Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4205–4211
Marseille, 20-25 June 2022

Assessing the Quality of an Italian Crowdsourced Idiom Corpus: the Dodiom Experiment

Giuseppina Morza, Raffaele Manna, Johanna Monti
UNIOR NLP Research Group - University of Naples "L'Orientale"
Via Duomo 219, Naples, Italy
{gmorza, rmanna, jmonti}@unior.it

Abstract

This paper focuses on the evaluation of linguistic data, concerning idioms examples collected and annotated through Dodiom, a GWAP environment, by Italian linguists. The paper provides an insight into the Dodiom project, the data collection through the contribution of the crowd, and, finally it specifically describes the annotation criteria used by the experts to estimate the quality of the collected data. The main scope of this paper is, indeed, the evaluation of the quality of the linguistic data obtained through crowdsourcing, namely to assess if the data provided by the players who joined the game are eligible and profitable for research and teaching purposes. This task concerns the development of a collection of idioms, namely a specific type of Multiword expressions which is usually hard to find in corpora and that contains words that may also be used in their literal meanings within a sentence. This is particularly important as these data may be used both for the training and the evaluation of NLP applications. Finally, results, as well as future work, are presented.

Keywords: crowdsourcing, data quality, idioms

1. Introduction

Natural Language Processing (NLP) has made significant advances in recent years, due to the introduction of statistical machine learning techniques. Human annotators, language experts, obviously play an essential role in the process of building language corpora, required to train statistical learners. However, this unavoidably results in high annotation costs and limited access to qualified annotators, which is a major hindrance to NLP research. As shown by popular platforms like Wikipedia, Duolingo among others, crowdsourcing presents a unique opportunity to obtain massive amounts of data in a relatively short time. The research task described in this contribution is based on (Eryiğit et al., 2021), who resorted to increasingly popular methodologies for data collection, namely crowdsourcing and gamification. The scholars developed a Game-with-a-purpose (GWAP) named Dodiom to elicit idiom examples and ratings from non-expert labelers reached via Facebook, Linkedin and word-of-mouth. For instance, given a MWE, Dodiom asks the non-expert user to create a sentence in which the MWE is used idiomatically or not and to provide a label related to one of the two uses (literal or idiomatic).

Our aim was to assess by means of a subsequent annotation task of the collected textual samples if the data obtained from non-experts (the crowd) in a cost-effective and time-saving way were valuable resources that could effectively be used as training and testing data for language understanding, in general, and idiom identification systems, as well as language learning material, or samples for lexicographic studies. In this paper, we describe the state of the art in idiom corpora development and gamification in Section 2. We then present the corpus collection in 3.1. In 3.2 we provide a description of the guidelines adopted for the data annotation task. The steps for evaluating the annotated data along with the results are shown in 4. Finally, we provide conclusions and a discussion on possible future work in 5.

2. Related Work

Games-with-a-purpose have come to play a central role in the development of linguistic resources for NLP (Chamberlain et al., 2013).

Gamification turned out to be an effective tool for the collection of language resources in a cost-effective way, proving on numerous occasions that even non-experts can successfully provide valuable scientific data. Snow et al. (2008) demonstrated the effectiveness of using crowdsourcing for a variety of natural language annotation tasks. In the case of crowdsourcing, the main strategy for achieving good quality in labeling is to aggregate results from many users to approximate the judgments of a single expert (Snow et al., 2008).

Below, we briefly outline the main resources created using GWAPs; we will then focus on reporting the methods used to evaluate aggregated data in the creation of Multiword expressions (MWEs) linguistic resources by means of crowdsourcing.

GWAPs have been popularized in the NLP field by early initiatives such as 1001 Paraphrases (Chklovski, 2005), Phrase Detectives (Chamberlain et al., 2008) and Dr. Detective (Dumitrache et al., 2013). One of the first GWAPs created and used to collect linguistic data, more specifically, corpora, was 1001 Paraphrases (Chklovski, 2005). In this game, participants were asked to produce paraphrases of expressions.

In Phrase Detectives (Chamberlain et al., 2008) participants annotate anaphoric coreference among phrases in sentences taken from English Wikipedia articles and narrative texts. Using ZombiLingo participants perform the annotation of dependency relations for French (Guillaume et al., 2016; Fort et al., 2014).

Idioms, that are at the *core* of this research project, are seen as a subcategory of multiword expressions (MWEs) which have been subject to many initiatives in recent years such as the Parseme EU COST Action (Savary et al., 2015; Savary et al., 2017), the MWE-LEX workshop series (Markantonatou et al., 2020) and the ACL special interest group SIGLEX-MWE (Cook et al., 2021). Traditional methods of collecting and annotating MWEs generally rely on using textual samples taken from large text corpora, with an effort by the team of researchers in manual annotation of multiword phenomena (Schneider et al., 2014; Savary et al., 2018). However, the scarcity of MWEs (especially idioms) in texts has presented obstacles to corpus-based studies and NLP systems addressing these specific linguistic phenomena (Losnegaard et al., 2016). Crowdsourcing MWE collections and annotations seems to provide a valuable alternative for the development of large-scale corpora of MWEs by leveraging the insights of the participants (Kato et al., 2018; Fort et al., 2018; Fort et al., 2020; Haagsma et al., 2020). RigorMortis (Fort et al., 2018; Fort et al., 2020) gamifies the MWEs collection and annotation processes in French corpora. In this context, gamification was deployed for MWE annotation with the aim of assessing the reliability of non-experts contributions in an MWEs annotation exercise at token level compared to a reference annotation. Taking into account the complexity in identifying types of MWEs (Schneider et al., 2014), (Fort et al., 2020) underline that the crowdsourcing approach can be considered as a valid choice to develop new annotated resources for MWEs identification.

Kato et al. (2018) combine automatic annotations and crowdsourcing for verbal MWEs including some idioms setting the annotation task as a multiword sense disambiguation problem. Participants from English-speaking countries were chosen on the basis of some requirements such as contributors with high accuracy on the CrowdFlower platform and contributors with a success rate higher than 70% in answering test questions. This procedure shows that annotators agree in approximately 67% on the same sense of verbal MWEs. Magpie (Haagsma et al., 2020) also showed the suitability of crowdsourcing for large-scale annotation of a variety of idiomatic expressions for English. Specifically, after picking up a set of idioms from three electronic dictionaries and extracting all forms of idiom types from the British National Corpus, they asked a crowd to annotate the data. The Magpie corpus was formed by aggregating annotations from participants along with a confidence score. This annotation procedure showed in Magpie confirmed that crowdsourcing is suitable, but both the participants and the procedure need to be carefully selected in order to gather reliable results. In this paper we evaluate and assess the quality of the data collected via the Dodiom gamification platform, a GWAP developed in the framework of Enet-Collect [1], and in particular we analyse if non-expert labelers can provide reliable natural language annotations specifically aimed at idiom corpora construction.

3. Dataset

In this section, we describe the idiomatic dataset collected through the Dodiom game. Specifically, we present the data collection process related to the submissions performed by players along with related features associated with Dodiom data. Then, we describe the data annotation process carried out to evaluate the quality of the data collected through the gamified crowdsourcing approach adopted in the Dodiom experiment: we present the guidelines adopted and the annotation process performed by the annotators.

3.1. Data Collection

We first provide a short outline of the Dodiom project ideation, participants, and data collection.[2] Subsequently, we describe the annotation process and the results concerning the quality of the collected data. Dodiom originates as a collaboration project between the NLP Research Group from the Department of Artificial Intelligence and Data Engineering of Istanbul University and the UNIOR NLP Research Group from the Department of Literary, Linguistic and Comparative Studies of the University of Naples *L'Orientale*.

The aim of the Dodiom project is the implementation of a gamified crowdsourcing approach for idiom corpora construction, where the crowd is actively taking a role in creating and annotating the language resource and rating annotations (Eryiğit et al., 2021). The Dodiom game has the major aim of collecting valuable usage samples for idioms which contain words that may also commonly be used in their literal meanings within a sentence, and for this reason make it difficult both for NLP systems and for language students to make sense of it. As an example, the idiom *gettare la spugna* (to throw the sponge) may have a literal reading and an idiomatic one, namely *to throw in the towel*, or *ammainare le vele* may mean *to furl the sails* or *to surrender*, depending on the context.

The game concerned two languages, namely Turkish

[1] European Network for Combining Language Learning with Crowdsourcing Techniques. EnetCollect had its main focus on combining the well-established domain of Language Learning with recent and successful crowdsourcing approaches. Official website available here: https://enetcollect.eurac.edu/.

[2] For a more detailed description of i) the gamified crowdsourcing approach used for collecting language learning materials for idiomatic expressions and ii) the design of the Dodiom messaging bot, an asynchronous multiplayer game for native speakers who compete with each other while providing idiomatic and nonidiomatic usage examples and rating other players' entries refer to (Eryiğit et al., 2021)

and Italian, even though it was designed with localization in mind, so as to collect idiom samples in multiple languages.

The Dodiom game, deployed between October and December 2020, aims at collecting idiomatic and non-idiomatic samples for specific idioms, which the players are required to submit when joining the game. Crowd-rating is also included in the game structure, as players are asked to express a positive or negative opinion upon other players' submissions ('likes' and 'dislikes' respectively). Improper use of the platform, as well as vulgar language, may also be reported by players (reports being later reviewed by moderators). For the Italian language, the overall Dodiom dataset includes a total amount of 6,730 samples, split into two sub-datasets: i) *with-reward* containing 5,286 samples, obtained during a session of the game where some monetary rewards were given to the best player of each day and ii) *without-reward* containing 1,444 sentences[3]. As shown in Figure 1, each provided example is displayed with the related idiom, the category (idiom/non-idiom) assigned by the player, the total number of likes/dislikes received from other players, any reports provided about vulgarity, improper usage of the platform etc., and the overall calculated rating (dislikes over likes).

3.2. Data Annotation

Since our goal is to evaluate the linguistic resources created and obtained through the use of the Dodiom game, we designed an annotation scheme aimed at estimating the reliability of the linguistic data collected. Namely, we devised a two-step annotation task in order to confirm whether the data provided by the players who joined the game are actually eligible and profitable for research and teaching purposes.

In the first step, we set up a list of guidelines along with 12 parameters which were employed to assess the sentences submitted by the players. Guidelines and parameters were set to aid the annotation exercise. Their purpose was to guarantee a consistent approach amongst annotators and to resolve ambiguous cases.

For eleven parameters, we selected a binary evaluation system for each of the categories: **0**, assigned as a default value to all examples provided by the crowd (that is, when they do not match the category under examination), whereas value **1** was assigned whenever the example was deemed to match the category. Instead, for a single parameter, **Quality**, we opted for using a scale associated with three values: 0, 1 and 2 which will be later further explained.

In order to have a clearer insight about what expert linguists evaluate in the collected crowdsourced samples, we hereby list the parameters taken into account.

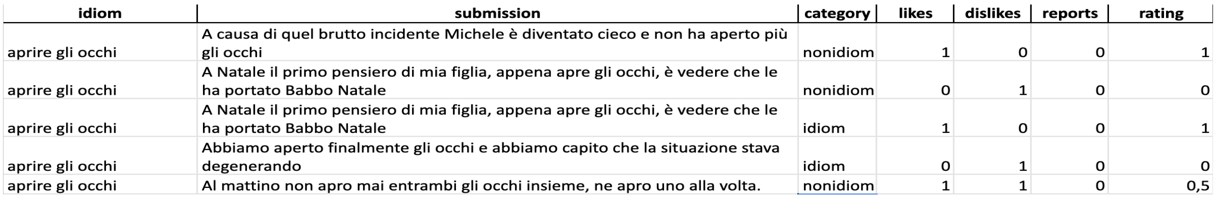
- **Wrong Category**: the sentence provided by the player is not correctly classified by the player, for instance it is classified by the user as having an idiomatic use while it is non-idiomatic.
- **Undecidable**: no sufficient context to decide if it is an idiomatic or non-idiomatic usage.
- **Low Context**: when no or only poor context has been provided.
- **Vulgar**: the proposed example is perceived as vulgar or offensive.
- **Incorrect Grammar - Word Order**: the grammar or the word order of the provided example is incorrect.
- **Incorrect Spelling**: the example contains spelling mistakes.
- **Meaningless**: the submitted sentence does not make sense to an average speaker.
- **Negative Sentiment**: the example arouses negative feelings.
- **Restricted Readers**: jokes related to the game itself. Some of the examples submitted were strictly related to the game itself and would therefore be unintelligible for an average reader. An instance for this was "A volte non so che esempi scrivere su dodiom, brancolo nel buio" (*Sometimes I don't know what examples to write on dodiom, I'm fumbling about in the dark*)
- **Not-Idiomatic & Not-Literal**: the players did not provide an example for the meaning required by the game but for one of the other possible meanings of the same idiom.
- **Quality** (0-2): *0 - No/poor quality*; *1 - Good quality* (good examples for machines and human learning purposes); *2 - Excellent quality* (very good examples which can be included in Dictionaries and Language learning resources).

The annotation scheme described above was applied to two random subsets of the Italian Dodiom dataset (10% out of the total collected samples). Specifically, 575 sentences from the *with reward* dataset and 154 from the *without-reward* dataset.

Subsequently, both subsets were labeled by three expert annotators with background in linguistics[4]: *A1*, *A2* and *A3*. They were tasked with labelling the same 729 sentences, each operating separately. In Figure 2 we provide some annotated sentences according to the twelve parameters and extracted from the *with reward* subset.

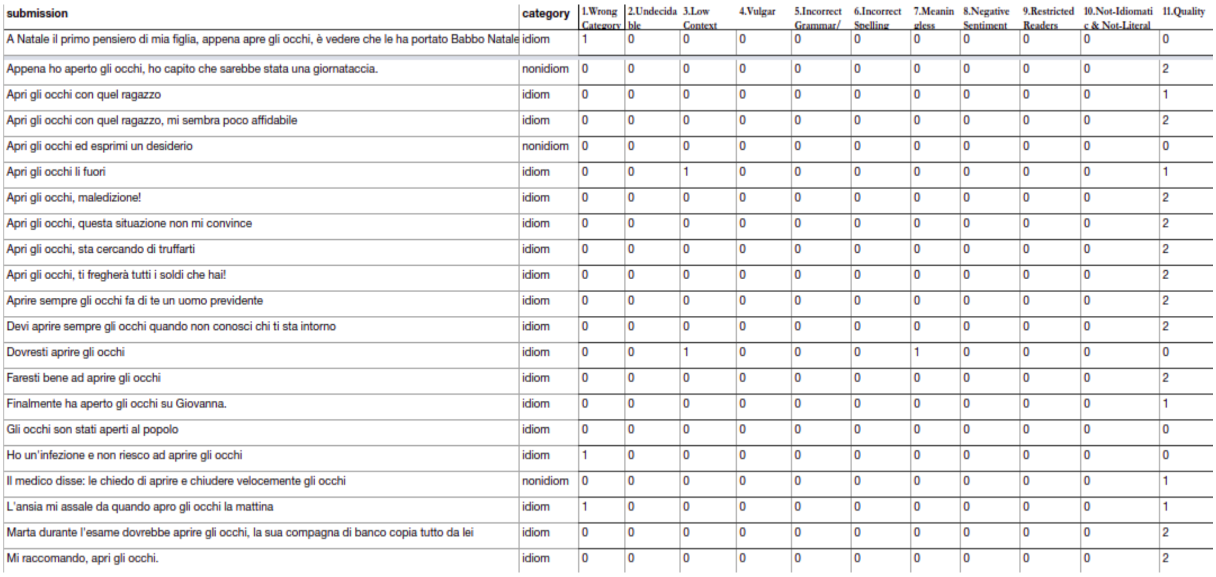
[3] The corpus is publicly available at: `https://github.com/unior-nlp-research-group/italian-dodiom-corpus`

[4] M.A. students in Linguistics from the Department of Literary, Linguistic and Comparative Studies at University of Naples "L'Orientale"



idiom	submission	category	likes	dislikes	reports	rating
aprire gli occhi	A causa di quel brutto incidente Michele è diventato cieco e non ha aperto più gli occhi	nonidiom	1	0	0	1
aprire gli occhi	A Natale il primo pensiero di mia figlia, appena apre gli occhi, è vedere che le ha portato Babbo Natale	nonidiom	0	1	0	0
aprire gli occhi	A Natale il primo pensiero di mia figlia, appena apre gli occhi, è vedere che le ha portato Babbo Natale	idiom	1	0	0	1
aprire gli occhi	Abbiamo aperto finalmente gli occhi e abbiamo capito che la situazione stava degenerando	idiom	0	1	0	0
aprire gli occhi	Al mattino non apro mai entrambi gli occhi insieme, ne apro uno alla volta.	nonidiom	1	1	0	0,5

Figure 1: Dodiom Italian dataset.



submission	category	1.Wrong Category	2.Undecidable	3.Low Context	4.Vulgar	5.Incorrect Grammar/	6.Incorrect Spelling	7.Meaningless	8.Negative Sentiment	9.Restricted Readers	10.Not-Idiomatic & Not-Literal	11.Quality
A Natale il primo pensiero di mia figlia, appena apre gli occhi, è vedere che le ha portato Babbo Natale	idiom	1	0	0	0	0	0	0	0	0	0	0
Appena ho aperto gli occhi, ho capito che sarebbe stata una giornataccia.	nonidiom	0	0	0	0	0	0	0	0	0	0	2
Apri gli occhi con quel ragazzo	idiom	0	0	0	0	0	0	0	0	0	0	1
Apri gli occhi con quel ragazzo, mi sembra poco affidabile	idiom	0	0	0	0	0	0	0	0	0	0	2
Apri gli occhi ed esprimi un desiderio	nonidiom	0	0	0	0	0	0	0	0	0	0	0
Apri gli occhi li fuori	idiom	0	0	1	0	0	0	0	0	0	0	1
Apri gli occhi, maledizione!	idiom	0	0	0	0	0	0	0	0	0	0	2
Apri gli occhi, questa situazione non mi convince	idiom	0	0	0	0	0	0	0	0	0	0	2
Apri gli occhi, sta cercando di truffarti	idiom	0	0	0	0	0	0	0	0	0	0	2
Apri gli occhi, ti fregherà tutti i soldi che hai!	idiom	0	0	0	0	0	0	0	0	0	0	2
Aprire sempre gli occhi fa di te un uomo previdente	idiom	0	0	0	0	0	0	0	0	0	0	2
Devi aprire sempre gli occhi quando non conosci chi ti sta intorno	idiom	0	0	0	0	0	0	0	0	0	0	2
Dovresti aprire gli occhi	idiom	0	0	1	0	0	0	1	0	0	0	0
Faresti bene ad aprire gli occhi	idiom	0	0	0	0	0	0	0	0	0	0	2
Finalmente ha aperto gli occhi su Giovanna.	idiom	0	0	0	0	0	0	0	0	0	0	1
Gli occhi son stati aperti al popolo	idiom	0	0	0	0	0	0	0	0	0	0	0
Ho un'infezione e non riesco ad aprire gli occhi	idiom	1	0	0	0	0	0	0	0	0	0	0
Il medico disse: le chiedo di aprire e chiudere velocemente gli occhi	nonidiom	0	0	0	0	0	0	0	0	0	0	1
L'ansia mi assale da quando apro gli occhi la mattina	idiom	1	0	0	0	0	0	0	0	0	0	1
Marta durante l'esame dovrebbe aprire gli occhi, la sua compagna di banco copia tutto da lei	idiom	0	0	0	0	0	0	0	0	0	0	2
Mi raccomando, apri gli occhi.	idiom	0	0	0	0	0	0	0	0	0	0	2

Figure 2: Example of Dodiom user sentences annotated following the twelve parameters.

4. Data Evaluation

In this section, we first describe the results obtained from the Inter-Annotator Agreement (IAA) between the three linguistic experts with respect to their judgments about the submissions of Dodiom players following the parameters described in section 3.2.

Subsequently, we will present the results obtained comparing the three expert linguists' judgments about quality and category on the common subsets of 729 sentences by the Dodiom players.

4.1. Inter-Annotator Agreement

Once the annotation process was completed, we compared the judgments expressed by the three expert linguists according to the twelve parameters chosen.

In the following section, we first show the results obtained by measuring the agreement among annotators using the Average Pairwise Percentage Agreement and Krippendorf's Alpha of the three annotators against the common *with-reward* subset according to the defined parameters, then we show also the results obtained for the common subset *without-reward*.

Concerning the IAA calculated on the subset *with-reward*, the agreement observed between the three linguistic experts and calculated with the average pairwise agreement was higher than 95% for the first eleven parameters, while a lower average was recorded for **Quality** with an average agreement of 53%. A plausible reason for the low agreement in the Quality criterion is related to the subjective nature encoded in the labels used to rate user submissions by experts. Furthermore, for the Quality parameter, the average of the agreement and the IAA shown in Table 1 do not take into account the ordinal nature of the labels. Consequently, an IAA was calculated that takes into account the closeness between the labels affixed by the annotators, i.e. the disagreement between 0 and 2 or 0 and 1 are penalized, while the disagreement between 1 and 2 is not penalized. We consider the agreement as the closeness between the two labels (1 and 2) that suggest good or excellent quality.

In addition, a lower reliability was recorded for **Not-Idiomatic** parameter with an alpha value of 0.48.

In Table 1, we show the results of the IAA calculated with Krippendorf's Alpha for all annotators and Pairwise Percentage Agreements for each pair of annotators.

Instead, concerning the IAA calculated on the subset *without-reward*, in Table 2 we show the results obtained using Pairwise Percentage Agreements for each pair of annotators.

In contrast, to measure the reliability of the three annotators' judgments we use Krippendorf's Alpha (Antoine et al., 2014; Zapf et al., 2016).

In this annotation exercise using *without-reward* subset, the reliability between annotators is high enough

	A1-A2	A1-A3	A2-A3	All
Wrong Category	0.982	1.0	0.982	0.988
Undecidable	1.0	0.998	0.998	0.998
Low Context	0.977	0.966	0.966	0.971
Vulgar	1	1	1	1.0
Incorrect Grammar	0.984	0.9846	1	0.993
Incorrect Spelling	0.997	0.995	0.996	0.995
Meaningless	0.981	0.987	1	0.98
Negative Sentiment	1	1	1	1
Restricted Readers	0.996	0.996	1	0.995
Not-Idiomatic	0.976	0.972	0.098	0.48
Quality	0.34	0.32	0.93	0.30
Quality (Closeness Agreement)	0.41	0.39	0.95	-

Table 1: IAA between Experts using with-reward subset.

for most parameters. Unlike the previous exercise, here we reach an alpha value of 0.97 for **Not-Idiomatic** while for the **Quality** parameter alpha is equal to 0.38. Also in the *without-reward* subset, the agreement for the Quality parameter reflects the subjective nature of the expert judgment.

	A1-A2	A1-A3	A2-A3	All
Wrong Category	0.987	0.987	1	0.98
Undecidable	1	1	1	1.0
Low Context	0.974	0.974	1	0.96
Vulgar	0.985	0.985	1	0.97
Incorrect Grammar	0.993	0.993	1	0.99
Incorrect Spelling	1	1	1	1.0
Meaningless	1	1	1	1.0
Negative Sentiment	1	1	1	1.0
Restricted Readers	0.941	0.941	0.987	0.91
Not-Idiomatic	0.985	0.985	1	0.97
Quality	0.60	0.50	0.65	0.38
Quality (Closeness Agreement)	0.72	0.74	0.56	-

Table 2: IAA between Experts using without-reward subset.

4.2. Linguistic Experts and the Crowd

In this section, we provide an insight on the results obtained by comparing the judgments made by Dodiom users regarding Idiom/Not-Idiom and Quality with the judgements made by the three linguistic experts.

Regarding the *with-reward* subset, linguistic experts annotated on average 14% of sentences as being incorrectly assigned by users to the idiomatic category (336 sentences). In contrast, they annotated on average 1.3% of sentences marked by users as being incorrectly assigned by users to the non-idiomatic category (239 sentences). In Table 3, we provide the results regarding the corrections on the Category. With category corrections, we are referring to the number of times that expert linguists labeled the sentences of Dodiom users as 1 in **Wrong Category** parameter.

While in Table 4, we show the number of corrections made on the subset *without-reward*.

Tables 3 and 4 show that the interventions of the linguistic experts in correcting the category of idioms were very few. This shows that the crowd is able to create and use sentences containing idioms, as well as to recognize non-idiomatic uses. In addition, the judgment left by the crowd about the category of the sentence submitted or proposed by the Dodiom game are not far from the judgment of an expert linguist.

	A1	A2	A3
Idiom	14	15	15
Not-Idiom	2	1	1

Table 3: Average corrections in the *with-reward* subset made by three expert linguists.

	A1	A2	A3
Idiom	0	1	1
Not-Idiom	0	1	1

Table 4: Average corrections in *without-reward* subset made by three expert linguists.

In Table 5, we show the results regarding the **Quality** judgments about idiomatic sentences by linguistic experts compared with those by Dodiom users. The results refer to the complete subset extracted from the Dodiom corpus. In fact, we show both the quality judgments about the *with-reward* subset and the *without-reward* subset.

Quality Evaluation	Expert	Crowd
2 - Excellent quality	278	315
1 - Good quality	177	195
0 - No/poor quality	274	219

Table 5: Dodiom Quality Evaluation - Comparison Expert vs. Crowd.

5. Conclusions

We demonstrate the effectiveness of gamification for the collection of valuable data related to idiom corpora construction. The guidelines we proposed to annotate the crowdsourced corpus proved that players have provided valuable results, which confirms the great advantage of deploying such strategies for the acquisition of precious language resources which, otherwise, could have taken much longer and would have been extremely expensive. Given these results, we decided to run the game again and have the crowd annotate a new idiom corpus focusing on idioms that turn out to be ambiguous, i.e. that have both literal and idiomatic readings, extracted from the online Italian dictionary[5] compiled under the direction of linguist Tullio De Mauro.

6. Acknowledgments

This research has been partly carried out within the framework of an Innovative Industrial PhD project sup-

[5] `https://dizionario.internazionale.it/`

ported by the POR Campania FSE2014/2020 fund. Our thanks go also to the Italian annotators Adriana Capasso and Giovanna Carandente, for their precious contribution. Autorship contribution is as follows: Giuseppina Morza is author of Sections 1 and 2; Johanna Monti is author of Section 3; Raffaele Manna of Section 4. Abstract and conclusions are in common.

7. Bibliographical References

Antoine, J.-Y., Villaneau, J., and Lefeuvre, A. (2014). Weighted krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *EACL 2014*, pages 10–p.

Chamberlain, J., Poesio, M., Kruschwitz, U., et al. (2008). Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, pages 42–49.

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 *The People's Web Meets NLP*, pages 3–44. Springer.

Chklovski, T. (2005). 1001 paraphrases: Incenting responsible contributions in collecting paraphrases from volunteers. In *AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors*, pages 16–20.

Paul Cook, et al., editors. (2021). *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, Online, August. Association for Computational Linguistics.

Dumitrache, A., Aroyo, L., Welty, C., Sips, R.-J., and Levas, A. (2013). "dr. detective": combining gamification techniques and crowdsourcing to create a gold standard in medical text. In *Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web (CrowdSem 2013), 12th International Semantic Web Conference*. Citeseer.

Eryiğit, G., Şentaș, A., and Monti, J. (2021). Gamified crowdsourcing for idiom corpora construction. *arXiv preprint*, arXiv:2102.00881.

Fort, K., Guillaume, B., and Stern, V. (2014). ZOMBILINGO: eating heads to perform dependency syntax annotation (ZOMBILINGO : manger des têtes pour annoter en syntaxe de dépendances) [in French]. In *Proceedings of TALN 2014 (Volume 3: System Demonstrations)*, pages 15–16, Marseille, France, July. Association pour le Traitement Automatique des Langues.

Fort, K., Guillaume, B., Constant, M., Lefebvre, 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.

Fort, K., Guillaume, B., Pilatte, Y.-A., Constant, M., and Lefèbvre, N. (2020). Rigor mortis: Annotating mwes with a gamified platform. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4395–4401.

Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *International Conference on Computational Linguistics (COLING)*.

Haagsma, H., Bos, J., and Nissim, M. (2020). MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 279–287, Marseille, France, May. European Language Resources Association.

Kato, A., Shindo, H., and Matsumoto, Y. (2018). Construction of large-scale english verbal multiword expression annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Losnegaard, G. S., Sangati, F., Escartín, C. P., Savary, A., Bargmann, S., and Monti, J. (2016). PARSEME survey on MWE resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2299–2306, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Stella Markantonatou, et al., editors. (2020). *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, online, December. Association for Computational Linguistics.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., Krstev, C., Vincze, V., Wójtowicz, B., Losnegaard, G. S., et al. (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)*.

Savary, A., Ramisch, C., Cordeiro, S. R., Sangati, F., Vincze, V., Qasemi Zadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., et al. (2017). The parseme shared task on automatic identification of verbal multiword expressions. In *The 13th Workshop on Multiword Expression at EACL*, pages 31–47.

Agata Savary, et al., editors. (2018). *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., and A Smith, N. (2014). Comprehensive annotation of multiword expressions in a social web corpus.

Snow, R., O'connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Zapf, A., Castell, S., Morawietz, L., and Karch, A. (2016). Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC medical research methodology*, 16(1):1–10.