Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4395–4401
Marseille, 11–16 May 2020

Rigor Mortis: Annotating MWEs with a Gamified Platform

Karën Fort[*†], **Bruno Guillaume**[†], **Yann-Alan Pilatte**[*], **Mathieu Constant**[‡], **Nicolas Lefèbvre**[†]
[*]Sorbonne-Université / STIH, [†]Université de Lorraine, CNRS, Inria, LORIA, [‡]Université de Lorraine, CNRS, ATILF
28, rue Serpente 75006 Paris, France, 54000 Nancy, France
karen.fort@sorbonne-universite.fr, bruno.guillaume@inria.fr, yann-alan.pilatte@etu.sorbonne-universite.fr,
mathieu.constant@univ-lorraine.fr, nico.nicolef@gmail.com

Abstract

We present here `Rigor Mortis`, a gamified crowdsourcing platform designed to evaluate the intuition of the speakers, then train them to annotate multi-word expressions (MWEs) in French corpora. We previously showed (Fort et al., 2018) that the speakers' intuition is reasonably good (65% in recall on non-fixed MWE). After a training phase using some of the tests developed in the PARSEME-FR project, we obtain 0.685 in F-measure at an experimentally determined 25% threshold (number of players who annotated the same segment).

Keywords: MWE, crowdsourcing, games with a purpose

1. Motivations

Multiword expressions (MWEs) can roughly be defined as combinations of several words that encompass some composition irregularity at one or several linguistic levels, including morphology, syntax and semantics. For instance, the meaning of the expression *cut the mustard* (succeed) cannot be derived from the meaning of its parts. Multiword expressions include multiple linguistic phenomena, as mentioned in (Sag et al., 2001), for example idioms (e.g. *add fuel to the fire*), phrasal verbs (e.g. *give up*), complex function words (e.g. *as soon as*), light-verb constructions (e.g. *take a bath*), adverbial and nominal open compounds (e.g. *by the way*, *dry run*).

Handling MWEs is a key challenge for natural language processing (Sag et al., 2001), on which researchers have long been working. Recently, significant advances have been made thanks to the availability of new MWE-annotated data that are used to learn state-of-the-art identification models (Schneider et al., 2016; Ramisch et al., 2018).

The construction of such annotated corpora is nonetheless costly. Indeed, they are mainly annotated by experts or linguistics-aware people long-trained by experts in order to guarantee the annotation quality. Indeed, MWEs are known to be hard to identify due to the fuzzy delimitation between compositional and non-compositional combinations of words. This difficulty is demonstrated in the modest average inter-annotator agreement (0.65 in F-score) for comprehensive MWEs annotation in English (Schneider et al., 2014).

In this paper, we propose a gamified platform for annotating MWEs. Experiments were carried out for French. The aim of this paper is to assess to what extent one can rely on corpora annotated in MWEs by the participants with respect to experts.

2. Related Work

The creation of linguistic resources for natural language processing using games with a purpose (GWAPs) has been tested for a while now. The first to be developed were `JeuxDeMots`, a game allowing the creation of a lexical network for French, which is more than ten years old now (Lafourcade, 2007; Lafourcade et al., 2018), closely followed by `Phrase Detectives` (Chamberlain et al., 2008), in which participants annotate co-reference relations in English corpora. Both games are still running and collecting language data as we write these lines.

Other GWAPs were then designed, addressing new tasks, like `ZombiLingo` (still running) for the annotation of dependency relations for French (Guillaume et al., 2016) or `Wordrobe` (no more active), for various semantic annotation tasks (Bos and Nissim, 2015).

Most of the active GWAPs in the domain now appear on the LDC LingoBoingo portal[1]. Apart from the already mentioned games, it presents `TileAttack` (Madge et al., 2017), `Wormingo` (Kicikoglu et al., 2019), `WordClicker` (Madge et al., 2019), `Name That Language!` (described as the `Language ID game` in (Cieri et al., 2018)) and `Know Your Nyms?`.

Many of these GWAPs were quite successful, both in terms of the quantity of created data and of the obtained quality (Chamberlain et al., 2013). However, despite the wide variety of available GWAPs, there is, to our knowledge, no other active (and open) gamified platform or game dealing with MWE identification. The only related work we found is a gamified interface which was developed as part of the PARSEME COST action[2], allowing selected participants (researchers) to guess the meaning of opaque MWEs in other languages (Krstev and Savary, 2018).

3. Rigor Mortis

3.1. A gamified platform

`Rigor Mortis`[3] is a gamified crowdsourcing platform, which aim is to allow the participants to annotate MWEs

[1] `https://lingoboingo.org/`.
[2] `https://typo.uni-konstanz.de/parseme/`
[3] `http://rigor-mortis.org`

in corpora. We instantiated it for French, but, in principle, it can be adapted to any language, provided the tests described in Section 3.3. are adapted.
The platform integrates different phases that the user can unlock in sequence:

1. the intuition test
2. the training phase
3. the annotation itself

The players gain points only from the third phase of the game.
The gamification layer is quite light for the moment, with a leaderboard and a very simple scenario revolving around the exploration of pyramids.[4] The design is thus inspired from the world of ancient Egypt (see Figures 1 and 4).

Figure 1: `Rigor Mortis`: phase 1 and leaderboard.

The annotation interface of the platform is directly inspired from that of `TileAttack` (Madge et al., 2017)[5], offering the possibility to annotate discontinuous segments (see Figure 2) and to include the same token in different annotated segments.

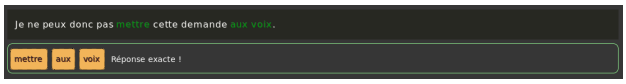

Figure 2: Annotation of a discontinuous segment (intuition phase).

3.2. Intuition phase

In order to prevent the participants from being influenced by the training phase, the intuition phase is the first step in `Rigor Mortis`. The players are asked to identify MWEs without any prior training and very little help, only a short description of what a MWE is (expressions made of several words, more or less fixed and non-compositional) and a couple of examples.
During this phase, the participants have to annotate ten sentences, taken from a corpus made from French political scandals Wikipedia articles, the *Affaires* corpus. These sentences were selected carefully for their simplicity and brevity: we did not want the players to feel bored and run away too quickly. In order to prevent bias, they do not receive any feedback on their annotations during this phase, but they can see how they performed once done (see Figure 3).

[4]We had more ideas, but had to renounce implementing them due to a lack of resources.

[5]`https://tileattack.com/`.

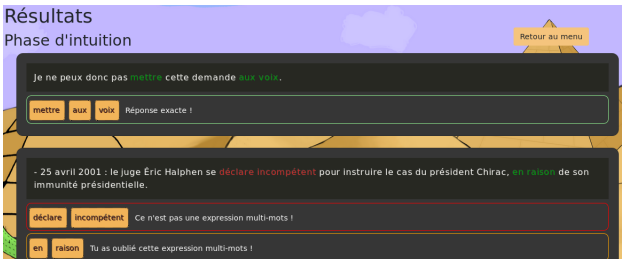

Figure 3: Intuition phase: the results can only be checked once finished.

This particular phase of the game has been presented in a previous publication (Fort et al., 2018), so we will not detail it here, except to mention that the intuition of the speakers is rather good for non-fixed MWEs (65% of recall).

3.3. Training phase

The training phase was split in steps, or "corridors" of the pyramid, each corresponding to a linguistic test allowing to identify one type of MWE. As we did not want the training phase to be too burdensome, both in terms of time and difficulty, we limited ourselves to five of the easiest and most productive tests from the PARSEME-FR project[6], which includes ten tests for the identification of non-verbal MWEs.[7]
For each test, we purposely crafted two short sentences, with only one possible annotation, so that the players can easily apply their new knowledge, gain confidence, and not be confused upon completing the training phase. During this phase, if the player gives a bad answer, their error is notified to them.
The five tests were selected for their productivity, their coverage of the different types of MWEs and their ease of comprehension. They encompass the following[8]: replacement [LEX], identification of "cranberry" terms [CRAN], insertion [INSERT], morphosyntactic features (such as singular/plural) [MORPHO] and "zero" determiner [ZERO]. All these tests try to capture some fixity of the expressions, that tend to entail idomaticity as productive composition rules can not be applied. In particular, the fixedness of a MWE can be identified by applying a transformation on the given MWE that leads to an unexpected meaning shift or an unacceptable sequence, with respect to a regular setting.

3.3.1. Replacement test [LEX]

We use the replacement test to show two properties of MWEs to our participants. Firstly, that one part of a MWE cannot be replaced by another word with a similar meaning or function (be it a synonym, a hypernym or analogous

[6]https://parsemefr.lis-lab.fr

[7]`https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/wikis/Guide-annotation-PARSEME_FR-chapeau`.

[8]We use the same codes as the PARSEME-FR project, between brackets.

linguistic items), without either modifying its meaning, or making it meaningless. Secondly, that the 'head' of a MWE is not itself the hypernym of that expression. In the following example, "eau de vie" (brandy) cannot be replaced by "liquide de vie" (liquid of life) and it is not "eau" (water) either, it is in fact alcohol :

Tu prendras bien un peu d'eau de vie.
You will take at least a little of water of life.
'I'm sure you'll take at least a sip of brandy.'

Figure 4: Training on the replacement test.

3.3.2. Cranberry test [CRAN]

The cranberry test consists in identifying terms which cannot exist outside of the MWEs they are part of. In the example below, "perlimpinpin" is such a word in "poudre de perlimpinpin" (snake oil):

[...] C'est de la poudre de perlimpinpin.
[...] This is powder of perlimpinpin.
'[...] This is snake oil.'

3.3.3. Insertion test [INSERT]

The insertion test should fail with most MWEs. When trying to insert an adjective or an adverb, the sentence would have a different meaning, or not mean anything at all, as in "Luc prend la puissante mouche" (Luc takes the powerful fly). Note that we can apply this test only in cases where it is possible to insert an adjective or an adverb in a similar regular linguistic context.

Luc prend la mouche.
Luc takes the fly.
'Luke flies off the handle.'

3.3.4. Morphosyntactic test [MORPHO]

The morphosyntactic test exposes how changing the number or the gender in the candidate MWE is either impossible, or changes the meaning of the sentence. In the following example, "ramener ses fraises" (plural) is possible in French, but with another, literal, meaning.

[...] Je savais qu'il ramènerait sa fraise.
[...] I knew he would bring back his strawberry.
'[...] I knew he would show up.'

3.3.5. Zero determiner test [ZERO]

The zero determiner test presents verbal expressions with no determiner. In French, one cannot "prêter la main forte" (to lend the strong hand):

Il prête main forte à ces gens [...].
He's lending hand strong to these people [...].
'He's lending these people a helping hand [...].'

3.4. Annotation phase

The annotation phase can only be unlocked once the intuition and training phases have been completed. As shown in Table 1, this phase allows to play 504 sentences from the *Affaires* corpus, made from Wikipedia pages on French political scandals. The same corpus was used for the intuition phase.
The bonus phase was added at our most productive player's request. It includes 743 sentences from Wikipedia pages on strikes and demonstrations against *Loi travail* (Law on work). We include it in the annotation phase, as it is merely an extension of the third phase.
The annotation phase contains 19 "control" sentences, i.e. reference sentences, annotated by two of the co-authors, which we use to check that the participant still remembers the training (see Figure5).

Figure 5: Feedback given on a wrong answer on a control sentence during the annotation phase: *encore venu* (come again) is not a MWE

.

In case of an error on a control sentence, the next sentence will also be a control sentence, and so on, until the given answer is the right one. Hopefully, with this system, the player will either learn about MWEs or get bored and renounce playing.

4. Participation

We advertised the platform on social networks and on the French NLP mailing list, LN.
It was therefore not entirely surprising to us that some colleagues specializing on the subject participated in the game. In particular, two members of the PARSEME-FR project appeared rapidly in the leaderboard.
As can be seen from the homepage of the platform, 121 persons registered and while 65 did not go further than the training and did not get any point, 57 users scored at least 1 point. However, we created two fake players to bootstrap the game, therefore there are in fact 55 real participants in the annotation part of the platform.

Phase	Source	Nb sentences	Nb tokens
Intuition (1)	Wikipedia *Affaires*	10	268
Training (2)	ad hoc	10	112
Annotation (3)	Wikipedia *Affaires*	504	16,753
Bonus annotation (4)	Wikipedia *Loi travail*	743	25,067

Table 1: Corpora used in `Rigor Mortis`.

One participant, `Wellington`, managed to annotate all the proposed sentences and ask us for more. We therefore added a bonus level, which was mainly played by two players, `Wellington` and `Methos31`.

5. Evaluating the annotations

5.1. Creating a reference *a posteriori*

62 sentences were played by at least two members of the PARSEME-FR project, including 32 by three of them. They did not always agree in their annotation, so we adjudicated the obtained results to create a reference.
It has to be noted that M. Constant participated both in the game as a PARSEME-FR expert and in the adjudication as a co-author of this paper. However, a couple of months passed in the meantime and he did not recall his annotations. Moreover, the adjudication was done collaboratively with all the co-authors.
The 62 reference sentences contain 61 MWEs with the distribution presented in Table 2.

Nb of MWEs	Nb of sentences with so many MWEs
No MWE	26
1 MWE	20
2 MWEs	9
3 MWEs	5
4 MWEs	2

Table 2: Distribution of MWEs in the reference sentences.

Unsurprisingly, no sentence contain more than four MWEs and very few of them (two) contain four MWEs. On the opposite, 26 sentences out of 62 do not contain any MWE. These reference sentences were played by between 22 and 51 different players with an average of 31.48 players by sentence.

5.2. Taking variations into account

Even when a MWE is recognised in a sentence, it is often difficult to decide the precise set of tokens that belongs to it. For instance, in the reference data, there is a sentence *Les inventeurs ont l'habitude de démontrer l'efficacité* ... [Inventors are used to demonstrate the efficiency...] with a MWE *avoir +habitude* [have + habit]. Following PARSEME-FR guidelines, the determiner *l'* is not part of the MWE because some examples like *Ils ont cette habitude* [They have this habit] can be found where the same MWE is present without the same determiner.
It was also decided in PARSEME-FR not to include the final preposition in MWEs. Thus, instead of annotating *au détriment de* [to the detriment of], the annotators are supposed to restrict the annotation to *au détriment*. One of the motivation of this choice is that one can find occurrences of this MWE followed by coordination like *au détriment de X et de Y*. If the preposition is included in the MWE, there is no easy way to deal with this case and to decide what to do with the preposition *de* just before *Y*.
In order to take these difficulties into account, two measures are used:

Exact an annotation from a player is considered as correct if it contains exactly the same set of tokens as in the reference.

Approx an annotation from a player is considered as correct if it contains exactly the same set of semantically full words: only nouns, verbs, adjectives and adverbs are considered in the comparison.

In the *Approx* setting, two annotations which differ only by some determiners, prepositions or other function words are considered as identical. In the two examples presented above, annotations *au+detriment+de* and *au+detriment* are equivalent for the **Approx** setting; similarly, *avoir + habitude* and *avoir + l' + habitude* are equivalent.

5.3. Empirically determining a threshold

Although the participants are trained, our platform is a crowdsourcing one, which means we have no prior knowledge of who is going to participate and how well they will perform. In order to benefit from the "wisdom of the crowd", we need to establish the minimum number of agreeing players necessary to obtain the best results.
Given a threshold α, the scores are computed as follows:

1. Projection:
   - For **Approx** scores, each annotation (a list of tokens of the sentence) from the reference and from the players is projected on the sub-list containing only the semantically full tokens.
   - For **Exact** scores, the full list of tokens is kept.
2. For each sentence played by n players, we keep only the annotations identified by at least $\alpha \cdot n$ players.
3. We use the usual metrics of precision, recall and F-measure to evaluate the quality of the annotations obtained at the previous step as compared to the reference ones.

Again, it should be reminded here that we consider as a full-fledged annotation the fact that a player chose not to

select anything and clicked the `Valider` (Validate) button (no MWE).
No threshold can be fixed *a priori*, so we observe below on the reference data (Figure 7 and 8) how this threshold impacts the quality of the produced resource.

6. Obtained results

6.1. Produced annotations

The participants identified 13,387 annotations in the annotation phase itself (15,693 in both the intuition and the annotation phase).
As shown in Figure 6 and as usual in voluntary crowdsourcing (Chamberlain et al., 2013), very few participants produced a lot of data. In our case, two annotators produced nearly 2,000 annotations (i.e. around 14%) each.

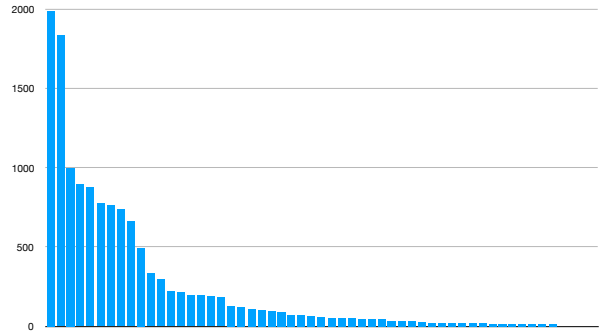

Figure 6: Number of annotations produced by player.

It has to be noted that we considered as an annotation the fact that a participant chose not to annotate anything in a sentence, explicitly pressing the `Validate` button without selecting any text segment.
Obviously, a number of the produced annotations are not real MWEs (noise), whereas some real MWEs were not identified (silence).

6.2. Quality of the produced annotations

3,124 annotations were produced on the reference sentences by all the participants, among which 229 were played by the three PARSEME-FR participants who produced 92, 90 and 47 annotations.
Below, we consider the 2,895 remaining annotations (produced by annotators who are not PARSEME-FR members), which were added to 1,952 sentences.
Figure 7 describes how the annotation scores of the players against the reference (precision, recall and F-measure) evolve for different values of the threshold we defined previously.
In the **Exact** setting (Figure 7), at the threshold 0% (no annotations are filtered out, so all MWEs annotated by at least one player are considered), we observe a high recall at 0.956 but a very low precision 0.138). This means that most of the MWEs of the reference are found but with a lot of noise. The best F-measure (0.618) is obtained with a 25% threshold. The precision becomes greater than the recall at a 35% threshold.
Figure 8 describes the same observation but with the **Approx** setting. In this case, at threshold 0%, the recall is

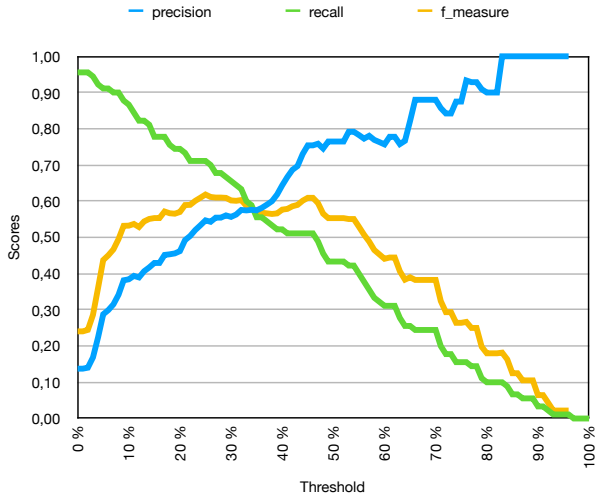

Figure 7: Exact scores depending on the threshold.

0.989 and the precision 0.170. The maximum F-measure, 0.685, is obtained at 25% threshold. The balanced value for precision and recall is obtained at a 36% threshold with 0.622.

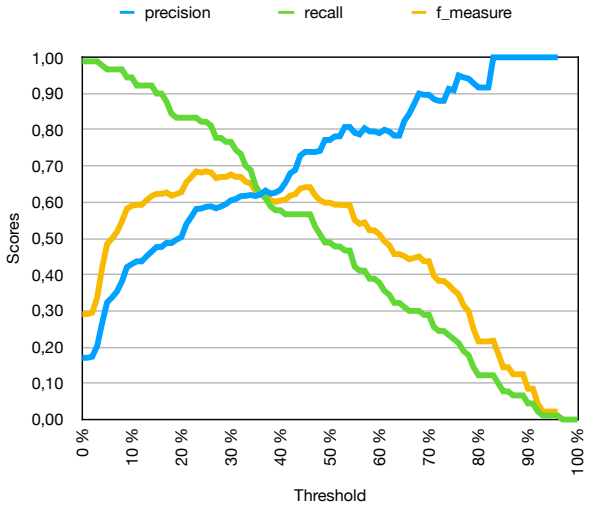

Figure 8: Approx scores depending on the threshold.

In the **Approx** setting, only one MWE of the reference is not annotated by any player. It is *pertes+subies* in *des pertes financières directes et subies de plus de 750 millions de francs* [direct and incurred financial losses of more than 750 million francs] which is an instance of the light verb construction *subir+pertes* [suffer+loss]. However, this one is very difficult to notice because the two tokens of the MWE are separated by three other words which are not included in the construction.
In Figure 9, we report the same scores for the PARSEME-FR experts annotations.[9] The four different values for the

[9]We report only the **Exact** measure for the expert; the **Approx** measure is very close and there is no significant difference

F-score are: 0.752 (threshold 0% to 33%), 0.788 (threshold 34% to 50%), 0.763 (threshold 51% to 66%) and 0.594 (threshold 67% to 100%). This indicates that experts obtained significantly higher scores at the same thresholds than other players. Nevertheless, for high thresholds (i.e. for cases where all experts are consistent), the F-score is closed to 0.6, which confirms the difficulty of the task.

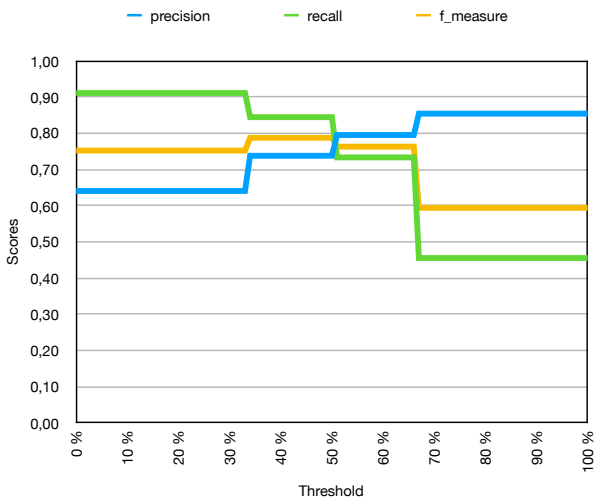

Figure 9: Exact scores depending on the threshold (for experts).

6.3. Impact of the training

Comparing the results between the intuition phase and the actual annotation phase performed post training could let one think that the training had very little impact. However, the sentences played in the intuition phase were very short and easy to annotate on purpose: we did not want to make players feel like this task would be too hard before they even started.

While the approach we took seemed reasonable through the GWAP lens, we cannot really reap the fruits of seeing the benefits of the training phase, because the annotation corpus was overall harder to play that the intuition phase sentences. The main take-outs when having a look at the results would be that a training was missing on verbal MWEs. For example, light verb constructions and reflexive verbal MWEs were not easily detected by our players, who could have benefited from more instructions on the matter.

7. Conclusion

We present in this article the results obtained with our gamified platform on MWEs identification. With the **Approx** measure and a well-chosen threshold, we obtained an F-measure of 0.685. Considering the difficulty of the MWE identification task as shown by the low inter-annotator agreement (0.65 F-score) reported in (Schneider et al., 2014), we believe that our first experiments shows that the crowdsourcing approach for MWEs identification is relevant and represents a valid option to build new annotated resources.

between the two settings.

The gamification of the platform should be improved, in order to attract and retain more players. In addition to usual gamification features, we could also add more interaction among the users, for instance with a new game mode where a user have to correct the output of another player (`Phrase Detectives` proposes a similar feature) or places where players can discuss about their annotations (as in the `ZombiLingo` forum).

As future work, we would like to use the data produced by the players of `Rigor Mortis` to quantify a degree of fixity of a MWE, depending of the percentage of players who annotated it.

Another interesting perspective would be to adapt `Rigor Mortis` to other languages. The PARSEME project defined general annotation guidelines[10] (with restriction to verbal MWEs) and applied them to 27 languages. In order to adapt `Rigor Mortis` for a new language, an extension of the PARSEME work on verbal MWEs to all MWEs is needed; the PARSEME-FR being a first step in this direction.

The code of `Rigor Mortis` as well as the created resources, are freely available on GitHub.[11]

8. Acknowledgements

We wish to thank all the participants in `Rigor Mortis`, especially Agata Savary and Carlos Ramisch, from the PARSEME-FR project, as well as `Wellington` and `Methos31`, our highly competitive best players. The game platform is hosted in the CPER LCHN (Contrat de Plan État-Région - Langues, Connaissances et Humanités Numériques) infrastructure. This research has been partially funded by the European Network for Combining Language Learning with Crowdsourcing Techniques[12] (enetCollect) COST action and the French Agence Nationale pour la Recherche, through the PARSEME-FR project (ANR-14-CERA-0001).

9. Bibliographical References

Bos, J. and Nissim, M. (2015). Uncovering noun-noun compound relations by gamification. In Beáta Megyesi, editor, *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 251–255.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Addressing the resource bottleneck to create large-scale annotated texts. In *STEP '08: Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 375–380, Morristown, NJ, USA. Association for Computational Linguistics.

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013). Using games to create language resources: Successes and limitations of the approach. In Iryna Gurevych et al., editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.

[10] `https://parsemefr.lis-lab.fr/parseme-st-guidelines/`

[11] `https://github.com/gwaps4nlp/rigor-mortis/tree/master/LREC_2020.`

[12] `https://enetcollect.eurac.edu/`

Cieri, C., Fiumara, J., Liberman, M., Callison-Burch, C., and Wright, J. (2018). Introducing nieuw: Novel incentives and workflows for eliciting linguistic data. In *Proc. of LREC 2018: 11th Edition of the Language Resources and Evaluation Conference,*, Miyazaki, Japan, May.

Fort, K., Guillaume, B., Constant, M., Lefèbvre, N., and Pilatte, Y.-A. (2018). "Fingers in the Nose": Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly Gamified Crowdsourcing Platform. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 207 – 213, Santa Fe, United States, August.

Guillaume, B., Fort, K., and Lefebvre, N. (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.

Kicikoglu, D., Bartle, R., Chamberlain, J., and Poesio, M. (2019). Wormingo: A 'true gamification' approach to anaphoric annotation. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*, FDG '19, pages 75:1–75:7, New York, NY, USA. ACM.

Krstev, C. and Savary, A. (2018). Games on multiword expressions for community building. *INFOtheca: Journal of Information and Library Science*, February.

Lafourcade, M., Joubert, A., and Le Brun, N. (2018). The jeuxdemots project is 10 years old: What we have learned. In *Proceedings of the LREC Games4NLP workshop*, Miyazaki, Japan, May.

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

Madge, C., Chamberlain, J., Kruschwitz, U., and Poesio, M. (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.

Madge, C., Bartle, R., Chamberlain, J., Kruschwitz, U., and Poesio, M. (2019). Incremental game mechanics applied to text annotation. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, pages 545–558, New York, NY, USA. ACM.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018). Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (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.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., and Smith, N. A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Schneider, N., Hovy, D., Johannsen, A., and Carpuat, M. (2016). SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California, June. Association for Computational Linguistics.