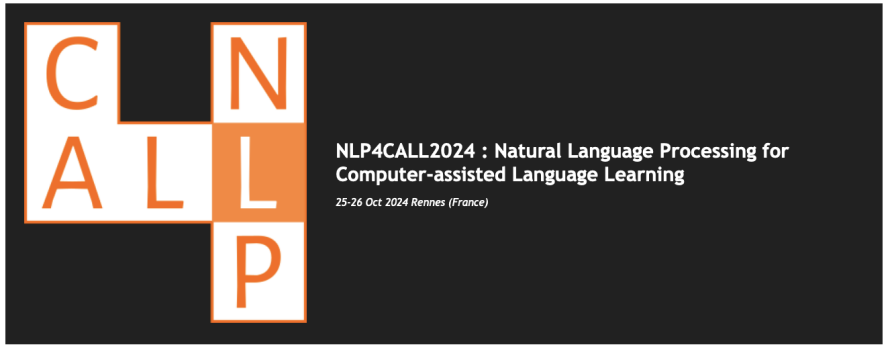
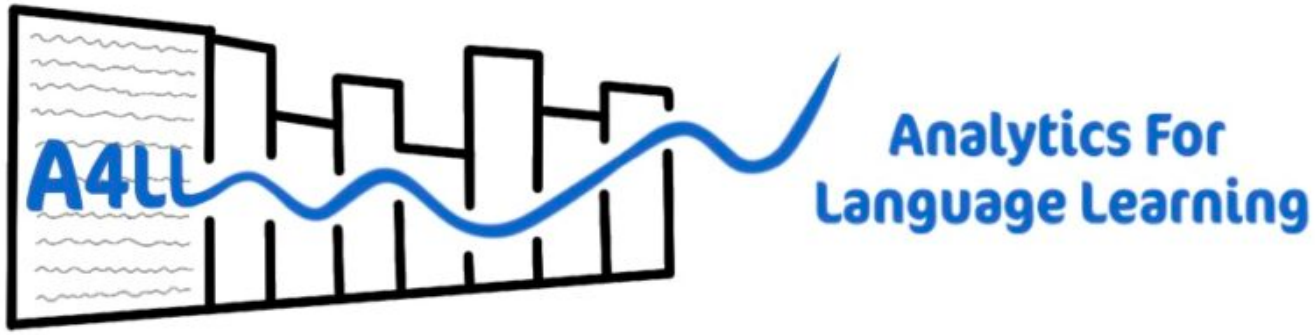
JOINT COLLABORATION

- Bernardo Stearns (code)
- Thomas Gaillat (co-supervision)
- A4LL project (LIDILE, Rennes) https://sites-recherche.univ-rennes2.fr/lidile/articles/a4all/





OUTLINE

- 14 h A4LL a linguistic analytics dashboard for teachers of L2 English Introduction and dashboard demo – Thomas Gaillat & Rémi Venant

- 14h 30 Overview of the linguistic features: creating measures – Joint presentation: Nicolas Ballier, Bernardo Stearns and Jen-Yu Li

- 15h Modelling learners' CEFR against features of their texts – Andrew Simpkin
- 15h30 A4LL architecture and modularity: minding collaboration and future steps in the system design – Cyriel Mallart

Overview of the linguistic features: creating measures : micro-systems and keylogs

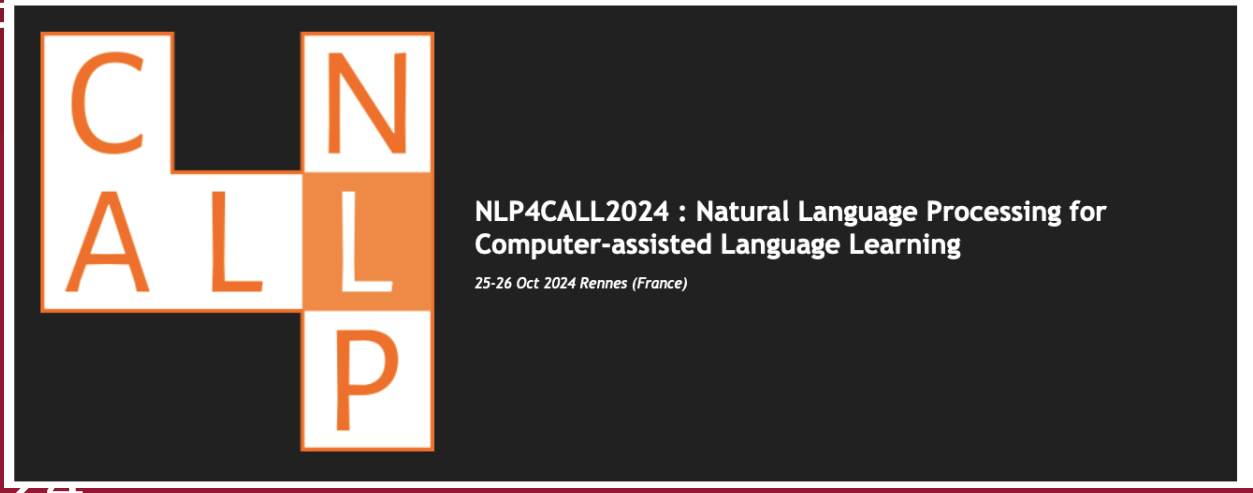
- Overview of the linguistic features: creating measures – Joint presentation: Nicolas Ballier, Bernardo Stearns and Jen-Yu Li

Université Paris Cité

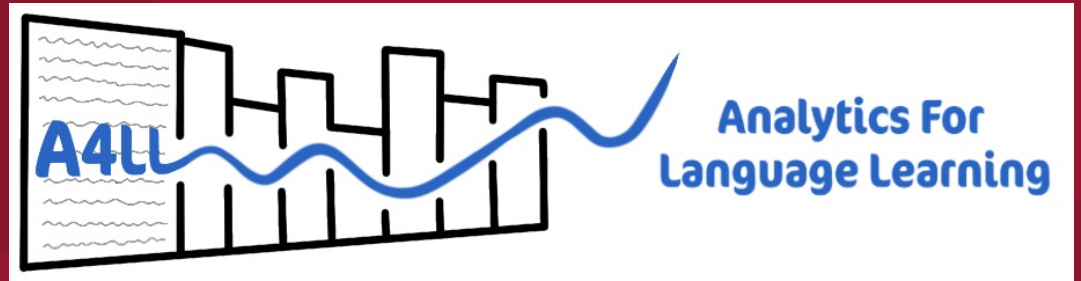
creating measures : micro-systems and keylogs

Nicolas Ballier

nicolas.ballier@u-paris.fr



NLP4CALL2024, Rennes, 24 Oct 2024



MICRO-SYSTEM-based Metrics

- Alternation research paradigm : competing structures
- (Bresnan on dative alternation): the double-object construction and the prepositional dative.

John gave Mary a book. (double-object construction) John gave a book to Mary. (prepositional dative)

Fred picked up the book vs. Fred picked up the book: The genitive alternation: This involves choosing between the s-genitive and the of-genitive.
The squirrel's nest (s-genitive) *The nest of the squirrel* (of-genitive)

- Particle placement: This alternation involves choosing between placing a particle before or after the direct object. *He picked up the book*. (particle before direct object) vs He picked the book up. (particle after direct object) multifactorial analysis of particle placement (Gries, 2003)
- That-complementation: This involves choosing between including or omitting the word "that" in a complement clause. For example:vI thought that the first officer likes the counselor. (inclusion of "that") I thought the first officer likes the counselor. (omission of "that")

Reduced relative clauses: This involves choosing between a full relative clause and a reduced relative clause (the newspaper he read / that he read)

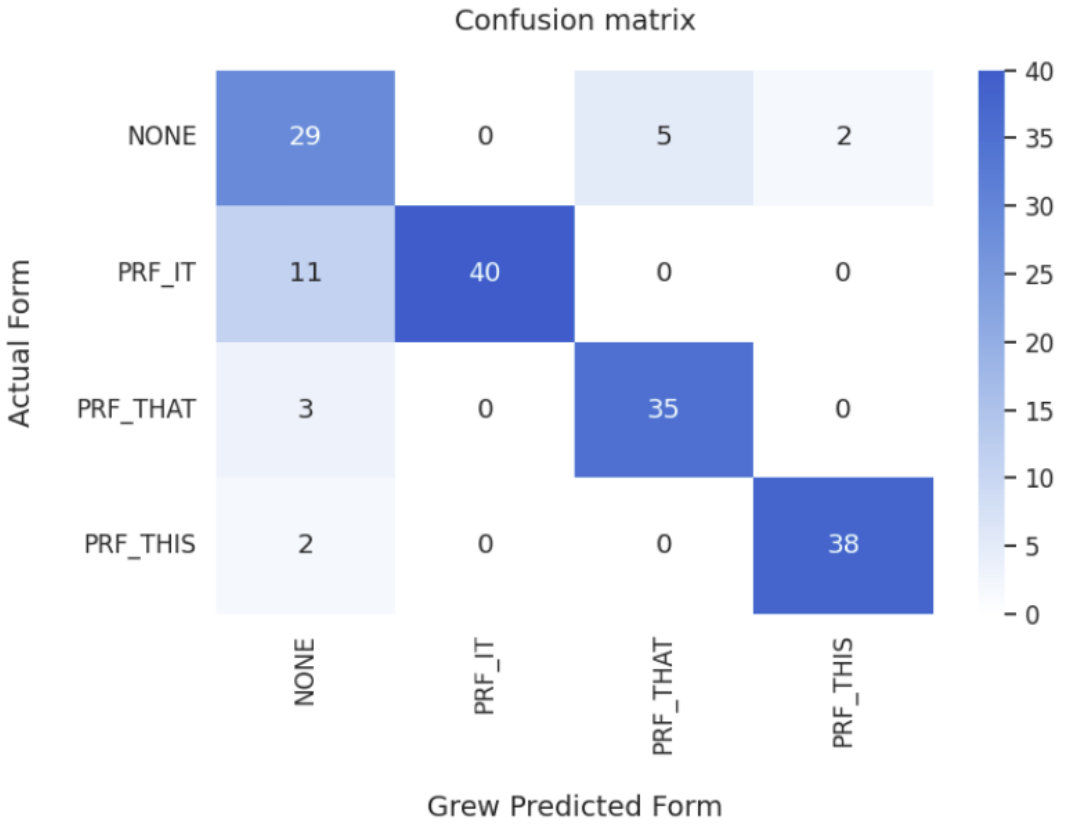
Competing constructions MS

Microsystems	Components	Function	Examples of confusions
Proforms	it, this, that	reference to entity	The student cares for this/that/it
Multi-noun	compound, genitive, prepositional	Pairs of nouns functioning as compounds, genitive or prepositional phrase	She took a student loan/a student's loan/the loan of a student.
Articles	a, the or	determining a noun	a/the/0 loan
Duration	for, since or during	complementing a verb with duration related information	The student has had this loan for/since/during 2 years.
Quantifier 1	any, some	determining a quantity: one or more or unspecified respectively	Any/some students could help.
Quantifier 2	many, much	determining an important quantity	Many/much hard-working students don't rest.
Relativiser	that, which, who	surbordinator refering to entity	The students who/that/which study.

MICRO-SYSTEMS (Gaillat et al, 2022)

Microsystems	Function	variables
Nominal constructs	Denomination	determiner genitive; noun-of/for-noun constructions, compound nouns
Modals for possibility	Possibility	*may*; *can*; *might*; *could*
Modals for obligation	Obligation	*must*; *have to*
Proforms	Reference	*it*; *this*; *that*
Articles	Determination	*a*; *the*; *Ø*
Relativisers	Reference	*that; which; who; 0*
Complementizer vs relativizer	Expressing hypotaxis	*that*
Duration/start/date	Expressing time	*For; since; ago; from; during*
Prepositional constructions	Linking entities	*For; to*
Quantifiers	Quantification (Neutral; large; small)	*Some* vs *any; many* vs *much* vs *most; few* vs *little*

MS are operationalisable with Grew-match extraction



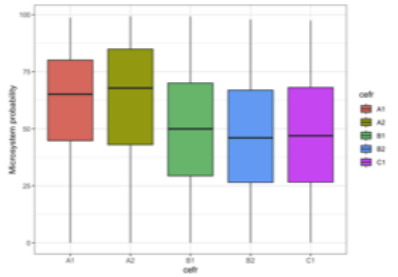
Confusion matrix for the extraction of IT, THIS and THAT proforms in the Gold Standard dataset

MS can be automatically extracted and computed (but issues remain with zero extractions)

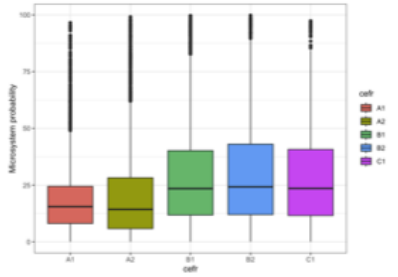
Table D4. Quality of relativizer MS extractions in the GS

	precision	recall	f1-score	support
NONE	0.62	0.88	0.73	32
REL THAT	0.90	1.00	0.95	36
REL WHICH	1.00	0.85	0.92	47
REL WHO	1.00	0.80	0.89	50
accuracy			0.87	165
macro avg	0.88	0.88	0.87	165
weighted avg	0.90	0.87	0.88	165

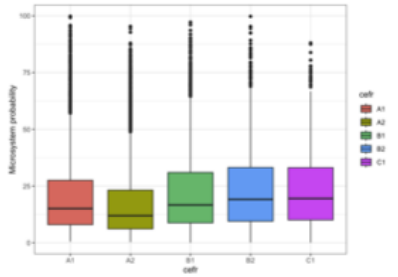
MS can be actionable for prediction level tasks



(a) Median probabilities of IT.



(b) Median probabilities of THAT.



(c) Median probabilities of THIS.

Typology of the keylog metrics

- Behavioural metrics (per text) : pauses, edits, r-burst vs. P-burst
- Behavioural metrics: inter-key intervals
- Behavioural metrics (biometry / typist identification): Tapper & Villani

- text-based metrics: writing bursts
-

P- bursts vs. Revision R-bursts (Pacquetet, 2024)

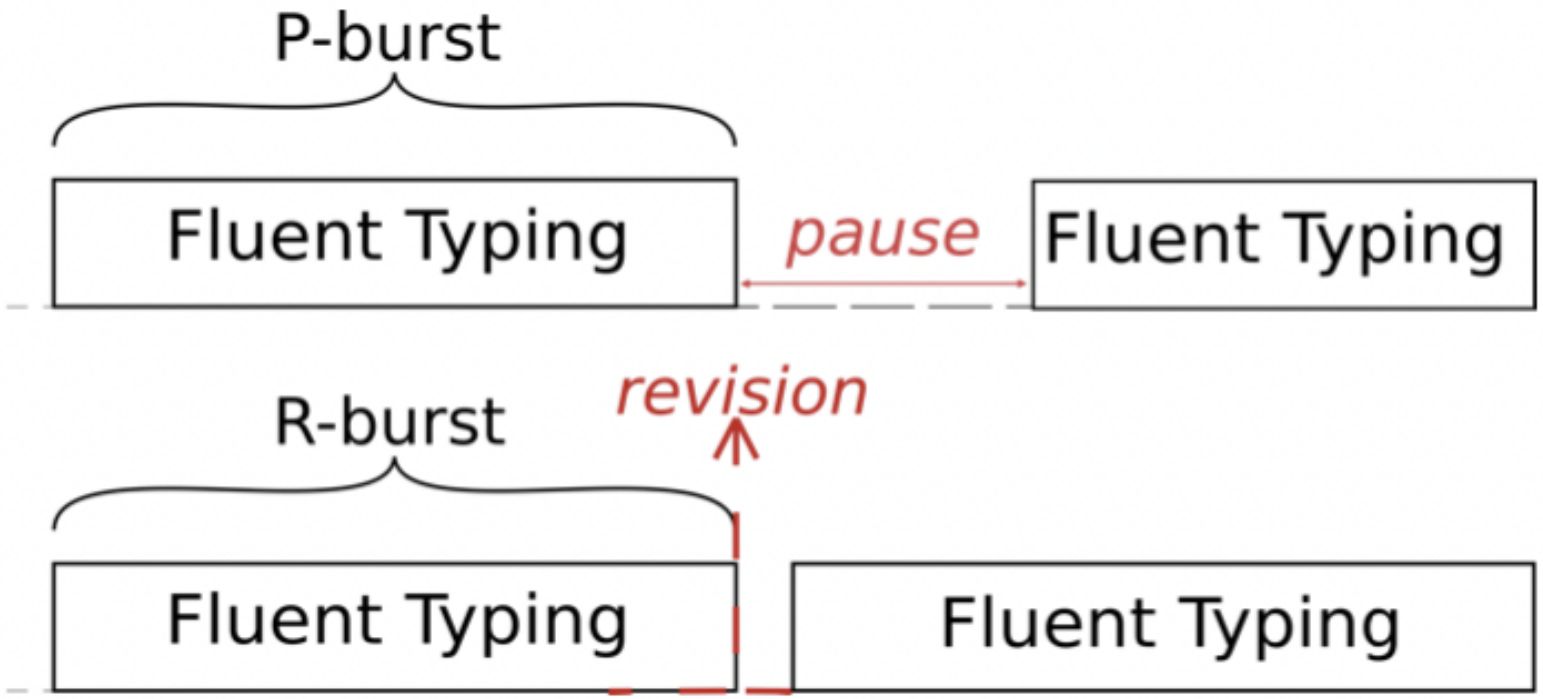


FIGURE 2.3: Types of typing bursts

50 KEYLOG METRICS FOR A4LL (a selection)

- Character-based metrics : (eg ratio_backspace_keys)
- Word-based metrics : mean_length_pauses_after_word
- burst metrics : mean_time_revision_burst
- Sentence-based metrics : ratio_nb_rev_burst_per_sentence
- Text-based metrics : total_nb_bursts_any_kind

References

- Pacquetet, E (2024) *The effect of linguistic properties on typing behaviors and production processes* , PhD. University of Buffalo
-

ALTERNATES : annotation guidelines

MS	Description
Quantifier 1	
any	as a determiner
some	as a determiner not as an adverbial
Articles	
A/an	Article A as a determiner
THE	Article THE as a determiner
Article 0	Nouns without any determiner. As a proxy we list *nouns* that have neither determiner nor possessive pronoun dependency relation. In case there is a THE or A article in front of that noun, select the value corresponding to that article. If it is introduced by a quantifier (fewer, many, any...), select none.
Proforms	
IT	It as an proform only, not extrapositional e.g. "it's ridiculous that they've given the job to PAt", nor impersonal e.g. "It seemed that / as if things would never get any better.". it-cleft constructions, e.g."It was your father who was driving - No it wasn't not, it was me." or weather/time it e.g. "It's only two weeks since she left." "It's raining."
THIS	only as proform, not as determiner, nor adverbial
THAT	only as proform, not as determiner, nor adverbial, nor relativizer nor complementizer.
Multinoun	For the multinoun MS, the *last* word of the pattern is between two stars *. For instance: The university *car*; The university's *car*; The car of the *university*
N of N	Any time a noun appears in a N of N construction
NN	In cases of NN it can be either first or second position. e.g "I am studying materials science in an *engineering* school .". Here consider that the target to evaluate is Engineering school even if it is the first N that is between stars. NOTE: this pattern does NOT include ADJ + NN of course.
N's N	Any time a noun appears in a N's N construction
Duration MS	
FOR	"For" used to express a lasting period of time (translates as "pendant" in FRench). Not to be confused with expression of purpose. e.g. "I want to do this for a gap year." or reason e.g. "thanks for doing xyz"
SINCE	"Since" used as a point of departure in time
DURING	"During" used for the expression of a lasting period of time
Quantification	
MUCH	Used to express quantity
MANY	used to express quantity
Relativizers	
THAT	Uses of "that" as relative pronoun only, NOT as proform, determiner, complementizer or adverbial.
WHICH	Uses of "which" as relative pronoun only, not as interrogative. NOTE: Watch relative pronouns as objects of verb.
WHO	Uses of "who" as relative pronoun only, not as interrogative. NOTE: be careful with cases where WHO has no apparent antecedent: A who relative clause introduced by verb, e.g. "You can meet who you like" (Larreya & Rivire, 1991)

Training Artificial Learners

Making Predictions in new learners texts

Extracting Metrics

Training Artificial Learners

EFCAMDAT dataset

1. LEARNER 18445817, LEVEL 1, UNIT 1, CHINESE

Hi! Anna,How are you? [redacted] you to sendmail to me. My name's Anfeng.I'm 24 years old.Nice to meet you !I think we are [redacted] already,I hope we can learn english toghter! Bye! Anfeng.

2. LEARNER 19054879, LEVEL 2, UNIT 1, FRENCH

Hi, my name's Xavier. My [redacted] days is saturday. I get up at 9 o'clock. I have a breakfast, I have a shower... Then, I goes to the market. In the afternoon, I play music or go by bicycle. I like sunday. And you ?

3. LEARNER 19054879, LEVEL 8, UNIT 2, BRAZILIAN

Home Improvement is a pleasant protest song sung by Josh Woodward. It's a [redacted] but realistic song that analyzes how rapid changes in a town affects the lives of many people in the name of progress. The high bitter-sweet voice of the singer, the smooth guitar along with the high pitched resonant drum sound like a moan recalling the past or an ode to the previous town lifestyle and a protest to the negative aspects this new [redacted] city brought. I really enjoyed this song.

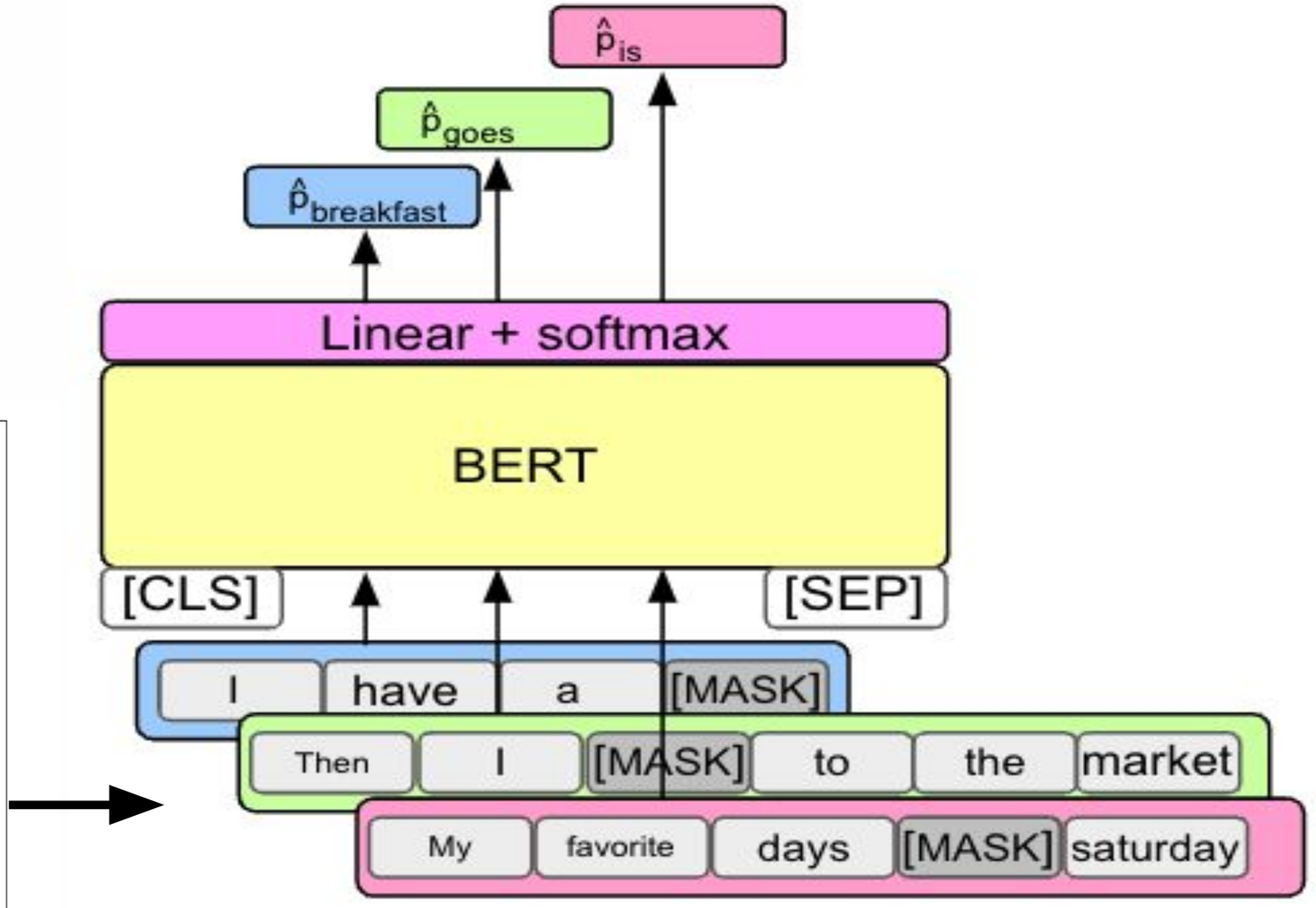
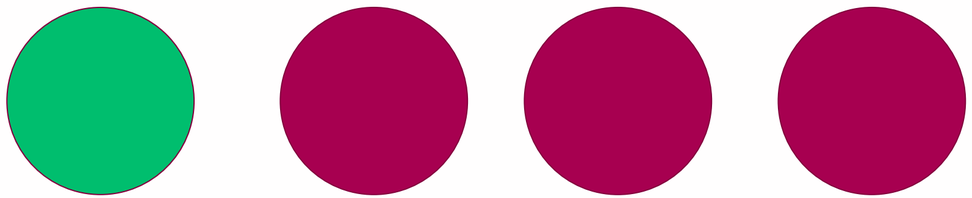


Figure 1: Three typical scripts, in which learners are asked to introduce themselves (1), describe their favourite day (2), and review a song for a website (3).

Making Predictions in new learners texts

A pipeline for extracting sophistication and vocabulary metrics (using the SELVA dataset)

Text	Vocab Range
The injuries who we can see are commotions, bruises, contusion, broken bones or muscle injuries because we can easly fall in a bad position.	B2



Native LM

A-level LM

B-level LM

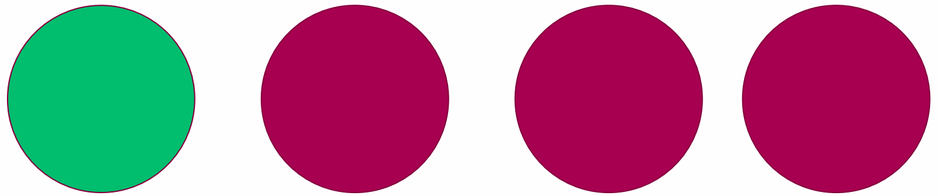
C-level LM

Making Predictions in new learners texts

Hypothesis masking + Artificial Learner predictions. e.g. Token Prediction

Text

The injuries [redacted] we can see are commotions, bruises, contusion, broken bones or muscle injuries because we can easly fall in a bad position.



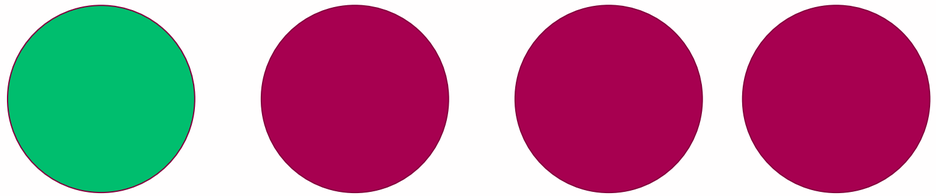
	Native LM	A LM	B LM	C LM
Predictions (top 5)	"that"	"who"	"that"	"as"
	"which"	"that"	"who"	"which"
	"what"	"in"	"in"	"that"
	"as"	"as"	"as"	"which"
	"whom"	"from"	"from"	"who"

Making Predictions in new learners texts

Hypothesis masking + Artificial Learner predictions. e.g. Token Prediction

Text

The injuries [redacted] we can see are commotions, bruises, contusion, broken bones or muscle injuries because we can easly fall in a bad position.



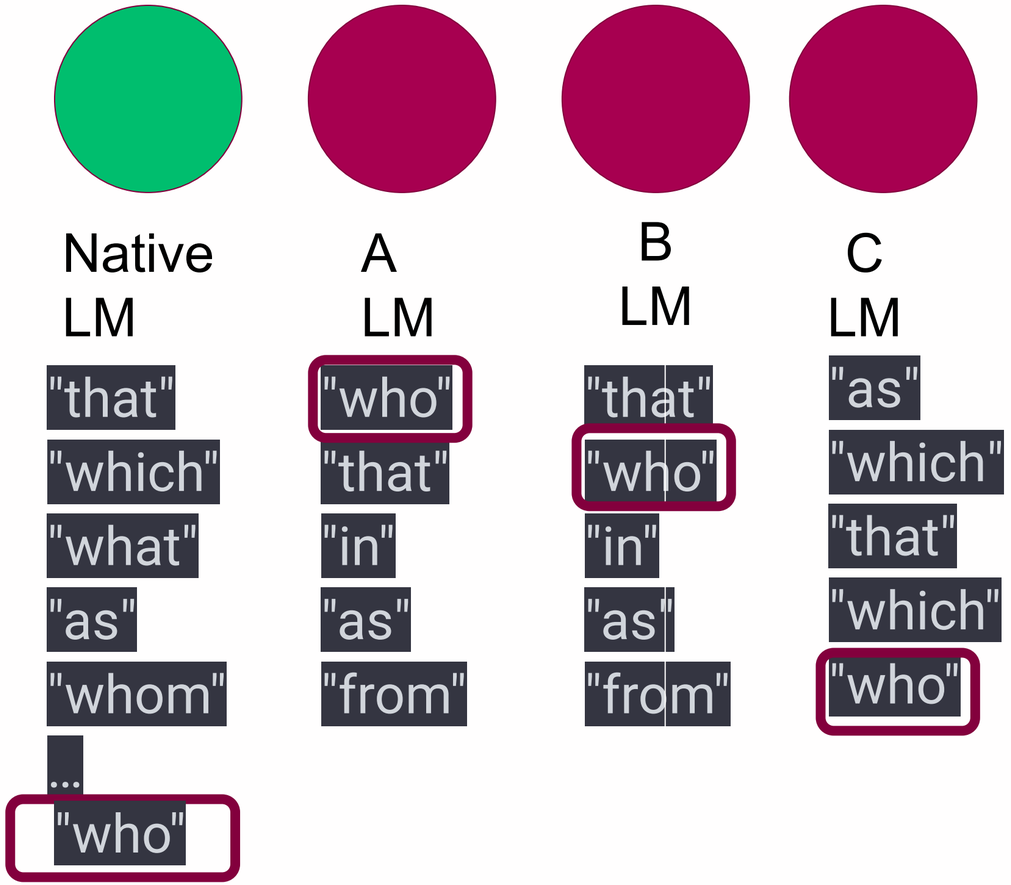
	Native LM	A LM	B LM	C LM
Predictions (top 5)	"pronoun"	"pronoun"	"pronoun"	"conj"
	"pronoun"	"pronoun"	"pronoun"	"pronoun"
	"pronoun"	"prep"	"prep"	"pronoun"
	"prep"	"conjuct"	"conjuct"	"pronoun"
	"pronoun"	"prep"	"prep"	"pronoun"

Extracting Metrics

On average, how likely is an artificial learner to use the learners' words?

Text

The injuries who we can see are commotions, bruises, contusion, broken bones or muscle injuries because we can easly fall in a bad position.



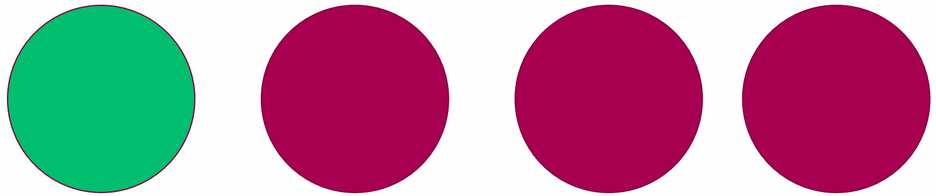
	Native LM	A LM	B LM	C LM
P(w = learner used word)	0.05	0.92	0.40	0.12

Extracting Metrics

On average, how likely is an artificial learner to use the learners' words?

Text

The injuries who we can see are commotions, bruises, contusion, broken bones or muscle injuries because we can easly fall in a bad position.



Native LM	A LM	B LM	C LM
0.96	0.78	0.15	0.67
0.41	0.99	0.34	0.56
0.05	0.92	0.40	0.12
0.02	0.85	0.29	0.74
0.61	0.37	0.88	0.03

Overview of linguistic features: Verb-Noun collocations

Jen-Yu Li
University Rennes 2
LIDILE (Linguistique, Ingénierie, Didactique des Langues), France

NLP4CALL 2024: 25-26 October, held in Rennes, France

Collocations

- Subset of phrasemes (Mel'čuk, 1998; Tutin, 2013)
- Component of lexical competence (Eguchi and Kyle, 2023)
- Second language (L2) learners usually encounter difficulties in collocations (Garner et al., 2020)

Examples of erroneous Verb Noun collocations:
**create a better material, *create a taller building, *reform the land*

Corpora

Learners Corpora

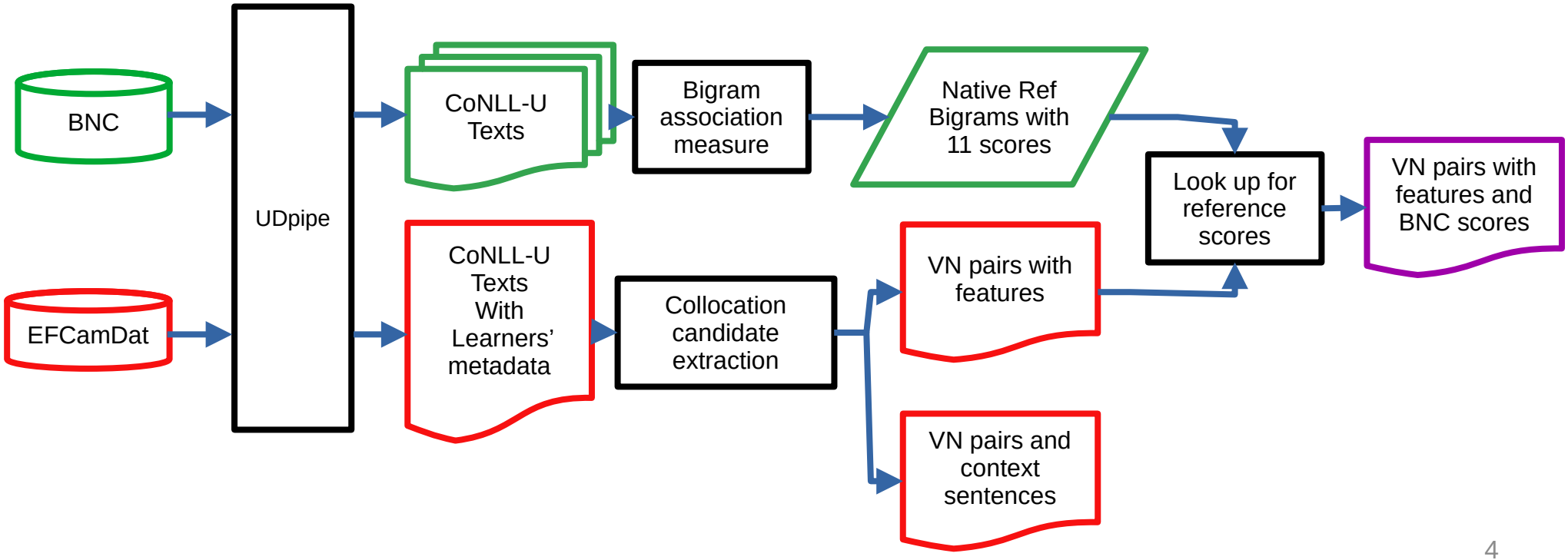
1)EF-Cambridge Open Language Database (EFCamDat) (Geertzen et al., 2013; Shatz, 2020)

2)National University of Singapore Corpus of Learner English (NUCLE) (Dahlmeier et al., 2013)

Native Speaker Corpus

3)British National Corpus (BNC) (BNC Consortium, 2007)

Collocation Extraction



Collocation Indicators

- Identification / Count
  collocations in the text

- Rate
  number of collocations by text length

- Diversity
  Type Token Ratio: the number of different collocations by the total number of collocations

Main references

BNC Consortium, 2007, The British National Corpus. Distributed by Bodleian Libraries, University of Oxford

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, 22–31.

Eguchi, M., & Kyle, K. (2023). L2 collocation profiles and their relationship with vocabulary proficiency: A learner corpus approach. Journal of Second Language Writing, 60, 100975.

Garner, J., Crossley, S., & Kyle, K. (2020). Beginning and intermediate L2 writer's use of N-grams: An association measures study. International Review of Applied Linguistics in Language Teaching, 58(1), 51–74. https://doi.org/10.1515/iral-2017-0089

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project, 240-254.

Mel'čuk, I. (1998), "Collocations and Lexical Functions", in A. Cowie (ed.), Phraseology. Theory, Analyses, and Applications, Oxford: Oxford University Press, 23-53

Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. International Journal of Learner Corpus Research, 6(2), 220–236.