UNIVERSITÉ RENNES 2

LIDILE

L'interopérabilité des corpus pour la modélisation des dynamiques d'acquisition de langue seconde

T. Gaillat

Team: C. Mallart, N. Ballier, A. Simpkin, R. Venant, A. Faugère, B. Stearns, J. Yu Li, P. Lissón

A4LL Analytics for Language Learning

anr

A4LL’ s context

- Second Language (L2) teaching.
- Need: monitor linguistic profiles of learner productions
- Objective: Create a language-learning analytics system with visualisations

Project's Research Questions

1. What are the **language features** related to specific **proficiency** levels?
2. How can these features be measured **automatically**?
3. How can measures be converted into **meaningful analytics** for descriptive feedback and teaching decisions?

A4LL’s approach

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

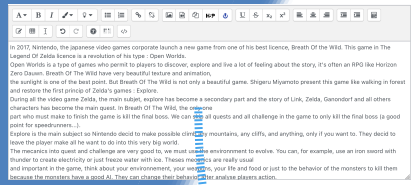
A4LL’s approach

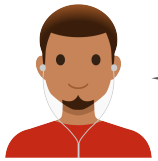
1. **Richly annotated L2 data (Rennes students)**
2. **Identifying and designing automatic measures in L2 writings**
3. **Modeling L2 writing & proficiency**
4. Defining visualisations
5. Creating and interoperable data pipeline

Methodological challenge

- Ensuring interoperability AND data quality
- Evaluating at different stages
  - How well are the data annotated and extracted?
  - How good are the models?

System overview



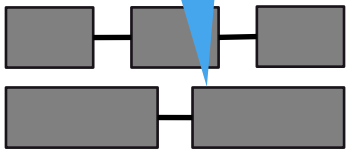


Mediated interactions

Class + Moodle

- Predictions
- Explanations
- Enriched texts

- Text flow
- Final production



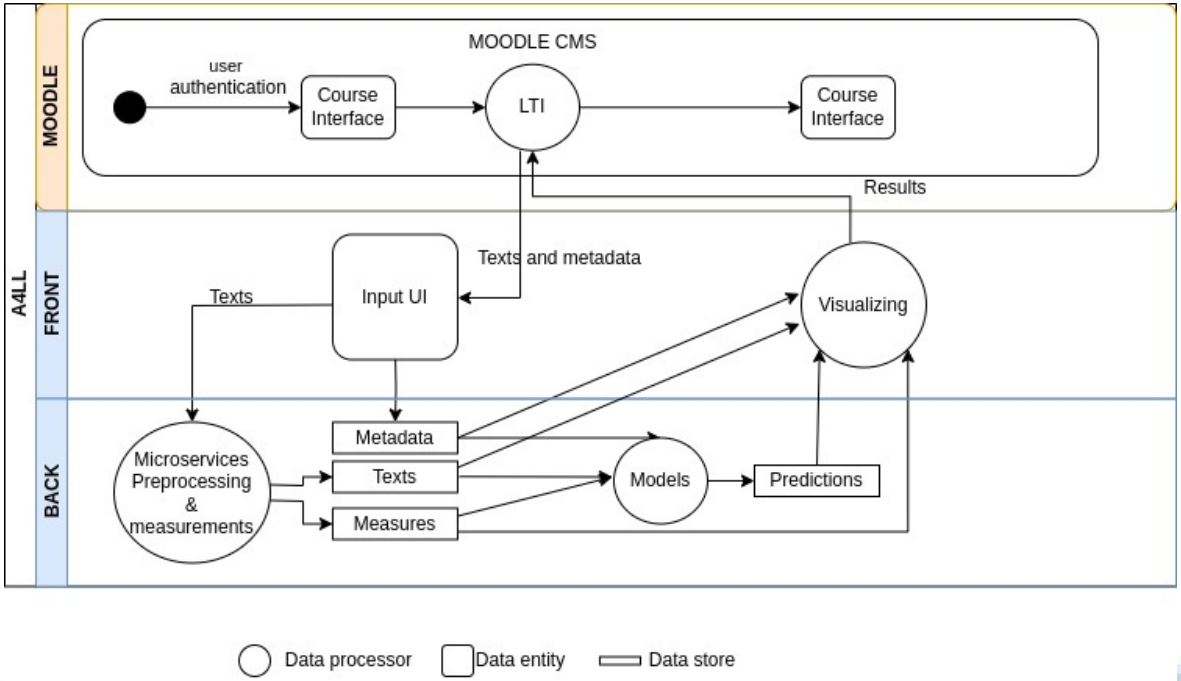
Linguistic feature extraction

Probabilistic models based on ling features

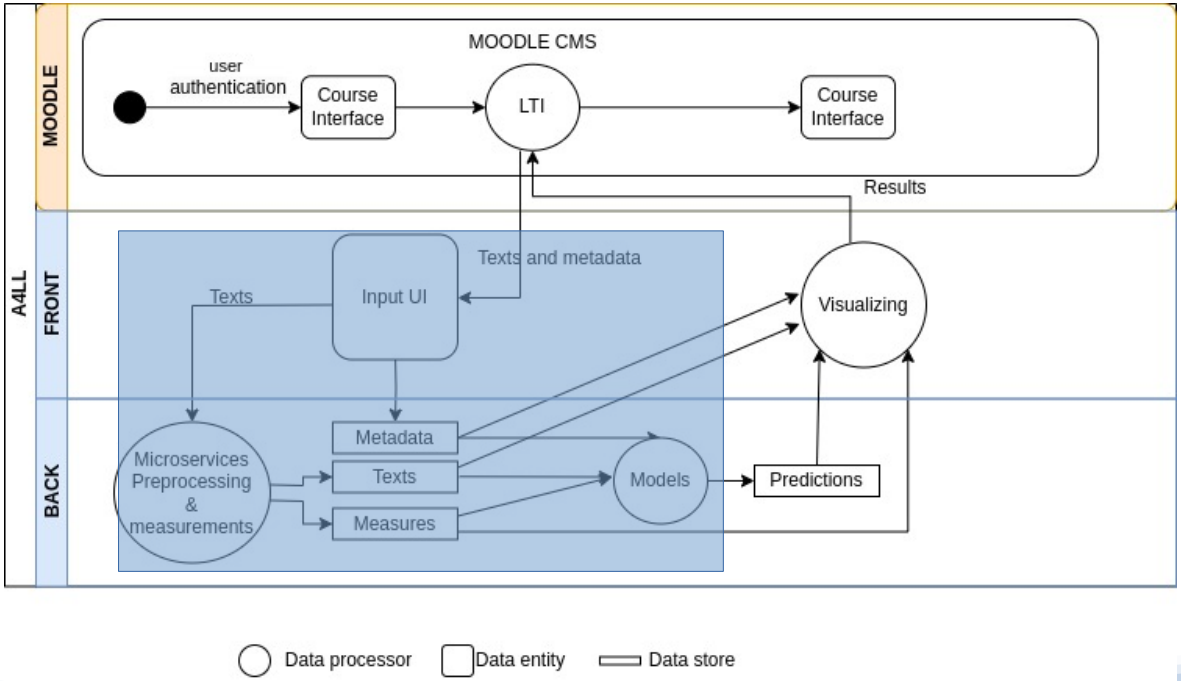
Probability-based metrics of ling features

Symbolic knowledge

Data interoperability



Data interoperability



Metric processing micro-services

- Microsystem tool - in dev
- Lexical profiling – in dev
- Collocation identification tool - in dev
- Keylogging behaviour - in dev
- Syntactic complexity tool with TAASSC (Kyle, 2016)
- Cohesion tool with TAACO (Kyle et al., 2018)

Approach: preparing and deploying

1. Modelling phase

   - Using corpus data to model learner output against proficiency

   - Validation of data **and** models

2. Deployment phase

   - Applying validated models to new incoming texts in a fully connected architecture

Approach: preparing and deploying

1. Modelling phase

   - Using corpus data to model learner output against proficiency

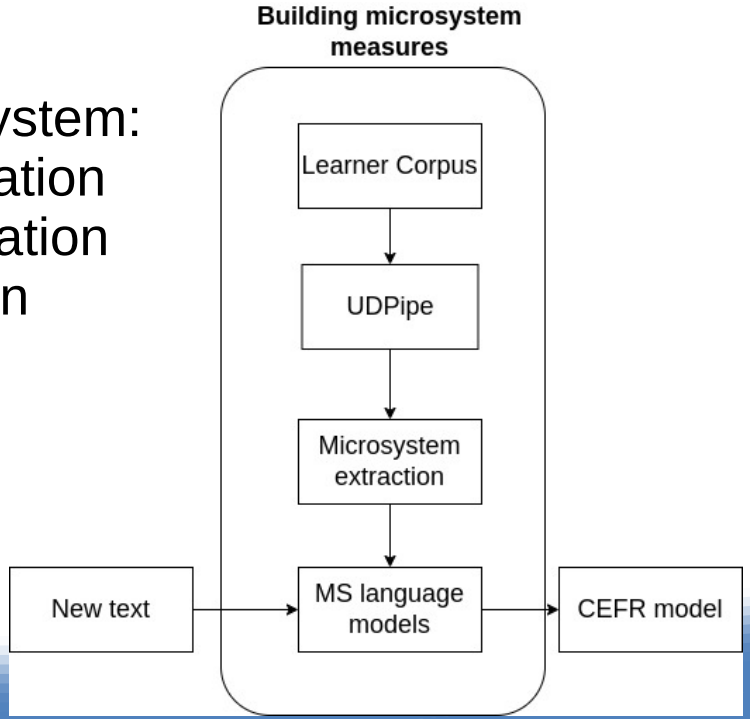
   - Validation of data **and** models

2. Production phase

   - Applying validated models to new incoming texts

The case of microsystem measures

A paradigmatic system:
This is a presentation
That is a presentation
It is a presentation



UDPipe microservice

1. Clean-up
2. UDPipe (Straka et al., 2016) & Universal Dependency annotation (de Marneffe et al., 2021)

Creates data frames of UD sentences :

- Input: texts
- Output: full CONLL-U annotated file

CONLL-U format

ID	token	lemma	UPOS	XPOS	Morphological features	Head index	Dependency relation with target	Morphological feats of head
1	This	this	PRON	DT	Number=Sing| PronType=Dem	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	_
5	question	question	NOUN	NN	Number=Sing	2	obl	_

Microsystem module

- IT THIS THAT proform microsystem
- Extract patterns with Grew-match
- IT without non-referential form
  - pattern { DEP [lemma="it"]; } without { GOV -[expl]-> DEP;}
- THAT
  - pattern { TH [lemma="that", upos=PRON, PronType=Dem] ; }
- THIS
  - pattern { TH [lemma="this", upos=PRON, PronType=Dem] ; }

minus_1_upos	Number	plus_2_upos	head_Case	pattern	plus_1_Tense	head_upos	minus_1_xpos	plus_2_lemma	head_SpaceAfter	plus_3_form	plus_2_textform	head_dependency_rel	head_xpos
NOUN	Sing	PRON		PRF_IT	Pres	VERB	NN	my		friends	my	nsubj	VBG
NOUN	Sing	PRON		PRF_IT	Pres	NOUN	NN	my		boyfriend	my	nsubj	NN
NOUN	Sing	PRON	Acc	PRF_IT	Pres	PRON	NN	I		in	me	nsubj	PRP
	Sing	PRON	Acc	PRF_THAT	Pres	PRON		I		but	me	nsubj	PRP
PUNCT	Sing	PRON		PRF_THAT	Pres	NOUN	,	my	No	boyfriend	my	nsubj	NN
SCONJ	Sing	ADJ		PRF_THAT	Past	NOUN	IN	sunny	No	day	sunny	nsubj	NN
	Sing	PRON		PRF_THIS	Pres	NOUN		I	No	in	I	nsubj	NN
	Sing	PRON		PRF_THIS	Pres	NOUN		my	No	friends	my	nsubj	NNS
	Sing	PRON		PRF_THIS	Pres	NOUN		my	No	family	my	nsubj	NN

Evaluation 1 & 2: annotation and extraction

- 165 randomly extracted evaluation set
  - External dataset (CELVA.Sp)
  - Manual annotation 2 experts
- Evaluation of annotation (Fleiss Kappa=0.95) + 1 consolidator
  - Automatic extraction
- Evaluation of extraction MS_IT_THIS_THAT

Precision	0.88
Recall	0.86
F1-score	0.86

Microsystem language model

- Outcome variable: Pattern
- Independent variables: morphosyntactic and dependency features
- Multinomial regression with Nnet library in R

Evaluation 3: model

- MS language model probabilities against CEFR levels
- Ordinal logistic regression

	Odds Ratio	95% CI	p-value
ms_it	0.985	0.984, 0.986	<0.001
ms_this	1.008	1.006, 1.009	<0.001
ms_that	1.020	1.019, 1.021	<0.001

Higher-level tasks

New texts: Internal & external test sets from different learner corpora

Task 1: Use MS probs as one of many other predictors in CEFR prediction model

Task 2: Use MS prob vectors to predict stability/instability of a microsystem

Evaluation T1: classification with CEFR classified texts

Evaluation T2: clustering and distance to nearest cluster

More data from UDPipe

- Collocations
- Keylogs
- Cohesion
- Fine-tuned BERT for vocabulary range assessment

Keylog measures


```
{
  "name_record": "keylogs-1645788314664",
  "measures": {
    "ratio_backspace_keys": 0.1556406685236769,
    "nb_backspace_sequences_longer_than_3": 64,
    "total_nb_bursts_any_kind": 204,
    "nb_p_bursts": 44,
    "nb_r_bursts": 79,
    "nb_revision_bursts": 80,
    "mean_nb_keystrokes_per_burst": 14.142857142857142,
    "mean_nb_keystrokes_per_p_burst": 25.954545454545453,
    "mean_nb_keystrokes_per_r_burst": 17.68354430379747,
    "mean_nb_keystrokes_per_revision_burst": 4.15,
    "proportion_p_burst": 0.21568627450980393,
    "proportion_r_burst": 0.3872549019607843,
    "proportion_revision_burst": 0.39215686274509803,
    "mean_time_p_burst": 13248.900000003252,
    "mean_time_r_burst": 6653.983544299874,
    "mean_time_revision_burst": 1367.2300000056625,
    "ratio_nb_p_burst_per_sentence": 3.3846153846153846,
    "ratio_nb_r_burst_per_sentence": 6.076923076923077,
    "ratio_nb_rev_burst_per_sentence": 6.153846153846154,
  }
}
```

```
{
  "name_record": "keylogs-1645788314664",
  "measures": {
    "mean_nb_characters_per_word": 4.706356311548792,
    "mean_length_pauses_before_word": 2618.1809870150064,
    "mean_length_pauses_after_word": 2779.3129588168517,
    "mean_nb_characters_per_word_in_p_bursts": 4.4862464183381086,
    "mean_length_pauses_before_word_in_p_bursts": 2357.590630370122,
    "mean_length_pauses_after_word_in_p_bursts": 2623.2672492796505,
    "mean_nb_characters_per_word_in_r_bursts": 4.933050658335193,
    "mean_length_pauses_before_word_in_r_bursts": 2841.953849582687,
    "mean_length_pauses_after_word_in_r_bursts": 2767.3436063385734,
    "mean_nb_characters_per_word_in_revision_bursts": 4.449740932642487,
    "mean_length_pauses_before_word_in_revision_bursts": 2521.532435227305,
    "mean_length_pauses_after_word_in_revision_bursts": 3399.244766837575,
    "nb_repairs": 137,
    "nb_revisions": 56,
    "nb_typos": 81,
  }
}
```

```
{
  "name_record": "keylogs-1645788314664",
  "measures": {
    "proportion_typos": 0.5912408759124088,
    "proportion_revisions": 0.40875912408759124,
    "nb_backspace_seq_longer_than_3": 45,
    "nb_backspace_seq_shorter_than_or_equal_3": 92,
    "ratio_backspace_seq_longer_than_3": 0.3284671532846715,
    "ratio_backspace_seq_shorter_than_or_equal_3": 0.6715328467153284,
    "mean_time_between_typing_first_char_and_first_backspace": 48239.7021897805,
    "nb_backspace_seq_longer_than_3_for_revision": 39,
    "nb_backspace_seq_shorter_than_or_equal_3_for_revision": 17,
    "ratio_backspace_seq_longer_than_3_for_revision": 0.6964285714285714,
    "ratio_backspace_seq_shorter_than_or_equal_3_for_revision": 0.30357142857142855,
    "mean_time_between_typing_first_char_and_first_backspace_for_revision": 53464.08928571428,
    "nb_backspace_seq_longer_than_3_for_typo": 6,
    "nb_backspace_seq_shorter_than_or_equal_3_for_typo": 75,
    "ratio_backspace_seq_longer_than_3_for_typo": 0.07407407407407407,
    "ratio_backspace_seq_shorter_than_or_equal_3_for_typo": 0.9259259259259259,
    "mean_time_between_typing_first_char_and_first_backspace_for_typo": 44627.7802469127
  }
}
```

Keylog measures


```
"name_record": "keylogs-1645788314664",
  "measures": {
    "mean_nb_characters_per_word": 4.706356311548792,
    "mean_length_pauses_before_word": 2618.1809870150064,
    "mean_length_pauses_after_word": 2779.3129588168517,
    "mean_nb_characters_per_word_in_p_bursts": 4.4862464183381086,
    "mean_length_pauses_before_word_in_p_bursts": 2357.590630370122,
    "mean_length_pauses_after_word_in_p_bursts": 2623.2672492796505,
    "mean_nb_characters_per_word_in_r_bursts": 4.933050658335193,
    "mean_length_pauses_before_word_in_r_bursts": 2841.953849582687,
    "mean_length_pauses_after_word_in_r_bursts": 2767.3436063385734,
    "mean_nb_characters_per_word_in_revision_bursts": 4.449740932642487,
    "mean_length_pauses_before_word_in_revision_bursts": 2521.532435227305,
    "mean_length_pauses_after_word_in_revision_bursts": 3399.244766837575,
    "nb_repairs": 137,
    "nb_revisions": 56,
    "nb_typos": 81,
```

Collocation measures

"text_ID": 7939,

“text”: I have been here 12 days, so I am very miss my little girl.,

"sent_id": 6,

"V": "miss",

"V_deprel": "acl",

"V_feats": "{'VerbForm': 'Inf'}",

"N": "girl",

"N_deprel": "obj",

"N_feats": "{'Number': 'Sing'}"

Candidate’s values calculated from actual texts and from BNC

'chi_sq': 5.666819082456944,

'dice': 0.3333333333333333,

'fisher': 0,

'jaccard': 0.19999999999999998,
'likelihood_ratio': 1.5390106706312134,

'mi_like': 0.03703703703703703,
'phi_sq': 0.10303307422648988,
'pmi': 4.196397212803503,
'poisson_stirling': 1.0654657376011676,

'raw_freq': .006060606060606061,
'student_t': .5458584363247371

Deliverables

sites-recherche.univ-rennes2.fr/lidile/en/a4ll/

- CELVA.Sp L2 corpus > Huma-num Nakala
- MOODLE corpus collection module > Gitlab
- Data Management Plan > Opidor > A4LL
- Python programs > Gitlab LIDILE

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, 34(2). https://doi.org/10.1017/S095834402100029X

•Dougiamas, M., & Taylor, P. (2003). Moodle: Using Learning Communities to Create an Open Source Course Management System. Proceedings of the EDMEDIA 2003 Conference, Honolulu, Hawaii, 171–178. https://www.learntechlib.org/primary/p/13739/

•Ellis, R. (1994). The Study of Second Language Acquisition. Oxford University Press.

•Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In R. T. Miller, K. I. Martin, C. M. Eddington, A. Henery, N. Miguel, A. Tseng, A. Tuninetti, & D. Walter (Eds.), Proceedings of the 31st Second Language Research Forum. Cascadilla Press.

•Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication [Dissertation, Georgia State University]. https://scholarworks.gsu.edu/alesl_diss/35

•Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. Behavior Research Methods, 50(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

•Py, B. (1980). Quelques réflexions sur la notion d'interlangue. Revue Tranel (Travaux Neuchâtelois de Linguistique), 1, 31–54.

•Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 4290–4297. https://aclanthology.org/L16-1680

Merci

- More info on:
- sites-recherche.univ-rennes2.fr/lidile/en/a4ll/