# "Fingers in the Nose": Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly Gamified Crowdsourcing Platform

**Karën Fort**
Sorbonne Université, STIH EA 4509
28, rue Serpente 75006 Paris, France
karen.fort@sorbonne-universite.fr

**Bruno Guillaume**
Université de Lorraine, CNRS,
Inria, LORIA, 54000 Nancy, France
bruno.guillaume@inria.fr

**Matthieu Constant**
Université de Lorraine, CNRS, ATILF
54000 Nancy, France
Mathieu.Constant@univ-lorraine.fr

**Nicolas Lefèbvre**
Université de Lorraine, CNRS,
Inria, LORIA, 54000 Nancy, France
nicolas.lefebvre@inria.fr

**Yann-Alan Pilatte**
Sorbonne Université
28, rue Serpente 75006 Paris, France
yann-alan.pilatte@etu.sorbonne-universite.fr

## Abstract

This article presents the results we obtained in crowdsourcing French speakers' intuition concerning multi-work expressions (MWEs). We developed a slightly gamified crowdsourcing platform, part of which is designed to test users' ability to identify MWEs with no prior training. The participants perform relatively well at the task, with a recall reaching 65% for MWEs that do not behave as function words.

## 1 Introduction and State of the Art

The identification of multi-word expressions (MWEs) is crucial in natural language processing (NLP) (Constant et al., 2017). Significant efforts have been made in recent years on the subject, in particular though the PARSEME international network (Savary et al., 2015). However, although some collective expert-based annotation initiatives have been successfully undertaken (Schneider et al., 2016; Savary et al., 2017), language resources are still limited in coverage and the need remains to identify newly-created MWEs. One potential solution is to exploit the so-called "wisdom of the crowd".

There have been several research papers on the interpretation of MWEs by native speakers, in particular by Gibbs (Gibbs, 1992; Gibbs et al., 1997). More recently, Ramisch et al. (2016) involved microworking crowdsourcing on Amazon Mechanical Turk. Finally, the experiment described in Krstev and Savary (2018) involves a gamified interface allowing MWE researchers to guess the meaning of opaque MWEs in other languages.

However, we could find no publication concerned with evaluating human ability to identify MWEs in a text, without taking their interpretation into account.

On the other hand, voluntary crowdsourcing, especially in the form of Games with a Purpose (GWAPs), has proven effective in terms of both the quantity and quality of the data produced. Successful examples of such platforms include JeuxDeMots (Lafourcade, 2007), Phrase Detectives (Poesio et al., 2013), and ZombiLingo (Guillaume et al., 2016).

We created a gamified platform named RigorMortis[1] (see Figure 1), the first of its kind, for MWEs annotation in French[2]. This platform includes a task enabling evaluation of the participants' intuition concerning MWEs, the results of which we present here.



[1] See: rigor-mortis.org

[2] We believe it is adaptable to any language.



*Proceedings of the Joint Workshop on*
*Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 207–213
Santa Fe, New Mexico, USA, August 25-26, 2018.


Figure 1: Interface for game selection.

For non-francophones, the phrase "Fingers in the nose" refers to the French idiom "Les doigts dans le nez" which means "without any difficulty", "with both hands behind one's back", etc.

## 2 Description of the Experiment

### 2.1 Reference corpus

As the participants are volunteers, we had to keep the intuition task short. We therefore created a rather small reference corpus. It is composed of ten sentences taken from articles on French political scandals from the French `Wikipédia`. Although their number is small, they were carefully selected to include MWEs corresponding to distinct identification criteria. The corpus was annotated and adjudicated in MWEs by experts. It contains 16 MWEs. One sentence, however contains no MWEs (see Table 1).

This reference corpus has been built following precise annotation guidelines inspired by those of the PARSEME shared task on verbal MWEs (Savary et al., 2017), which includes a French dataset (Candito et al., 2017). The criteria used for identifying MWEs, and in particular for detecting their morphosyntactic, syntactic and semantic idiosyncrasies are purely formal: for instance, no possible lexical substitution of a component by a synonym, presence of a "cranberry" word, no possible insertion of plausible material. We therefore discard semantically and syntactically compositional expressions that display statistical idiosyncrasy: for instance, institutionalized phrases in the sense of (Sag et al., 2001), like *traffic light*. Further, only fixed lexical components of the expressions are annotated, so final prepositions in MWEs that can be considered part of the MWE valency frame (like simple verbs) are not annotated. For instance, only *en raison* (because) is annotated in *en raison de* (because of) where *de* (of) is a preposition. Such subtlety is unknown and not natural for participants. This is why we took both variants into account when evaluating the task.

The reference MWEs can be divided into two subtypes: MWEs that behave as function words (later called *functional MWEs*) and MWEs that do not. In French, functional MWEs are mainly fixed in the sense of Sag et al. (2001). Such fully lexicalized expressions are immutable: they can undergo neither morphosyntactic nor syntactic variations, and insertion of plausible material is impossible. Notice that not all fixed expressions are necessarily functional MWEs: *dommages et intérêts* (damages), for example, functions as a noun. Furthermore, some of the functional MWEs that we consider are not entirely fixed: for instance, ***aux yeux*** *des enfants* (**in the eyes** of the children) = *à leurs yeux* (in their eyes). In our corpus, 7 MWEs are functional and 9 MWEs are not.

Our notion of functional MWEs can also be related to the category of fixed MWEs defined in the Universal Dependencies (Nivre et al., 2016) where the guidelines state that: *"[A fixed MWE] is used for certain fixed grammaticized expressions that behave like function words or short adverbials."*

| MWEs | Glosses | Translations |
|---|---|---|
| se voiler la face | to cover one's face | to bury one's head in the sand |
| file d'attente | queue of waiting | waiting line |
| dommages et intérêts | damages and interests | damages |
| mode de vie | way of life | way of life |
| le président de la République | the President of the Republic | the President of the Republic |
| ministre des finances | Minister of finances | Secretary of the Treasury |
| extrême droite | extreme right | far right (political) |
| chef d'État | chief of state | head of the State |
| mettre aux voix | put to the voices | put to the vote |
| entre autre | between other | among others |
| au-delà | further | beyond |
| en raison (de) | in reason (of) | because (of) |
| peut-être | may-be | maybe |
| aux yeux (de X) | at the eyes of X | in X's view |
| d'ailleurs | from elsewhere | by the way |
| dans le but (de) | in the goal (of) | in order to |

Table 1: Glosses and translations of the MWEs of the experiment.

### 2.2 Crowdsourcing platform

We ask the participants to find "expressions multi-mots" (multi-word expressions) and give them a couple of examples, explaining that MWEs are non-compositional. We also mention that more "functional" expressions can also be MWEs and should be annotated. The participants are then directly asked to identify the MWEs in the sentences we propose, without any prior training.

The interface, inspired by that of `TileAttack` (Madge et al., 2017), allows users to annotate multiple and discontinuous MWEs (see Figure 2). It should be noted that the intuition task is part of a larger gamified platform and that, while the participants did not gain points in this phase, they did in the other phases (see Figure 1).


Figure 2: Interface for the annotation of the MWEs.

It is important that, during this phase, we give no feedback to the participants on their annotations. They can see feedback concerning their results only once they are finished annotating all ten sentences (see Figure 3), so that they are not biased.

The crowdsourcing interface was publicized mainly on social networks and natural language processing lists.

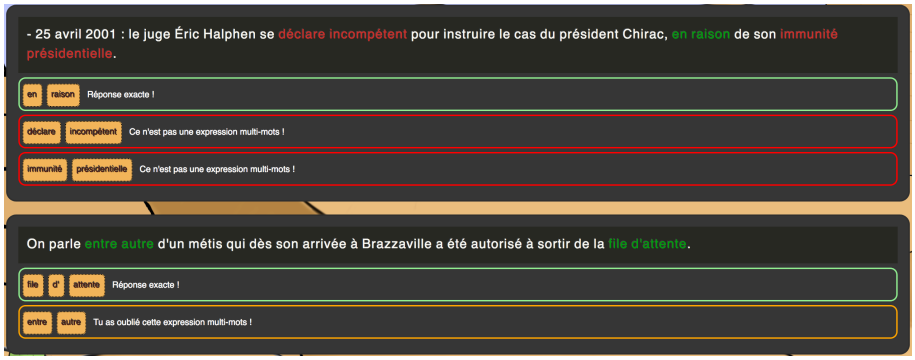


Figure 3: Feedback given to the participant *after* the intuition phase.

## 3 Results Obtained

### 3.1 Global Results

The ten reference sentences were played by 65 to 68 players (66.8 participants on average)[3].

Table 2 shows the precision, recall and F-measure for the whole annotation with two settings:

- **Perfect** match: the player selected exactly the same tokens as in the reference;
- **Approximate** match: the difference between the selection of the player and the reference includes only function words; for instance, [*président+République*] is accepted for [*président+de+la+République*], as *de* (of) is a preposition and *la* (the) is a determiner.

| | Precision | Recall | F-measure |
|---|---|---|---|
| Perfect | 48.13 | 41.22 | 44.41 |
| Approximate | 58.22 | 49.86 | 53.72 |

Table 2: Global results.

It should be noted that 40 out of 68 participants (i.e. 58.82% ) correctly annotated the sentence which contains no MWE.

These global results show that when the participants identify an MWE, they are often right (in 58.22% of the cases), but that they are less good at finding them all (less than 50% were found).

### 3.2 Results for Individual MWEs: the Impact of Functional MWEs

To determine if an MWE is more or less easy to find, we computed the recall for each MWE separately (see Figure 4).

Again, we show the two values: i) perfect (dark blue) and ii) approximate (light blue) match.

We observe in Figure 4 that there is a significant difference between the subset of *functional* MWEs (on the left) and that of *none-functional* MWEs (on the right).

Table 3 gives the recall value for these two subsets. The last column gives the overall value (already given above) for comparison. The recall for non functional MWEs reaches 65.05, which is more than twice that of functional ones (30.41).

In our own experience (and that of some participants), this difference in the identification of functional MWEs arises because we are so accustomed to them (they are so familiar) that we simply do not "see" them and forget to annotate them.

These results are encouraging, as they show that the participants can be rather efficient at identifying at least some types of MWE.

[3]We removed from this analysis four participants who are experts in the MWE subdomain.However, we kept the participants from the more general NLP and linguistics domains.

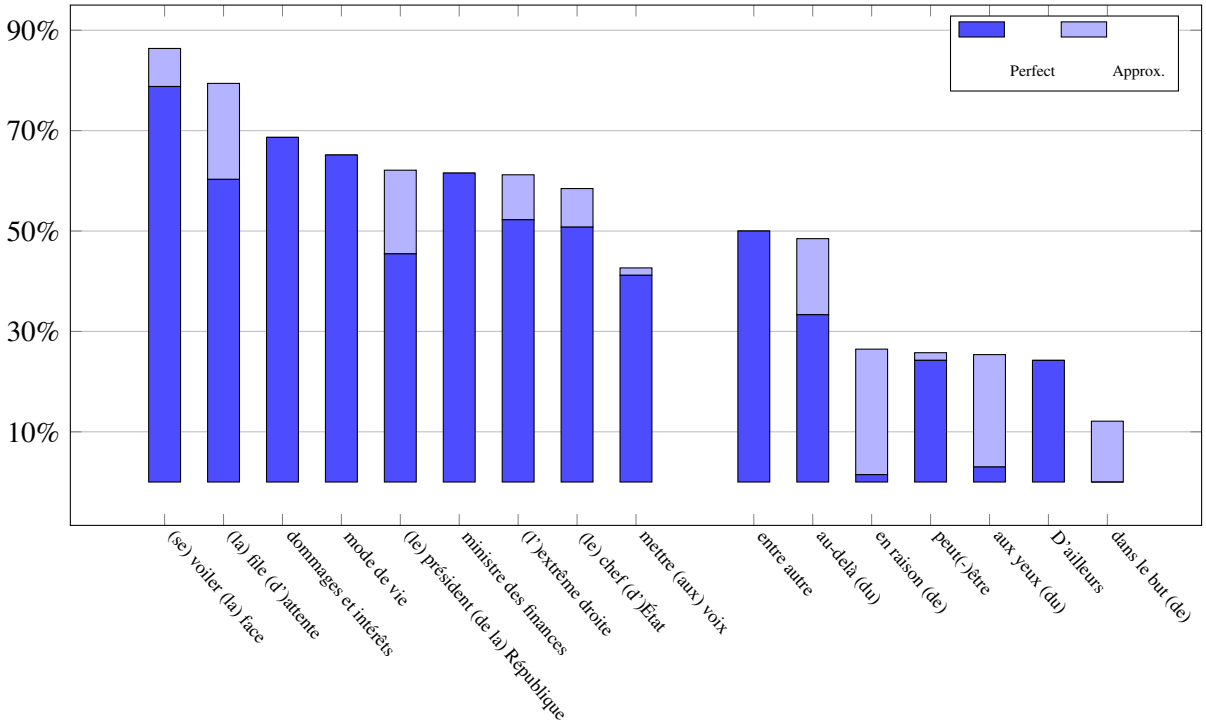


Figure 4: Recall of the participants on each MWE (left: non-functional; right: functional).

| | *non-functional* | *functional* | All |
|---|---|---|---|
| Perfect | 58.19 | 19.48 | 41.22 |
| Approximate | 65.05 | 30.41 | 49.86 |

Table 3: Recall for functional and non-functional MWEs.

### 3.3 Analysis of the Noise Produced

Table 4 lists the ten expressions that were identified by more than 10% of the participants and which are not in the reference.

Note that we reserve the term collocation to refer to any statistically significant co-occurrence, including all forms of MWEs as described above and compositional phrases which are predictably frequent

It shows that, unsurprisingly, the participants had difficulty distinguishing between MWEs and compositional expressions exhibiting statistical idiosyncrasy (in six cases). This can be explained by the fact that our definition of MWEs partially overlaps with the notion of collocation, as defined in (Sag et al., 2001). Collocations refer to "any statistically significant co-occurrence", that includes both syntactically/semantically compositional and non-compositional expressions.

Other mistakes include boundary errors (two cases) and common civilities annotated as MWEs (two cases). These could probably be avoided if the participants were properly trained.

## 4 Conclusion and perspectives

Although it was carried out on a small corpus[4], this experiment gathered results from a satisfying number of participants and showed that volunteers with no prior training can help identify at least some MWEs in texts. It is encouraging that the most difficult MWEs to find are the functional ones, as these are usually the first to be listed and are the least prone to neologism.

However, while the participants' intuition proves valuable, it should be complemented by proper training, using at least some of the tests defined by the PARSEME network.

[4]A way to increase the size of the corpus without making the task longer could be to randomize the sentences proposed to the participants.

| Noisy expressions | % of participants | Comment |
|---|---|---|
| immunité présidentielle (presidential immunity) | 55.88% | collocation (not a MWE) |
| élection présidentielle (presidential election) | 34.85% | collocation (not a MWE) |
| aux yeux du public (to the eye of the public) | 28.36% | boundary error |
| destin tragique (tragic faith) | 23.88% | collocation (not a MWE) |
| Monsieur le Président (Mister President) | 19.70% | common civilities |
| affaire politique (political scandal) | 15.38% | collocation (not a MWE) |
| chers collègues (dear colleagues) | 15.15% | common civilities |
| instruire le cas (investigate the case) | 14.71% | collocation (not a MWE) |
| se déclare incompétent (withdraw from the case) | 11.76% | collocation (not a MWE) |
| aux voix (to the vote) | 10.29% | boundary error |

Table 4: Identified expressions which were not in the reference.

A complementary experiment on the subject is work in progress, as the platform also enables researchers to train participants and collect annotations from them.

The reference corpus is freely available under a CC BY-SA license, on the platform itself.

## References


Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Yannick Parmentier, Caroline Pasquer, and Jean-Yves Antoine. 2017. Annotation d'expressions polylexicales verbales en français. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, pages 1–9, Orléans, France, June.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Raymond W. Gibbs, Josephine M. Bogdanovich, Jeffrey R. Sykes, and Dale J. Barr. 1997. Metaphor in idiom comprehension. *Journal of Memory and Language*, 37:141–154.

Raymond W Gibbs. 1992. What do idioms really mean? *Journal of Memory and Language*, 31(4):485 – 506.

Bruno Guillaume, Karën Fort, and Nicolas Lefebvre. 2016. Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING): Technical Papers*, pages 3041–3052, Osaka, Japan, December.

Cvetana Krstev and Agata Savary. 2018. Games on multiword expressions for community building. *INFOtheca: Journal of Information and Library Science*, February.

Mathieu Lafourcade. 2007. Making people play for lexical acquisition. In *Proceedings of the 7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thailand, December.

Chris Madge, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2017. Experiment-driven development of a gwap for marking segments in text. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*, pages 397–404. ACM.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia, May. European Language Resources Association (ELRA).

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, Aline Villavicencio, and Rodrigo Wilkens. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword Expressions: A Pain in the Neck for NLP. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pages 1–15.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznan, Poland, November.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain, April. Association for Computational Linguistics.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). In *Proceedings of SemEval*, San Diego, California, USA, June.