Assessing BERT's sensitivity to idiomaticity

Li Liu, François Lareau

Observatoire de
linguistique Sens-Texte

Université
de Montréal

Compositionality in idioms: a continuum

Idiomaticity		
Weak idiom	***étoile** de **mer***	‘**star**-shaped **marine** animal’
	(lit.) star of sea	
Semi-idiom	*fruit de **mer***	‘food that comes from **sea**’
	(lit.) fruit of **sea**	
Strong idiom	*noyer le poisson*	‘obfuscate things’
	(lit.) drown the fish	

(Mel’čuk 2014)

Research question

Question

Are LLMs like BERT sensitive to the degree of idiomaticity in idioms?

Task

Fill-mask task with CamemBERT-base on a French dataset

Hypotheses

BERT should be better at predicting:

- **tokens within idioms**, compared to simple lexemes
- **tokens within idioms with higher idiomaticity**, compared to those with lower idiomaticity

BERT vs. Idioms

- BERT can distinguish between the literal and idiomatic usage of potential idiomatic expressions. **(Tan and Jiang, 2021)**
- BERT-like language models represent idioms differently from their literal counterparts at both sentence and word levels. **(Tian et al., 2023)**
- BERT incorporates information from idioms and their surrounding context to process them. **(Nedumpozhimana and Kelleher, 2021)**
- Vector space models including BERT can not represent appropriately idiomaticity in noun compounds in English and Portuguese. **(Garcia et al. 2021b)**

Dataset : French Lexical Network (LN-Fr)

Lexical unit	Idiomaticity	POS	Example(s)
pomme 'apple'	simple lexeme	N	*À la fin du repas, on a parfois droit à un petit morceau de brie et, en guise de dessert, selon la saison, des **pommes**, des noix,…*
pomme de terre (lit.) apple of ground 'potato'	weak idiom	N Prep N	*Ils prenaient une demi-heure à midi pour manger un œuf sur le plat, une **pomme de terre**, du fromage blanc.*

Dataset : French Lexical Network (LN-Fr)

Type	Lexical units	Examples	Tokens
Simple lexeme	22551	42849	45563
Idiom	3127	4546	13529
Weak idiom	592	916	2425
Semi-idiom	589	899	2408
Strong idiom	1946	2731	8696
Total	**25678**	**47395**	**59092**

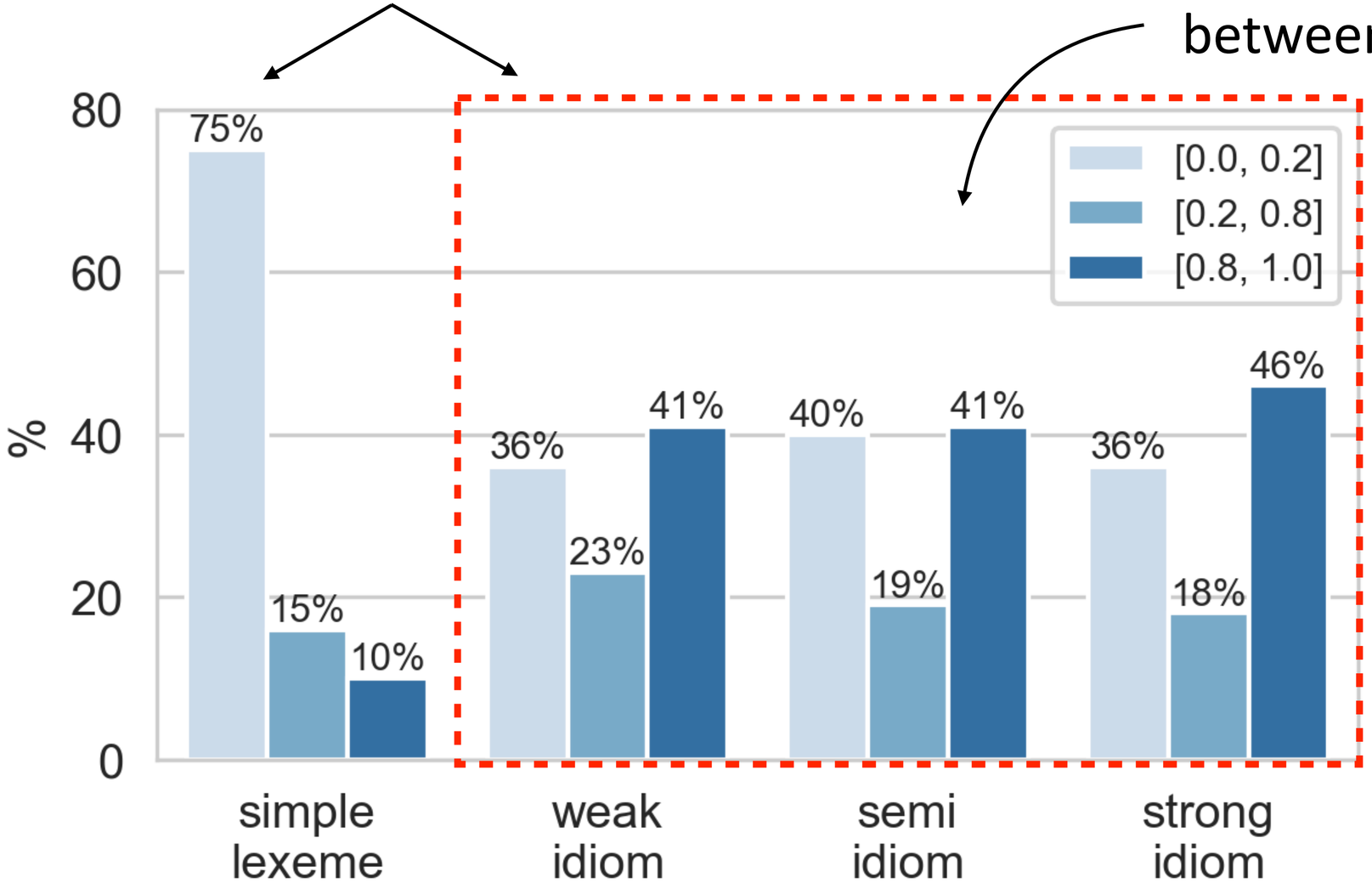
Fill-mask experiment : inputs and outputs

Lexical unit	Token	POS	Example(s)	Score	R1
pomme	*pommes*	N	*À la fin du repas,…, en guise de dessert, selon la saison, des* ***<mask>****, des noix,…*	0.10	F
pomme de terre	*pomme*	N	*Ils prenaient une demi-heure à midi pour manger un œuf sur le plat, une* ***<mask>*** *de terre, …*	0.99	T
	de	Prep	*… une pomme* ***<mask>*** *terre*	0.99	T
	terre	N	*… une pomme de* ***<mask>*** *…*	0.99	T

Analysis 1: idiomaticity levels

Significant difference: free vs. idiomatic tokens

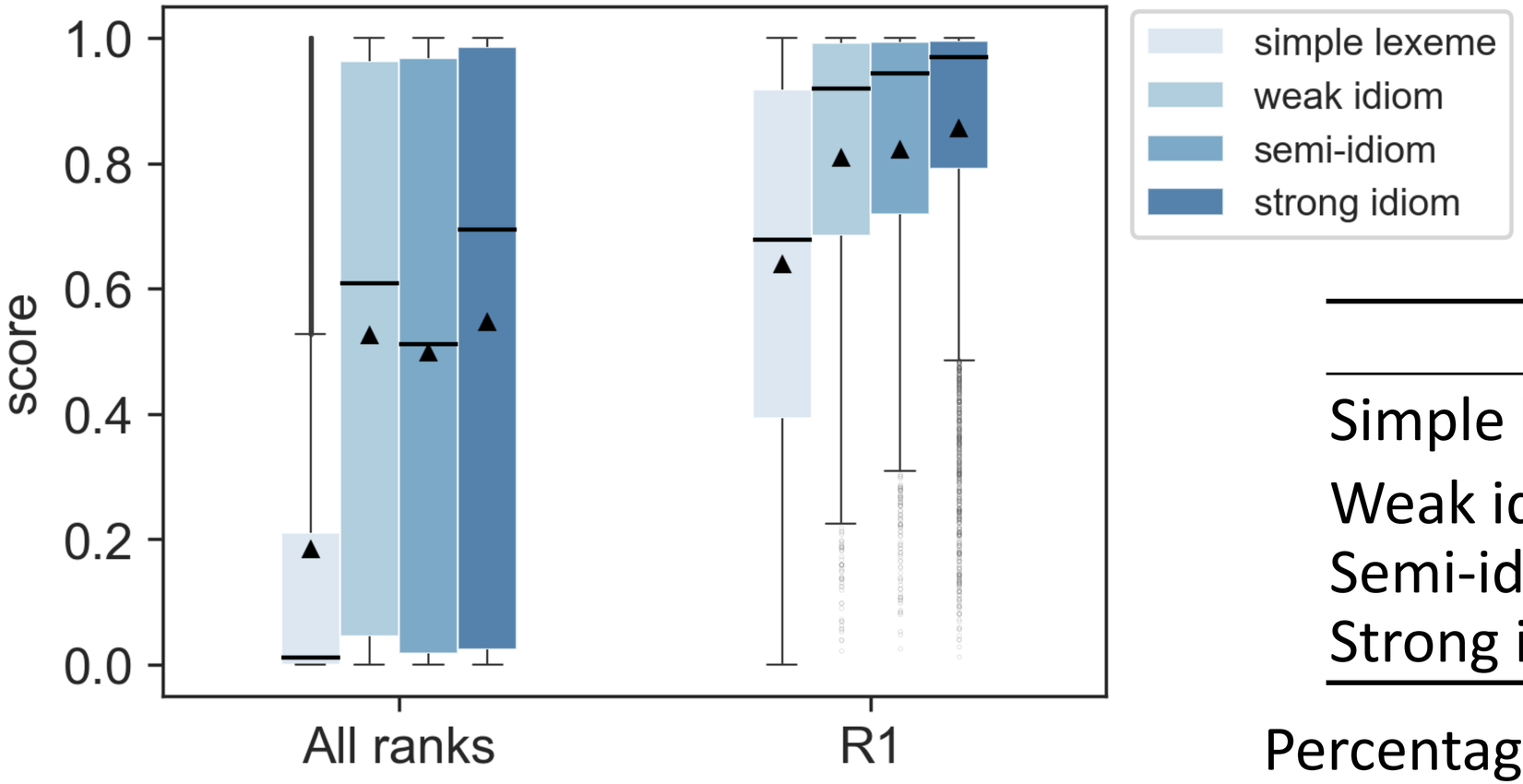
No significant difference between tokens within idioms



Score distribution
(Kruskal-Wallis test)

Analysis 1: idiomaticity levels

Idiomaticity levels (all) vs. Pred scores : moderately positive correlation



Score at all ranks and at R1

(— median, ▲ mean)

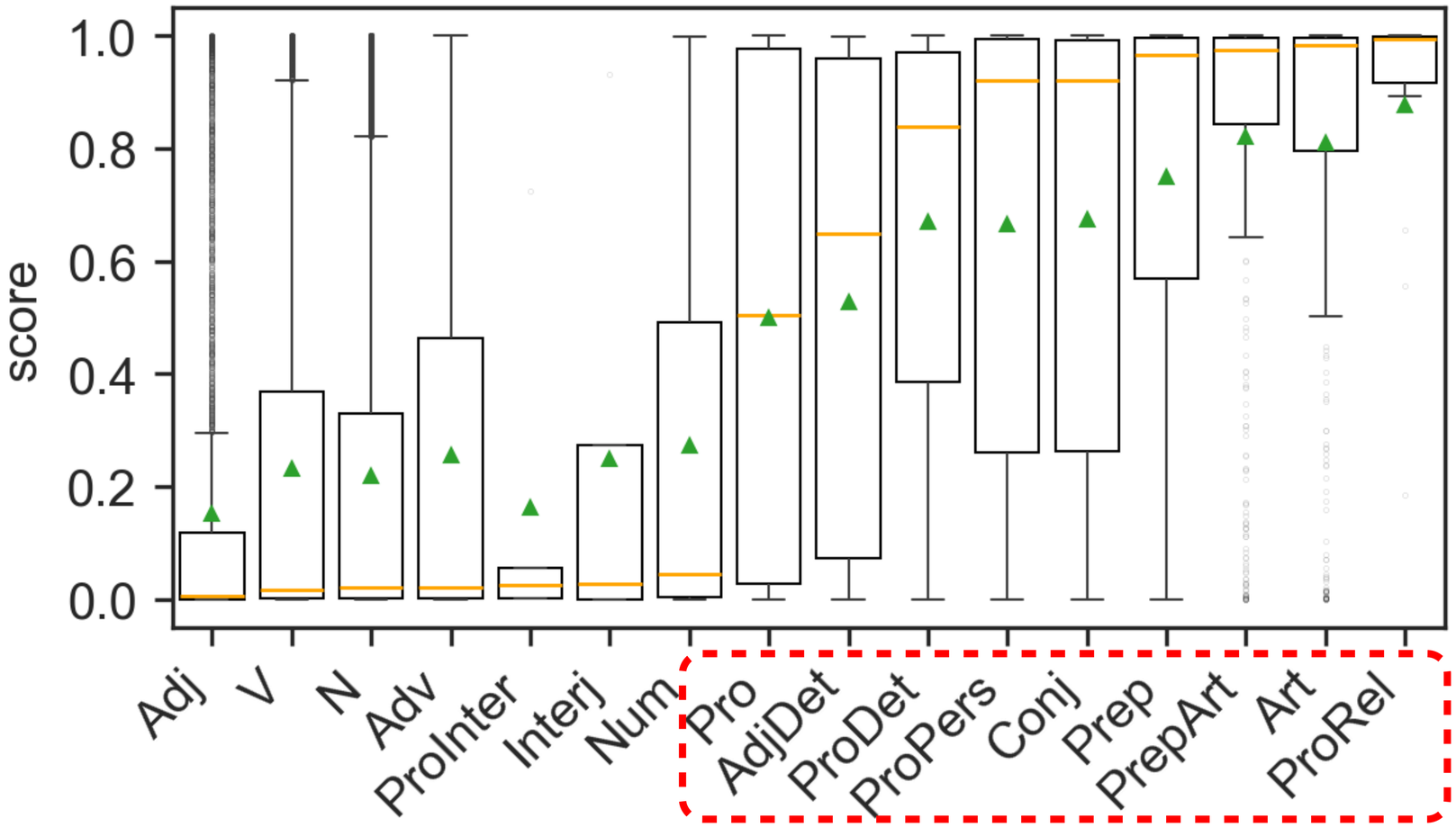
Same conclusion

	%R1
Simple lexemes	25 %
Weak idioms	62 %
Semi-idioms	58 %
Strong idioms	62 %

Percentage of **correctly predicted tokens (%R1)** by idiomaticity levels

Analysis 2: tokens within idioms (POS)

POS vs. Pred scores : moderately positive correlation



Score by token POS

16 token types:

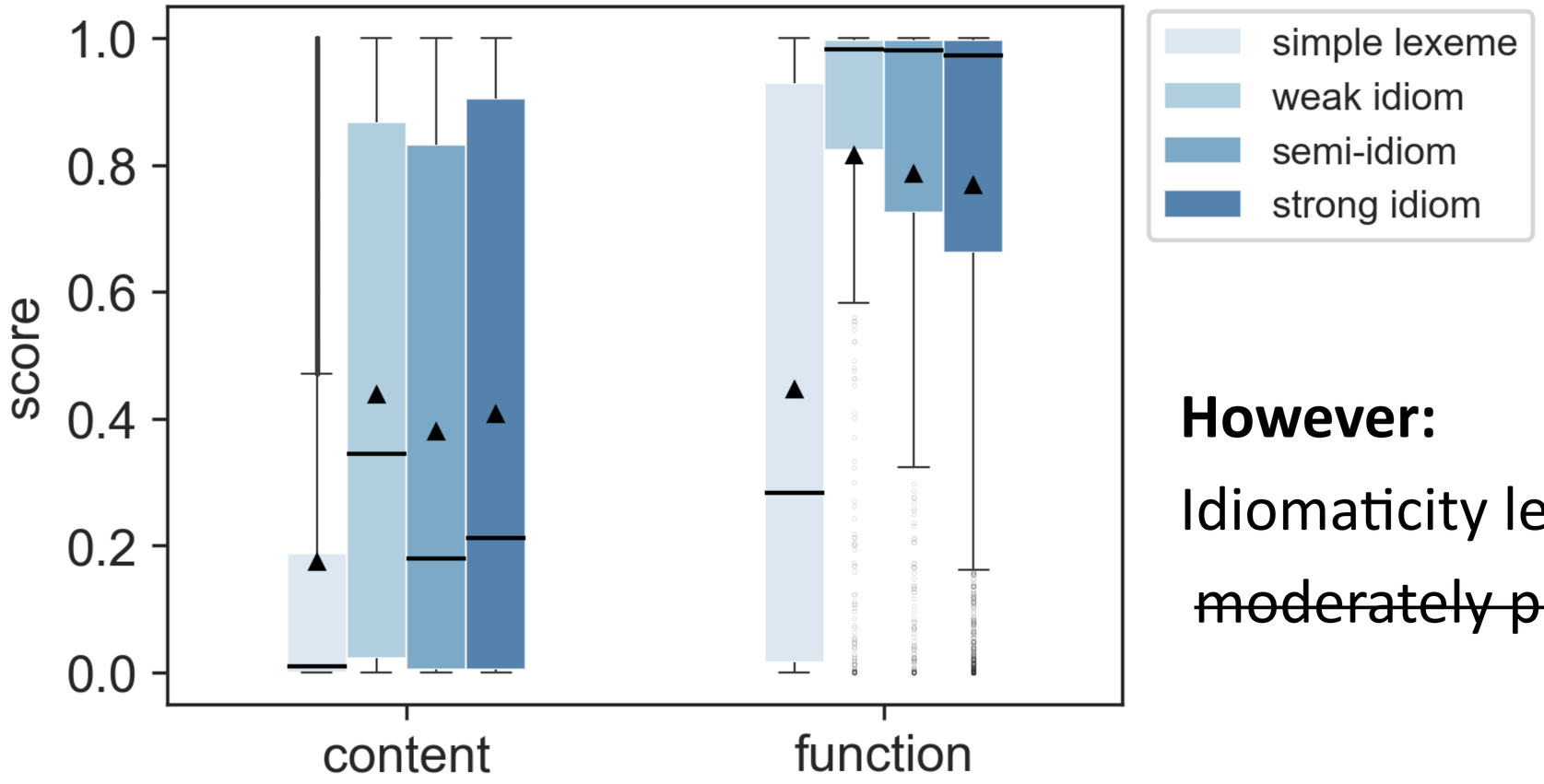
7 content token types

9 function token types

	Content	Function
Simple lexemes	99.52%	0.48%
Idioms	71.36%	28.64%

Analysis 2: tokens within idioms (POS)

Back to Analysis 1 …



Scores for content and function tokens

Same conclusion

However:

Idiomaticity levels (all) vs. Pred scores :

~~moderately positive correlation~~

Analysis 2: tokens within idioms (POS)

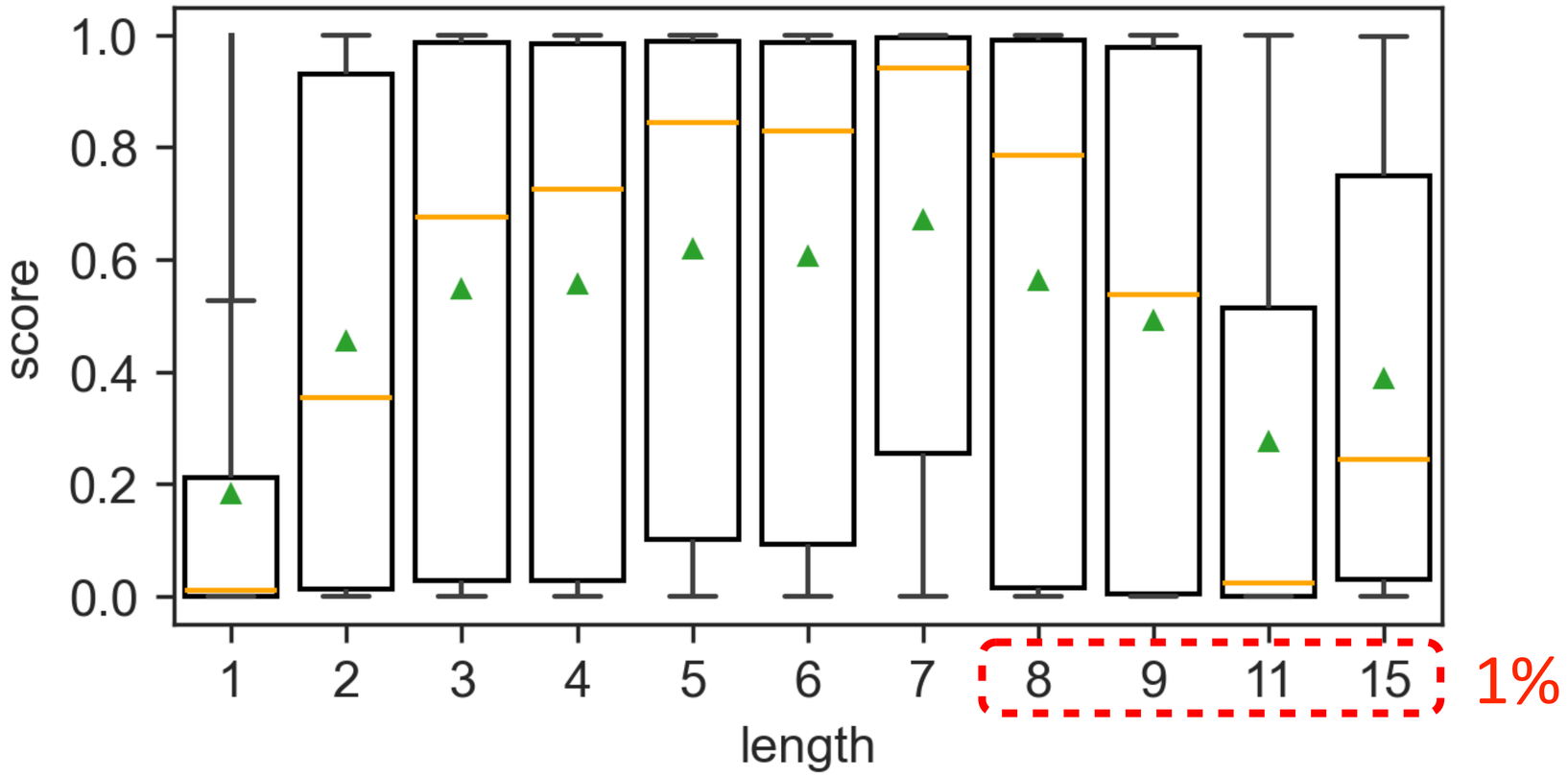
Same conclusion

	All	Content tokens	Function tokens
Simple lexemes	25 %	24 %	50 %
Weak idioms	62 %	55 %	86 %
Semi-idioms	58 %	48 %	83 %
Strong idioms	62 %	49 %	81 %

Percentage of **correctly predicted tokens (%R1)**
for content and function tokens

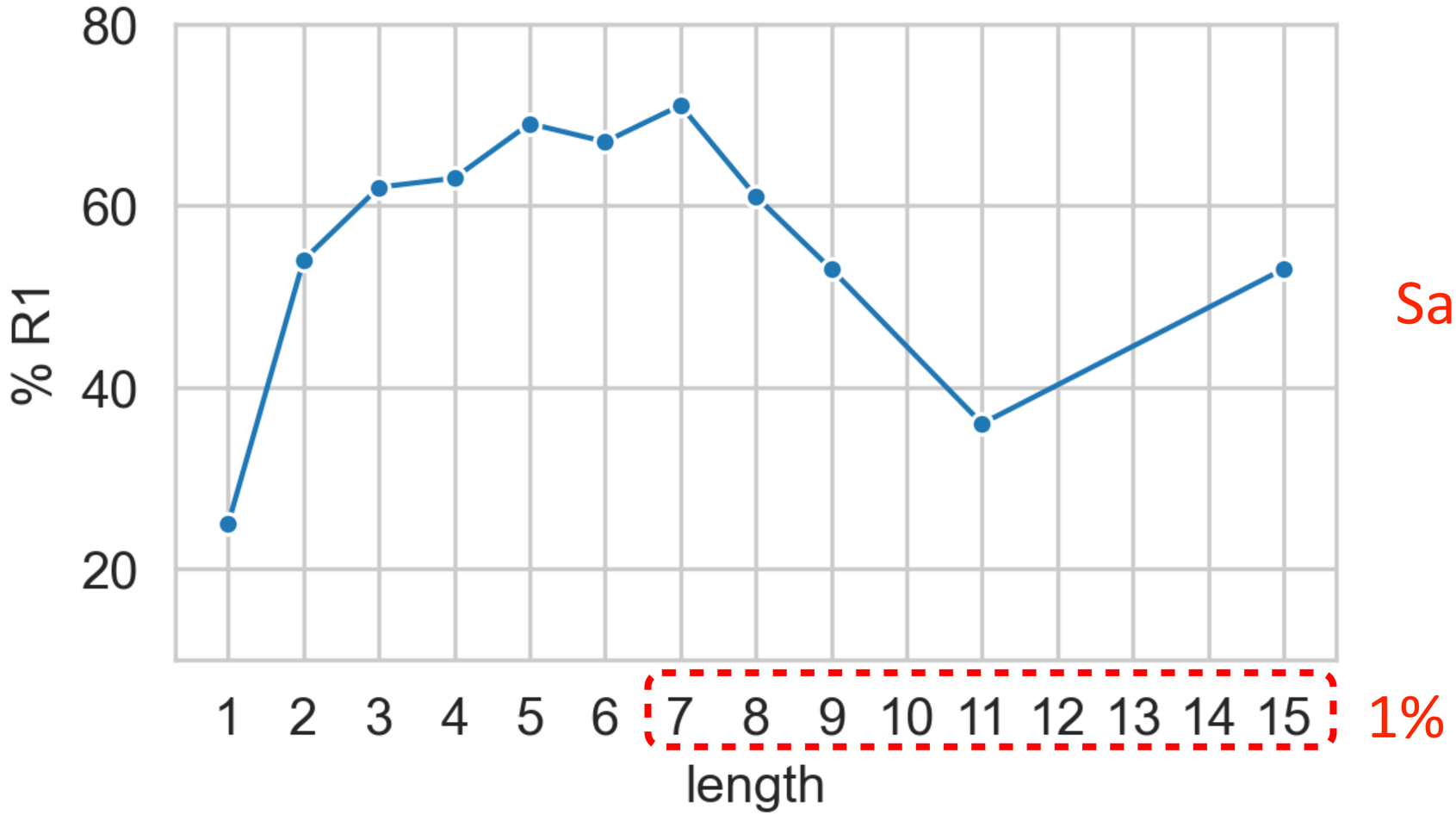
Analysis 3: idiom length

Idiom length vs. Pred scores : moderately positive correlation



Scores by lexical unit length

Analysis 3: idiom length



Same conclusion

1%

Percentage of **correctly predicted tokens (%R1)** by lexical unit length

Conclusion

- The model is significantly better at predicting tokens that belong to an idiom as opposed to simple lexemes. **(Hypothesis 1)**
- It is not sensitive to varying levels of idiomaticity among subtypes of idioms. **(~~Hypothesis 2~~)**
- It exhibits a heightened performance in predicting function words, regardless of idiomaticity.
- There is a positive correlation between idiom length and performance.
- **CamemBERT is more sensitive to lexical idiomaticity than semantic idiomaticity.**

Future work

- Other types of MWEs : collocation
- Other forms of idiomaticity
- Other language models
- Available dataset in other languages
- Additional potential influencing factors such as idiom frequency, etc.
- More complex tasks
- …

Thank you for your attention!

Contact us for more information:

{li.liu.2, francois.lareau}@umontreal.ca

Our dataset is available on github:

https://github. com/liliulng/idiomaticity-dataset