

WG3 meeting : shared task(s) organisation brainstorming

January 30, 2025

Two shared tasks have been proposed:

- The PARSEME 2.0
- AdMIRe

- Go on with the UD and PARSEME research questions
 - **universality**, etc.
- Attract **mainstream** research (LLMs)
- Have a good ground to apply **diversity measures**
 - Inter-linguistic
 - Intra-linguistic
- Improve quality of MWE representations
 - Avoid construction artifacts
- Expand coverage of MWE-related tasks
 - More language diversity

SemEval 2025 Task 1. Evaluation phase ends 31st January.

<https://semeval2025-task1.github.io>

Items in EN and PT-BR.

Problem: Existing idiomaticity detection tasks vulnerable to **dataset construction artifacts** (Boisson et al 2023).

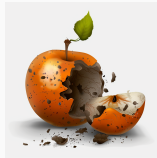
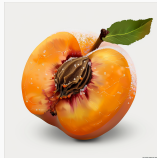
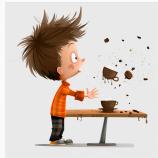
Challenge: Can we construct something which avoids this?

Idea: **Multi-modal** task might do the trick.

Subtask A - Static Images

Which of these images best represents the meaning of **bad apple** as used in the following sentence?

"However, if ethylene happens to be around (say from a bad apple), these fruits do ripen more quickly."

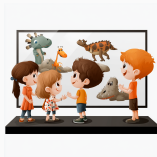
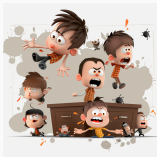


Rank the images according to how well they represent the given expression as used in context.

Subtask B - Image Sequences



Which of the following best completes the sequence representing the meaning of **bad apple**?



Is the sense represented **idiomatic** or **literal**?

Some Results (Preview!)

Note: These are not all from the same system.

Subtask A	EN	PT-BR	177 participants 1150 submissions (all phases)
Items (total)	100	55	
Best accuracy - test set	0.87	0.92	
Best accuracy - extended	0.81	0.67	
Subtask B	EN	PT-BR	69 participants 78 submissions (all phases)
Items (total)	30		
Best accuracy - test set	0.6	-	
Best accuracy - extended	0.23		

Text-only results similar, lower on the extended evaluation set.

Next: Human evaluation (tmrpickard1@sheffield.ac.uk)

Goals :

- Evaluate systems' ability to **predict** MWEs (non-verbal included)
- Assess the ability of Large Language Models (LLMs) to **understand and/or predict** MWEs
- Evaluate the **diversity** of the MWEs predicted
 - Continuity of **unseen MWEs** evaluation
 - Training on more diverse data covers more diverse phenomena
 - To promote **rare** phenomena

Idea 1 : Idiomatic language **generation** by LLMs

Given a one paragraph context on the task, what will the LLM predict ?

- PROMPT: “It’s *raining cats* and ...”
- PROMPT: “The child has been *crying* ...”

Idea 2 : Idiomatic language **comprehension** by LLMs

Given a one paragraph context on the task, can the LLM predict if the tokens ***pulling the strings*** are literal or idiomatic ?

- PROMPT: “[...] who is going to be ***pulling the strings*** in the Europe of tomorrow”
- PROMPT: “[...] the artist was ***pulling the strings*** to move the marionette”

And can the LLM explain the meaning of an MWE?

Need for a community effort

- MWE annotation calls for native language skills
- We want many languages

Problem : SemEval 2026

What to do ?

Similarities :

- Multiword expressions
- Native language skills & good understanding of objectives needed

Differences :

- Text only vs multimodal
- Emphasis on diversity / representation quality

Should we :

- Merge them?
 - Keep them separate ?
 - Do something else ?
-

Let's brainstorm together !

4 breakout groups to discuss the Shared Tasks (ST):

1. How to **join** the two STs?

→ Which subtasks? How to organize a joint UniDive shared task? How to select or/and merge different tasks? Do you have other ideas of MWE/syntax-oriented tasks which answer the wishes above? How to optimally organize the community effort for these tasks?

2. How to organize the STs in a **sequential way**?

→ Same questions, but for 2 separate shared tasks

3. How to **evaluate the models generation**?

→ How to evaluate generation-oriented tasks? What is the necessary quantity of data? How to avoid contamination of models by the publicly available test data?

4. How to make the STs **more trendy**?

→ How to use the shared tasks to shed light on the universality of LLMs? How to make the tasks more interesting for the LLM community? Or rare language phenomena in general?

Reporting : 15 minutes at the end, ~3-4 minutes per group

PARSEME :

- Multiword identification in text
- Multilingual
- Use of diversity measures
- 2 subtasks to attract mainstream research

(LLMs)

- Generation task
- Comprehension task

ADMIRE

- Use multimodality to force better representations of **meaning**
- Subtask 1 : Sense identification for MWEs + selection of representative static image
- Subtask 2 : Complete image sequences representing MWEs + identify sense
- Text-only version available using image descriptions
- Possibly more appealing to computer vision researchers?

-
- Group 1 : Joining the shared tasks as one
 - Group 2 : How to organize these task separately ?
 - Group 3 : Generation tasks evaluation
 - Group 4 : How to attract the LLM community ?
-

It's conclusion time !