Linguistic interoperability within a unified architecture

Thomas Gaillat, Cyrielle Mallart, Andrew J. Simpkin, Rémi Venant, Nicolas Ballier, Jen-Yu Li, Bernardo Stearns

Analytics for Language Learning

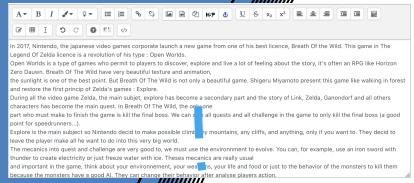
Linguistic interoperability within a unified architecture

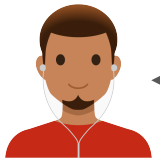
Mallart, Cyriel, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen-Yu Li, and Thomas Gaillat.

Context

- Second Language (L2) teaching
- Correction of writings = intuitive + manual => slow, not frequent, no objective overview
- Goal: application for linguistic diagnosis of L2 writings by teachers
- Project : automatic extraction of L2 linguistic features & generation of linguistic analytics

System overview



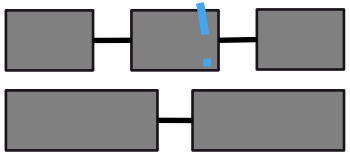


Mediated interactions

Class + Moodle

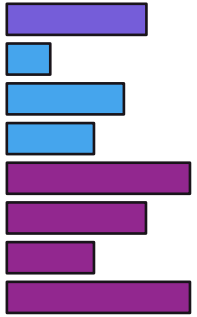
- Predictions
- Explanations
- Enriched texts

- Text flow
- Final production



Linguistic feature extraction

Probabilistic models based on ling features



Probability-based metrics of ling features

Symbolic knowledge

Our 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

Challenge: Creating and interoperable data pipeline
RQs:

1. **How to have the same data representation for any corpus?**
2. **How to make data and models compatible?**
3. **How to organise the flow of data?**

RQ1 An interoperable data representation

All texts = same annotation scheme

- Universal Dependencies
- CoNLL-U format

Linguistic measures based on:
Pos, dependencies and morphosyntactic features

Measures stored in a Mongo DB

RQ1 Linguistic pre-processing microservices

The Universal Dependency annotation scheme (de Marneffe et al., 2021) with:

- UDPipe (Straka et al., 2016)
- Stanford Stanza (Qi et al., 2020)
- In Docker containers on Virtual Machine

RQ1 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	_

RQ1 Processing microservices

Different linguistic dimensions:

- Lexical complexity: sophistication and diversity
- Syntactic complexity: systemic and structural
- Cohesion
- Phraseology with “collocationness”
- Lexical semantic complexity

RQ1 Processing microservices

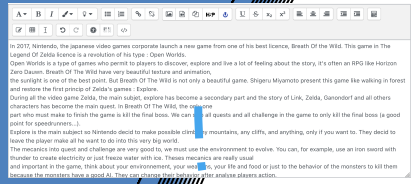
Text level measures:

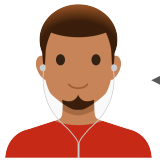
- Microsystem: probabilities of words belonging to same paradigm
- Lexical semantic complexity – probabilities of producing a word/word class at a certain proficiency level
- Collocation identification – probabilities of a V-N syntagm being a collocation
- Syntactic complexity with TAASSC (Kyle, 2016) – Frequency–based ratios of syntactic components
- Cohesion tool with TAACO (Kyle et al., 2018) - Frequency–based ratios of cohesive components
- Keylogging behaviour - Frequency–based ratios of typing behaviour components

RQ2 Interoperability between measures and models

- Models trained on texts, annotations and measures stored in the DB with specific features
- Models’ expected features selected by a data service module

RQ2 Model interoperability



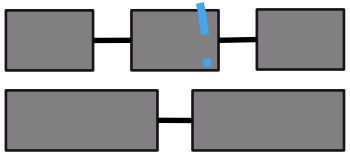


Mediated interactions

Class + Moodle

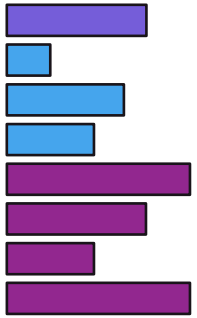
- Predictions
- Explanations
- Enriched texts

- Text flow
- Final production



Linguistic feature extraction

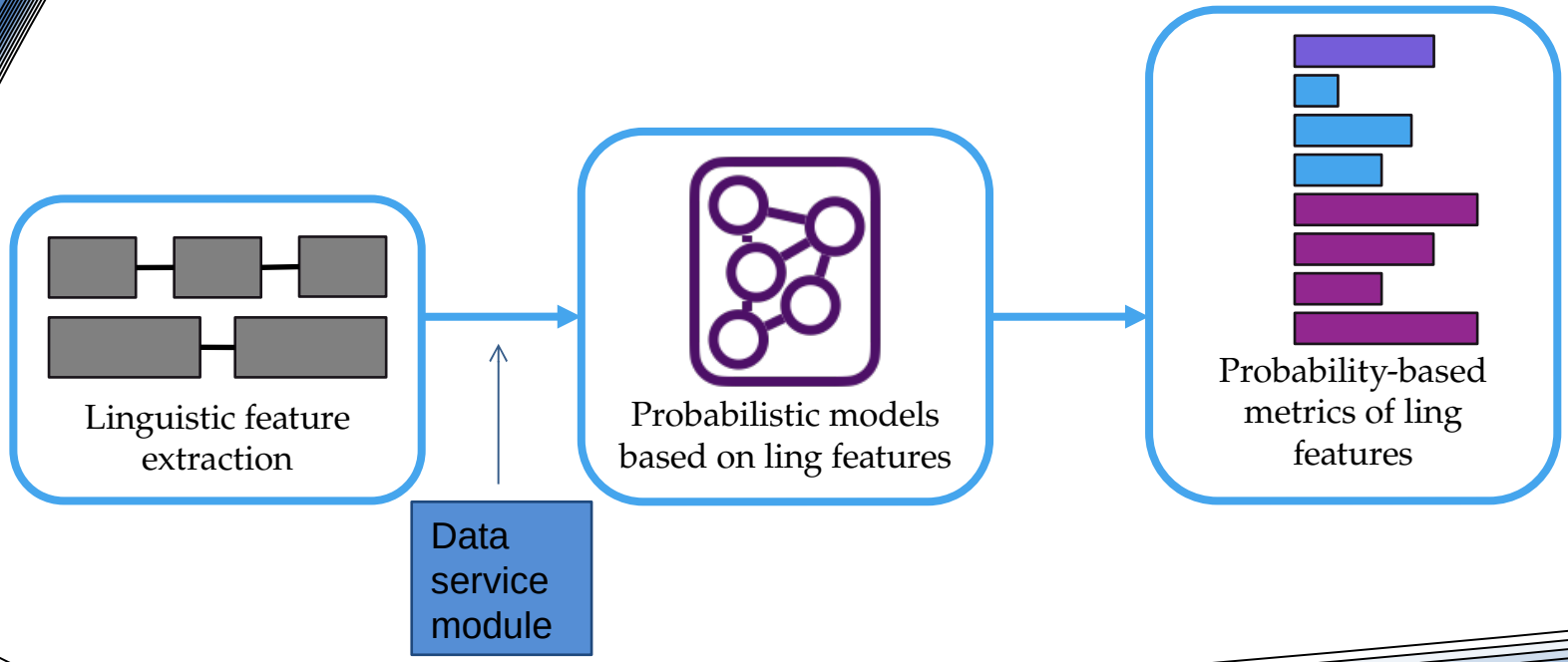
Probabilistic models based on ling features



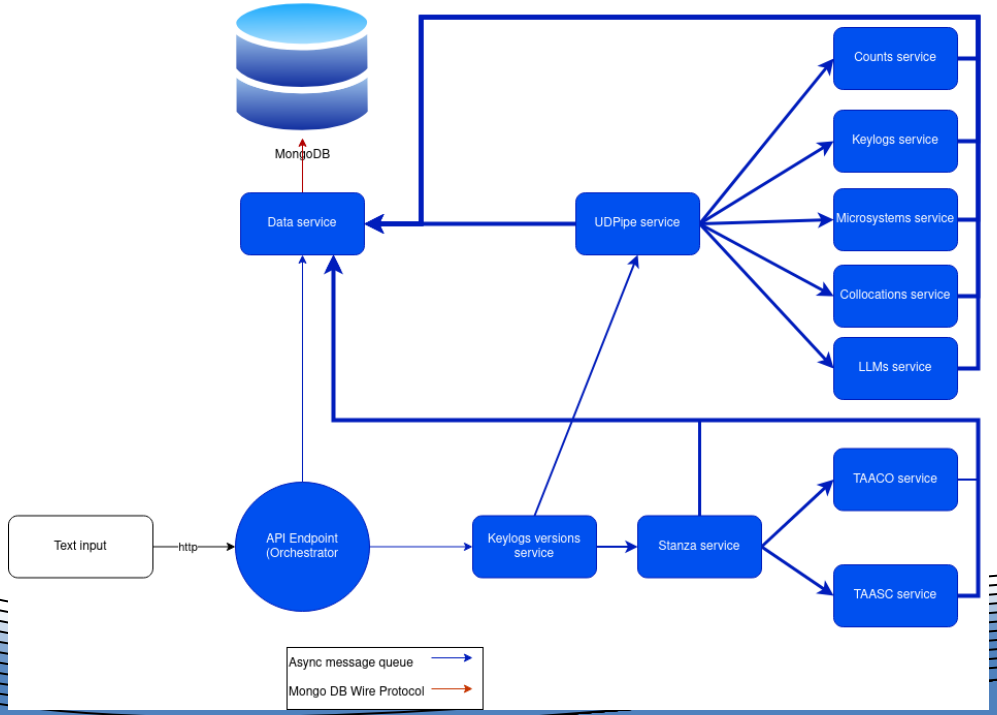
Probability-based metrics of ling features

Symbolic knowledge

RQ2 Model interoperability



RQ3 Architecture and data flow



RQ3 Architecture and data flow

- Hosted on a Huma-Num Virtual Machine running Linux OS
- Docker containers holding microservices

- University MOODLE server + LTI protocole for user interface

Conclusion

How to have the same data representation for any corpus?
How to make data and models compatible?

- Independent & dedicated services
- Data exchanges relying on same standards & protocoles
- Models features filtered in

Deliverables

https://sites-recherche.univ-rennes2.fr/lidile/articles/a4all/

CELVA.Sp L2 corpus > Huma-num Nakala

MOODLE corpus collection module > Huma-num Gitlab

DMP > Opidor

Python programs > Huma-num Gitlab

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/