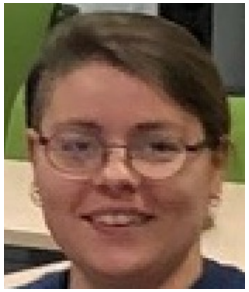
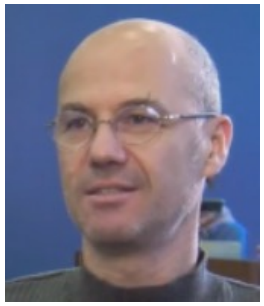
Grammatical profiling with UD annotation (WiP)

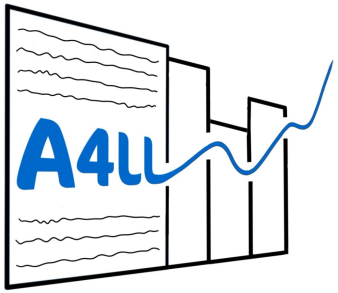
Nicolas Ballier (Université Paris Cité, CLILLAC-ARP and LLF)
Joint Research
with Cyriel Mallart and
Thomas Gaillat
(University of Rennes,
LIDILE lab)

Nicolas Ballier (UPCité, France) Joint Research with Cyriel Mallart and Thomas Gaillat (Rennes, France)

Grammatical profiling with UD annotation (WiP)







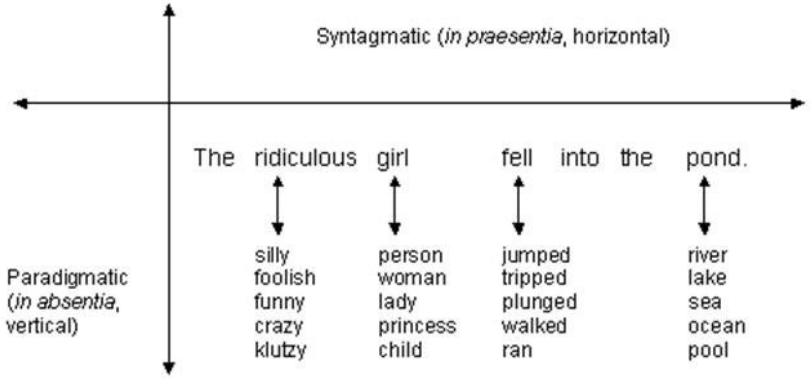
Analytics for Language Learning (Rennes)

Deep learning for language assessment (UPCité/KCL)

UD annotation as a window for grammatical profiles

Modelling stages of learner competence:

- UD for syntagmatic axis (proxy for complexity)
- UD for paradigmatic axis (micro-system approach)



From http://courses.nus.edu.sg/course/elltankw/history/Vocab/B.htm

Syntagmatic axis

Treebanks in UD ->

- Extract regularities in learner production
- Query syntax: GREW-match

http://universal.grew.fr/?corpus=UD_English-GUM@2.10

Extraction tool (GRE) Grammatical rule extraction
https://github.com/santiagohy/grammar-rules-extraction

Queries for error candidates

- Dependency relation labels : Parataxis

pattern { GOV -[parataxis]-> DEP }

- Compound

pattern { GOV -[compound]-> DEP }

- Combining features : (*the admission**s** committee*)

```
pattern {
  DEP [Number=Plur];
  GOV -[compound]-> DEP;
}
```

Candidates for metrics

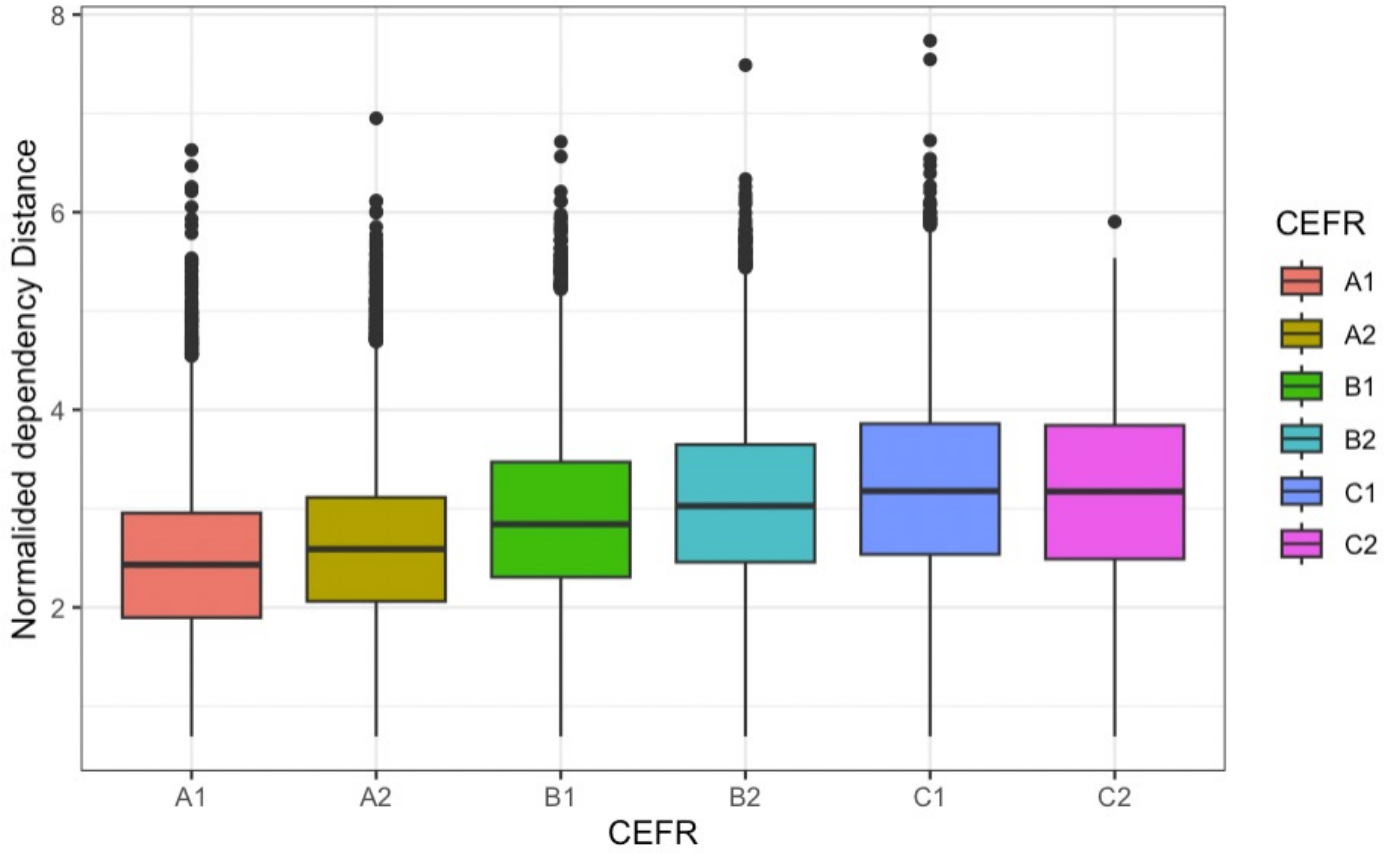
Text number	Sentence number	Word order	Word	POS	Governor order	Governor	Dependency relation	Dependency distance
t10	s7	1	She	PRP	3	does	nsubj	2
t10	s7	2	always	RB	3	does	advmod	1
t10	s7	3	does	VBZ	3	does	root	0
t10	s7	4	homework	NN	3	does	dobj	1
t10	s7	5	on	IN	3	does	prep	2
t10	s7	6	the	DT	7	weekend	det	1
t10	s7	7	weekend	NN	5	on	pobj	2
t10	s7	8	.	.	3	does	punct	5

(Ouyang, 2020)

Relevance of UD distance (normalised or not)

EFCAMDAT Spanish component

Spanish component of the EFCAMDAT



Deep Learning for Language Assessment (aims)

- Metrics and features
- Data collected in Rennes and on Prolific
- Keylog data and CEFR prediction

A4LL Scientific challenges

Analytics for LL : Create a language-learning analytics system

Three main research questions to uncover some of the features of Interlanguage

i) what are the **language features** related to specific **proficiency** levels?

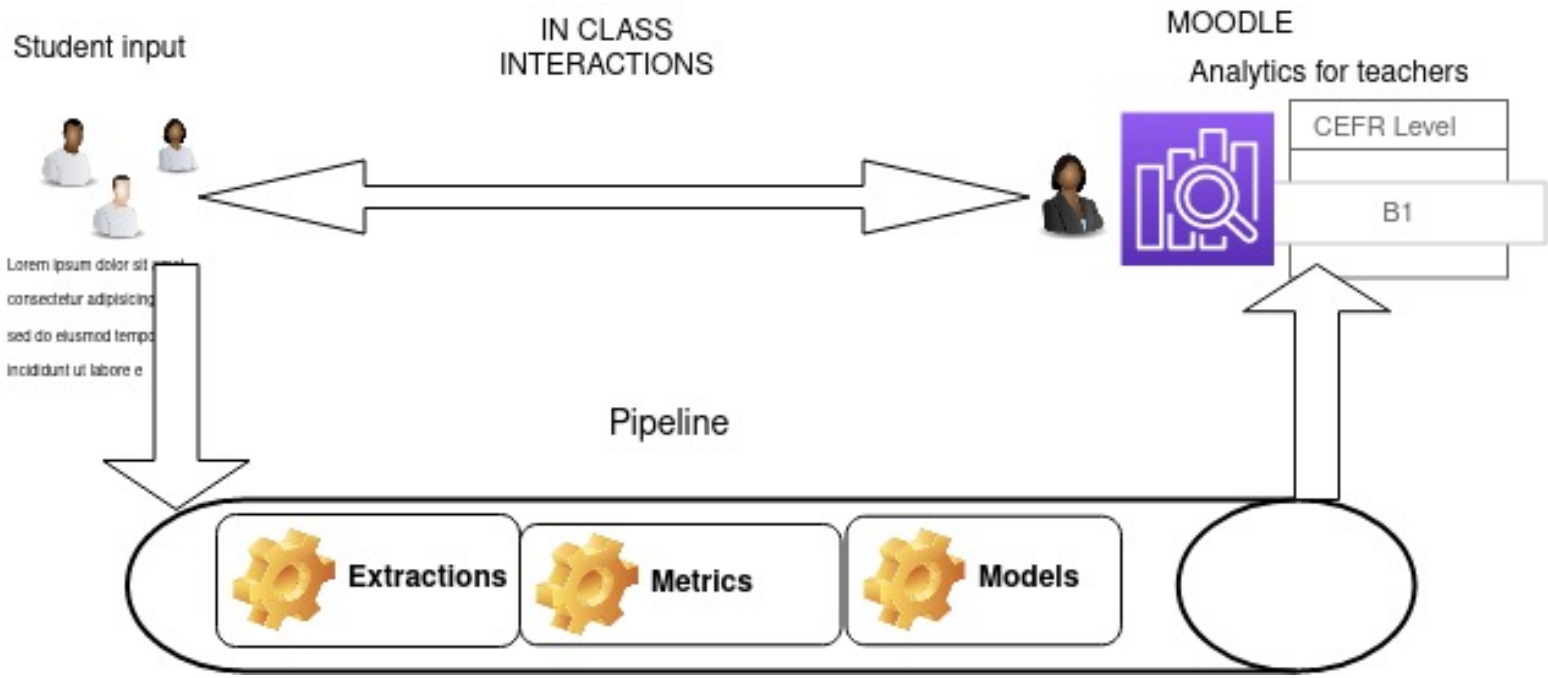
ii) how can these features be measured **automatically**?

iii) how can measures be converted into **meaningful analytics** for descriptive feedback and teaching decisions?

Our approach

1. Richly annotated L2 data (Rennes students)
2. Identifying and designing automatic measures in L2 writings
3. Modeling L2 writing & proficiency
4. Creating an interoperable data pipeline

The system


Student input
IN CLASS
INTERACTIONS
MOODLE
Analytics for teachers
CEFR Level
B1
Pipeline
Extractions
Metrics
Models

Metric processing tools

- Microsystem tool - in dev
- Collocation tool - in dev
- Error detection - to be dev
- Keylogs - in dev
- Syntactic complexity tool – to be adapted
- Cohesion tool - to be adapted

Micro-system concept

- Paradigmatic relations and functions
- Determination MS with articles A, THE and 0

1 "Ladies and Gentlemans, My flat was robbed the previous evening. In coming back at my home, I saw that the window was broken." (EFCAMDAT writing ID: 2498)

2 "What do you think about positive discrimination in **the*** companies?" (EFCAMDAT writing ID: 569744)

3 "Why the* gender's discrimination is still a problem in our society?" (EFCAMDAT writing ID: 579779)

Research Question

Which linguistic representation can be defined for a micro-system?

Which UD feature could be mobilised for the analysis?

Pipeline

1. **Annotation**: create_data frames of UD sentences :
   - Output: full CONLL-U annotated file
2. **Extraction**: GREW graph queries
   - Output: Set of forms making up a MS

- main node = target Microsystem word

CONLL-U sentence

sent_id = GUM_academic_huh-31

s_prominence = 4

s_type = decl

transition = establishment

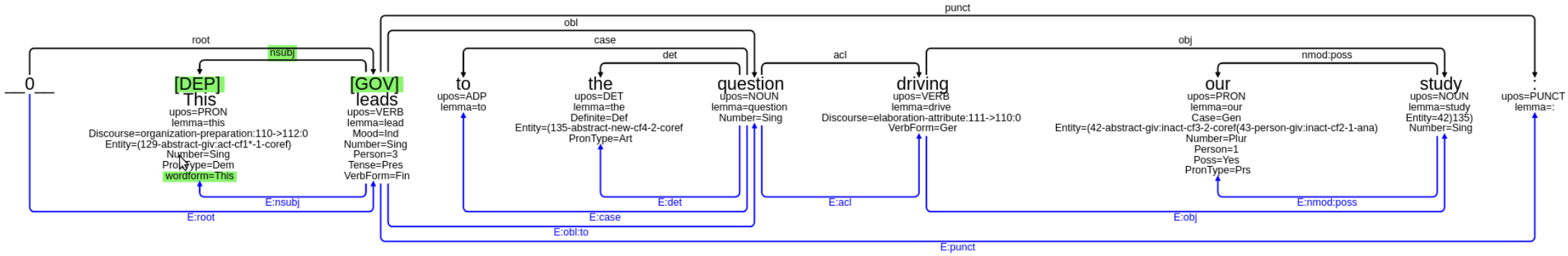
text = This leads to the question driving our study:

1	This	this	PRON	DT	Number=Sing|PronType=Dem	2	nsubj	2:nsubj	Discourse=organization-preparation:110->112:0|Entity=(129-abstract-giv:act-cf1*-1-coref)
2	leads	lead	VERB	VBZ	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	0	root	0:root	_
3	to	to	ADP	IN	_	5	case	5:case	_
4	the	the	DET	DT	Definite=Def|PronType=Art	5	det	5:det	Entity=(135-abstract-new-cf4-2-coref
5	question	question	NOUN	NN	Number=Sing	2	obl	2:obl:to	_
6	driving	drive	VERB	VBG	VerbForm=Ger	5	acl	5:acl	Discourse=elaboration-attribute:111->110:0
7	our	our	PRON	PRP$	Case=Gen|Number=Plur|Person=1|Poss=Yes|PronType=Prs	8	nmod:poss	8:nmod:poss	Entity=(42-abstract-giv:inact-cf3-2-coref(43-person-giv:inact-cf2-1-ana)
8	study	study	NOUN	NN	Number=Sing	6	obj	6:obj	Entity=42)135)|SpaceAfter=No
9	:	:	PUNCT	:	_	2	punct	2:punct	_

Extraction query

pattern {DEP [wordform="this"|"these"|"This"|"These"] ; GOV -[nsubj|obl|nsubj:pass|nmod|obj|nsubj:outer|conj|root]-> DEP; }

Searching and identifying “this” proform in the graph:



Micro-system data representation

- Features collected in a table (.CSV)
- Includes metadata in

nb_annees_L2 L1 Sejours_duree_semaines Sejours_frequence Lang_exposition L2 **Note_dialang_ecrit(CEFR)** Lecture_regularite autre_langue tache_ecrit Texte_etudiant Date_ajout pseudo

- Includes linguistic annotation

Proform: ([-5;+5] tokens, lemma, UPOS, morphological features), dependency_distance_to_root

Its Head: head_form head_lemma head_textform head_upos head_wordform head_xpos head_dependency_rel

Future work: operationalise co-occurrence restrictions

- Developing multi-node extraction
  - Which node as main element (GOV?)?
  - Adjacent slots of in the micro system (paradigm) as syntagmatic dimension of the learner production (* *was seen yesterday*)
- To capture POS or UD error patterns for micro-systems / articulate with repertoire of forms (see *WOULD* in plenary talk) & lexical routines

Case study

- Modelling CEFR according to occurrence of proforms in texts
- NLP4CALL 2023

Expected outcomes

- Identifying features of L2 developmental stages (profiles/interlanguage strata)
- Mapping stages to proficiency

- A MOODLE module for L2 Analytics with actionable visualizations (A4LL)

ACKNOWLEDGEMENTS

ANR (ANR-22-CE38-0015-01)
https://sites-recherche.univ-rennes2.fr/lidile/en/a4ll/

Université Paris Cité / King's College (London) funding Deep Learning for Language Assessment [DLLA] (PI for KCL: Helen Yannakoudakakis)

References

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. (2021). Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 1-17. doi:10.1017/S095834402100029X

Questions?

nicolas.ballier@u-paris.fr

THANK YOU !

DLLA Project

Micro-system paper