpus (Garcia-Debanc et al., 2017)



A complex instruction designed to tackle discourse coherence

Tell a story by inserting in it in the same order the three following predefined sentences:

- She was living in this house for a long time.
- **S2** He turned around when he heard this loud noise.
- **S3** Since this adventure, (the)* children no longer go out at night.

Cut and glue the strips or copy each sentence, taking care to not change any single letter.

A complex task that implicitly asks to resolve a series of cohesion problems

- Who is She and He? Are they children? Are they the children of \$3?
- How to make accessible and salient these characters when gluing the strip?
- How introduce this house or this great noise?

^{*}In French, determiner is required whether the NP is specific or generic

Describing the acquisition of "discourse organisation" skills

Focus on referential cohesion ties

- Tracking the mentions to the main characters (topics/participants continuity [Givón, 1983])
- The Text-world Theory [Werth, 1999]: The normal procedure for introducing information is [...] a reference is first established (typically with an indefinite NP), and then is subsequently maintained (with some kind of anaphora). This essentially simple picture is complicated by three things: (i) evoked knowledge differs enormously, both qualitatively and quantitatively, from one person to another; (ii) if an entity is manifest in the environment, then it is often unnecessary to go through the whole reference procedure; and (iii) as we have seen, information can be introduced non-standardly, by way of accommodation.

Describing the acquisition of "discourse organisation" skills

Focus on referential cohesion ties

- Tracking the mentions to the main characters (topics/participants continuity [Givón, 1983])
- The Text-world Theory [Werth, 1999]: The normal procedure for introducing information is [...] complicated by three things: (i) evoked knowledge differs enormously [...] from one person to another; (ii) [...] it is often unnecessary to go through the whole reference procedure; and (iii) [...] information can be introduced non-standardly [...]

Questions

- What kind of procedure the students will follow?
- Are there "normal" (or more frequent) procedures acc. to education grades?

The E-CALM referential continuities annotation process

Annotation of all referring expressions related to the characters imposed by the instruction (not exactly referred to) : continuity vs. coreference

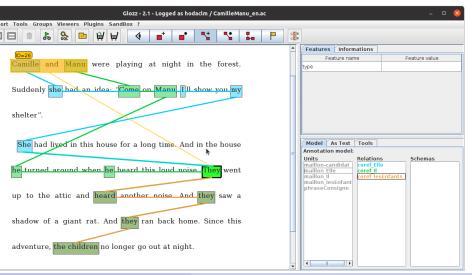
- \bullet Referring expressions delimitation as defined in $\operatorname{DEMOCRAT}$ guidelines [Landragin, 2015]
- Considering transition from individual reference to group reference : "group" mentions
 - "Camille and Manu were playing ... They went up to..."
- Considering transitions from narrative to direct speech : 1st and 2nd pronouns
 - "Camille and Manu were playing. She said: «Come on Manu. I will ...»"
- Expert annotators only

90% annotations strictly similar* based on $100~{\rm texts}$ annotated by $3~{\rm coders}$

(*same number of mentions, delimitation, and group feature)

Example from the ResolCo corpus (4th grade)

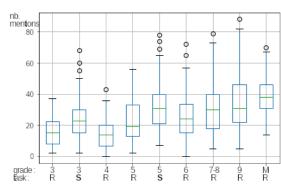
http://redac.univ-tlse2.fr/corpus/resolco



Overview of the annotated E-CALM subpart

	nbTexts	1 char.	2 char.	3 char.	4 char.	nbMentions
SCOLEDIT	348	21	273	41	13	9,465
ResolCo	383	6	14	362	0	9,935
E-Calm	731	27	287	403	13	19,400

- Preferred choice in **S**coledit texts: 2 characters
- The higher the grade, the more the number of mentions
- The easier the task (S), the more the number of mentions
- A frequent use of "group" mentions: 45% S texts, 75% ResolCo texts



What kind of procedure the students will follow?

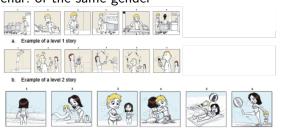
What referential marker for Introduction, Maintain, Shift

Using [Sandoz et al., 2023] study as a guide, a kind of benchmark
 Mélanie Sandoz, Katia Iglesias, Amélie M. Achim & Marion Fossard
 (2023) The contribution of discursive and cognitive factors in
 referential choices made by elderly people during a narrative task,
 Aging, Neuropsychology, and Cognition,

DOI:10.1080/13825585.2022.2150141

[Sandoz et al., 2023] study

- Oral interactions ("Tell successively each of the stories depicted in the nine sequences
 to your addressee which has the same set of images but in a random order and has to
 recreate the story told")
- 3 level of complexity : 1 character, 2 char. of different genders, 2 char. of the same gender



- 3 "discourse" stages : introduction, maintain, shift
- 3 types of referential markers : **IN**definite NP, **PR**onouns (clitic and zero pro.) and **D**+ NP (other kind of NP and stressed pronouns)

Example of a level 3 story

[Sandoz et al., 2023] results

No surprises

"results showed a significant effect of the discourse stages, [...]. Indeed, production of indefinite markers was favored when introducing a new character, while pronouns were mainly used to maintain a character in focus. For the shift of the character who is in focus, participants produced preferentially definite markers, [...].

An effect of the age

"[...] we also found that the effect of the discourse stage was modulated by age, suggesting that increasing age has an effect on referential choices. More particularly, older participants tended to produce less definite markers and more pronouns at the shift stage, compared to the other stages."

An effect of the number of characters but no effect of their gender

"A higher probability of producing **definite markers at the maintaining stage for the levels 2 and 3**, than at the level 1 [...]. No significant interaction was observed for the effect of gender."

Confronting E-CALM data to [Sandoz et al., 2023]

Differences with [Sandoz et al., 2023]

- Corpus vs. Experimental Data
- Written texts produced in reply to two tasks with another kind of level of complexity
- More than 2 characters with no control on their gender but
 - no gender effect acc. to [Sandoz et al., 2023]
 - the majority of Scoledit texts have only 2 characters
- No marking of the different stages
- More diverse referential markers (proper nouns, 1st/2nd pers. pronouns) + plural ("group reference" mentions)

E-CALM annotated data processing decisions

Discourse stages location

- intro First mention of the characters of the prompt (she, he, the witch, the wolf, the robot, the cat)
- maint Succession of mentions to the same character without interruption by other characters (of the prompt)
 - shift Mention of an already introduced character after the mention of another character, even if the mention is a "group" pronoun:
 - There is a witch [...]. A wolf [...]. **They** [...] Shift for the witch, Maintain for the wolf (considered as 2 mentions)

referential/task complexity (gender not taken into account)

- level1 Scoledit with 1 character
- level2 Scoledit with 2 characters
- level3 Resolco (always 2 or 3 characters, depending on the "Children" status)

E-CALM annotated data processing decisions

Referential markers = all annotated mentions (subject position or not)

IN indefinite NP

PR clitic pronouns, relative pronouns, zero anaphora and possessive determiner ^a

PN proper nouns

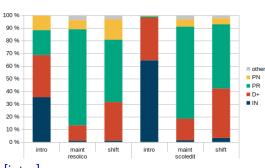
D+ all remaining mentions that contains at least one NOUN

Ref. marker	Dis			
	intro	total		
IN	742	133	149	1,024
D+	500	1663	2246	4,409
PR	154	8625	3149	11,928
PN	90	762	586	1,438
all	1491	11,607	6302	19,400

a. The witch.... Her cat...

Which ref. marker at each discourse stage acc. to the task

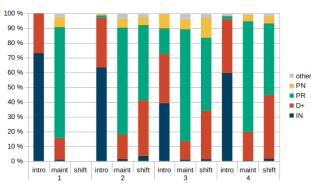
intro less IN in [R] (headings and narrative "style") maint mostly PR shift D+, PR or PN



Main effects of the task on the [intro]

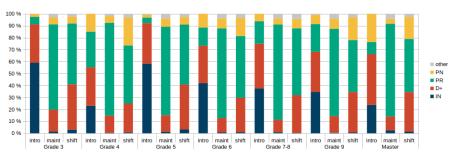
Which marker at each discourse stage acc. to the number of the characters

 No more D+ at the [maint] stage for more than 2 char. (cf. Sandoz et al. (2023))



No obvious effect excepted when 3 char. texts (i.e. RESOLCO) for [intro]

Which marker at each discourse stage acc. to the grade?



intro The higher the grade, the less the proportion of IN in [intro] (more PN or D+)

maint No variation acc. to the grade

shift Variations that required more qualitative analyses

Conclusion

- Results in the line with the literature and [Sandoz et al., 2023]
- First results on the acquisition of writing skills for building cohesion and producing coherent texts
- Q : What kind of procedure the students will follow?
- Q : Are there "normal" (or more frequent) procedures acc. to education grades?
- A : The same than (young) adults : [intro] with IN or PN, |maint] with PR, shift with D+ or PN

A resource waiting to be investigated

- Sequences of mentions in these referential continuities
- Strategies for managing several referents (step by step, all together, etc.)[Ho-Dac et al., 2021]
- Specific study on the transitions from individual reference to group reference ("group" feature)

A resource waiting to be further investigated

Correlation with the other annotation layers

- Revisions (e.g. erasures and additions during the writing process)
- Misspellings
- Teacher comments
- Rhetorical structure

New insights for studying discourse organization and written skills acquisition

- Is there a correlation between error rates and discourse organisation?
- Is there a correlation between revisions and discourse organisation?

Some Examples

- 6th grade text (R), minimalist ref. continuities http://redac.univ-tlse2.fr/corpus/ resolco/copies/html_normalise/CO-6e-2016-VTAC603-D1-R14-V1_N.html
- 9th grade text (R), she [intro] with PR (the prompt's first sentence):
 http://redac.univ-tlse2.fr/corpus/resolco/copies/html_normalise/CO-3e-2018-VTAC305-D1-R10-V1_N.html
- 9th grade text (R), he [intro] with PR (the prompt's second sentence): http://redac.univ-tlse2.fr/corpus/resolco/copies/html_normalise/ EC-CM1-2016-VRX-D1-R6-V1_N.html
- 3rd grade text (S) with a heading introducing characters with D+: http://scoledit.org/scoledition/production.php?id=1101&niv=CE2
- 3rd grade text (S) with 4 characters and a D+ for introducing the wolf http://scoledit.org/scoledition/production.php?id=1197&niv=CE2
- 5th grade text (S) with a lot of "group" mentions: http://scoledit.org/scoledition/production.php?id=828&niv=CM2
- Master narrative "style" text, she [intro] with PN: http://redac.univ-tlse2.fr/corpus/resolco/copies/html_normalise/UN-M2-2021-UCL-D1-R18-V1_N.html



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 Ide, N. and Pustejovsky, J., editors, Handbook of Linguistic Annotation, pages 1241–1264. Springer Netherlands, Dordrecht.



Doguet, C., David, J., Fleury, S., and (Eds) (2017). 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.



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

Corpus, 16.



Givón, T. (1983). Topic continuity in discourse: an introduction.

In T.Givon, editor, Topic continuity in discourse: a quantitative cross-language study, pages 1-42. John Benjamins: Amsterdam/Philadelphia.



Halliday, M. and Hasan, R. (1976).

Cohesion in English.

Longman: London.



Ho-Dac, L.-M., Garcia-Debanc, C., and Rebeyrolle, J. (2021).

La continuité référentielle dans le corpus RÉSOLCO : méthode d'annotation et premières analyses.

Langue française, 211(3):99–114.



Landragin, F. (2015).

Description, modélisation et détection automatique des chaînes de référence (democrat).

Bulletin de l'AFIA, 92 :11–15.



Sandoz, M., Iglesias, K., Achim, A. M., and Fossard, M. (2023). The contribution of discursive and cognitive factors in referential choices made by elderly people during a narrative task.

Aging, Neuropsychology, and Cognition, 0(0):1–22.



Werth, P. (1999).
Text worlds: Representing conceptual space in discourse.
Longman: London.

The $\operatorname{RESOLCo}$ task : a series of cohesion problems

Example of primary source collected at Grade 4



ognesse
Il étaiume fois une petite fiell qui s'apelle Baba Yaga
elle habitait dans cette maison antoi depuis long temps.

Deux petit enfan san parti lui rendre visite Ila et
Vax, Ilse sont retoura en entendant ce grand bruit
d'ognése ensuite les enfants ont courvét Depui cette
aventure les enfants me sortent plus la nuit.

The $\operatorname{RESOLCo}$ task : a series of cohesion problems

Example of primary source collected at Grade 4

01/02/2019
qu'arrive t.l.
aux enfants? (
que fait l'ogresse

Once upon a time there was an ogress called Baba Yaga. She was living in this house for a long time.

Two small children have gone to visit her. Lola and Max. He turned around when he heard this ogress great noise. Then the children ran and since this adventure, the children no longer go out at night.

Teacher comments

- Who is "He"?
- What happens to the children?
- What is the ogress doing?

Deletion, Addition as a trace of hesitations

(ResolCo only, Stanza UD POS tagging) Very few on connectives

	Revisions	Tokens	% all Revisions	% this POS
allPOS	1944	102192	100 %	2 %
VERB	374	14305	19 %	3 %
NOUN	328	17408	17 %	2 %
DET	233	15015	12 %	2 %
PRON	206	11887	11 %	2 %
PUNCT	178	10763	9 %	2 %
ADP	157	11392	8 %	1 %
PROPN	91	2614	5 %	3 %
ADV	90	6289	5 %	1 %
ADJ	89	4029	5 %	2 %
CCONJ	74	3264	4 %	2 %
AUX	74	2961	4 %	2 %
SCONJ	23	1265	1 %	2 %

Most frequent revised DM

et, longtemps, mais, depuis, puis, comme, jusque, soudain, après, ...

The E-Calm referential continuities annotation process

Context and previous work for French



ANNODIS [Asher et al., 2017]

- A diversified corpus of written French texts annotated with various discourse structures including "Topical Chains"
- Top-down annotation: from multi-level structures to cues
- Focus on topical chains i.e. chains involving the main "characters"
- Assisted annotation : pretagging + Glozz
- naive coders



ANR Democrat

- A diversified corpus of written French texts annotated with all referring expressions
- Bottom-up annotation : from all ref. expressions to chains
- Annotation of all ref. expressions
- Assisted annotation : TXM-URS + building chains automatically
- Expert coders

The E-Calm referential continuities annotation process

Context and previous work for French



ANNODIS [Asher et al., 2017]

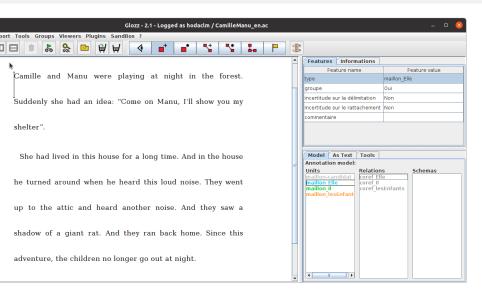
- A diversified corpus of written French texts annotated with various discourse structures including "Topical Chains"
- Top-down annotation: from multi-level structures to cues
- Focus on topical chains i.e. chains involving the main "characters"
- Assisted annotation : pretagging + Glozz
- naive coders

dem crat DEMOCRAT [Landragin, 2015]

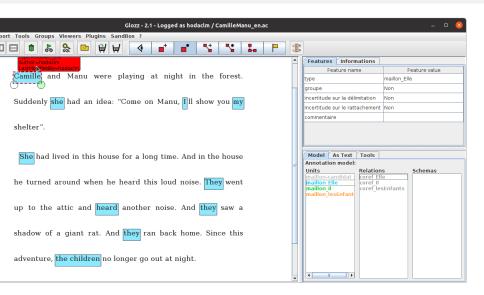
ANR Democrat

- A diversified corpus of written French texts annotated with all referring expressions
- Bottom-up annotation : from all ref. expressions to chains
- Annotation of all ref. expressions
- Assisted annotation : TXM-URS + building chains automatically
- Expert coders

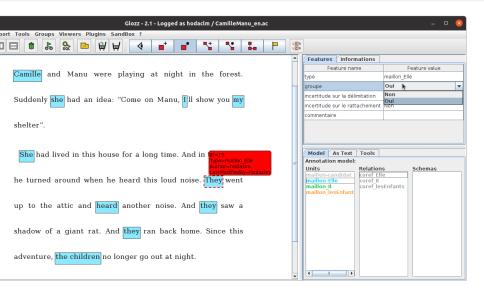
Computer Assisted Annotation



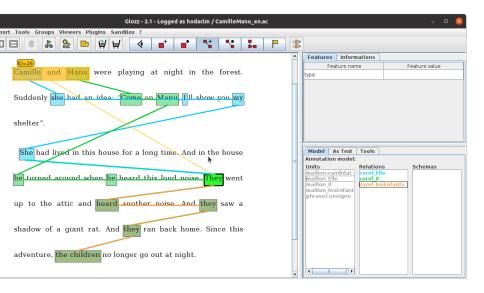
Each referent, one after the other



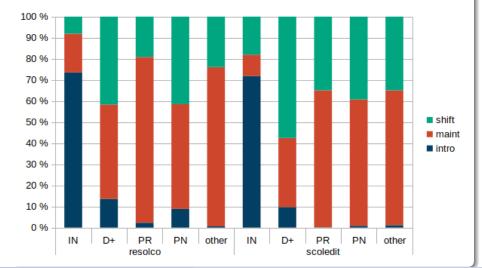
Including transitions from individual to group



Linking mentions to explicit ref. continuities



Which discourse stage for each Ref. Marker acc. to the task?



Which discourse stage for each Ref. Marker acc. to the grade?

