Natural Language Engineering (2022), 1–33
doi:10.1017/S1351324921000401

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Gamified crowdsourcing for idiom corpora construction

GülŞen Eryiğit[1,2*], Ali Şentaş[1] and Johanna Monti[3]

[1]Faculty of Computer and Informatics, Istanbul Technical University, Istanbul, Turkey, [2]Department of Artificial Intelligence and Data Engineering, Istanbul Technical University, Istanbul, Turkey, and [3]Department of Literary, Linguistic and Comparative Studies, University of Naples L'Orientale, Naples, Italy
*Corresponding author. E-mail: gulsen.cebiroglu@itu.edu.tr

(Received 15 January 2021; revised 8 November 2021; accepted 15 November 2021)

Abstract

Learning idiomatic expressions is seen as one of the most challenging stages in second-language learning because of their unpredictable meaning. A similar situation holds for their identification within natural language processing applications such as machine translation and parsing. The lack of high-quality usage samples exacerbates this challenge not only for humans but also for artificial intelligence systems. This article introduces a gamified crowdsourcing approach for collecting language learning materials for idiomatic expressions; a messaging bot is designed as an asynchronous multiplayer game for native speakers who compete with each other while providing idiomatic and nonidiomatic usage examples and rating other players' entries. As opposed to classical crowd-processing annotation efforts in the field, for the first time in the literature, a crowd-creating & crowd-rating approach is implemented and tested for idiom corpora construction. The approach is language-independent and evaluated on two languages in comparison to traditional data preparation techniques in the field. The reaction of the crowd is monitored under different motivational means (namely, gamification affordances and monetary rewards). The results reveal that the proposed approach is powerful in collecting the targeted materials, and although being an explicit crowdsourcing approach, it is found entertaining and useful by the crowd. The approach has been shown to have the potential to speed up the construction of idiom corpora for different natural languages to be used as second-language learning material, training data for supervised idiom identification systems, or samples for lexicographic studies.

Keywords: Crowdsourcing; Gamification; Game with a purpose (GWAP); Idiomatic expressions; Language resources

1. Introduction

An idiom is usually defined as a group of words established by usage as having a idiosyncratic meaning not deducible from those of the individual words forming that idiom. It is possible to come across situations where components (i.e., words) of an idiom appear in a sentence without forming an idiomatic expression. This ambiguous situation poses a significant challenge for both foreign language learners and artificial intelligence (AI) systems since it requires a deep semantic understanding of the language.

Idiomatic control has been seen as a measure of proficiency in a language both for humans and AI systems. The task is usually referred to as idiom identification or idiom recognition in natural language processing (NLP) studies and is defined as understanding/classifying the idiomatic (i.e., figurative) or nonidiomatic usage of a group of words (i.e., either with the literal meaning arising from their cooccurrence or by their separate usage). Two such usage examples containing different surface forms of the lemmas {"hold," "one's," and "tongue"} are provided below:

"Out of sheer curiosity I *held my tongue*, and waited." (idiomatic) (meaning "stop talking")

"One of the things that they teach in first aid classes is that a victim having a seizure can swallow his tongue, and you should *hold his tongue* down." (nonidiomatic)

Learning idiomatic expressions is seen as one of the most challenging stages in second-language learning because of their unpredictable meaning. Several studies have discussed efficient ways of teaching idioms to second-language (L2) learners (Vasiljevic 2015; Siyanova-Chanturia 2017), and obviously, both computers and humans need high-quality usage samples exemplifying the idiom usage scenarios and patterns. When do some words occurring within the same sentence form a special meaning together? Can the components of an idiom undergo different morphological inflections? If so, is it possible to inflect them in any way or do they have particular limitations? May other words intervene between the components of an idiom? If so, could these be of any word type or are there any limitations? Although it may be possible to deduct some rules (see Appendix for an example) defining some specific idioms, unfortunately, creating a knowledge base that provides such detailed definitions or enough samples to deduct answers to these questions is a very labor-intensive and expensive process, which could only be conducted by native speakers. Yet, these knowledge bases are crucial for foreign language learners who do not have as much time and encounter as many examples as native language learners to implicitly acquire idiom structures in the target language. Unfortunately, traditional dictionaries usually do not provide all the information and in-context examples needed to correctly interpret and use idioms by L2 learners (Moon 2015).

Due to the mentioned difficulties, there exist very few studies that introduce an idiom corpus (providing idiomatic and nonidiomatic examples), and these are available only for a couple of languages and a limited number of idioms: Birke and Sarkar (2006) and Cook, Fazly, and Stevenson (2008) for 25 and 53 English idioms, respectively, and Hashimoto and Kawahara (2009) for 146 Japanese idioms. Similarly, high coverage idiom lexicons either do not exist for every language or contain only a couple of idiomatic usage samples, which is insufficient to answer the above questions. Examples of use were considered as must have features of an idiom dictionary app in Caruso *et al*. (2019) that tested a dictionary mockup for the Italian language with Chinese students. On the other hand, it may be seen that foreign language learning communities are trying to fill this resource gap by creating/joining online groups or forums to share idiom examples.[1] Obviously, the necessity for idiom corpora applies to all natural languages and we need an innovative mechanism to speed up the creation process of such corpora by ensuring the generally accepted quality standards in language resource creation.

Gamified crowdsourcing is a rapidly increasing trend, and researchers explore creative methods of use in different domains (Morschheuser *et al*. 2017; Morschheuser and Hamari 2019; Murillo-Zamorano, ángel López Sánchez, and Bueno Muñoz 2020). The use of gamified crowdsourcing for idiom corpora construction has the potential to provide solutions to the above-mentioned problems, as well as to the unbalanced distributions of idiomatic and nonidiomatic samples, and the data scarcity problem encountered in traditional methods. This article proposes 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. The approach is experimented on two languages and evaluated in comparison to traditional data preparation techniques in the field. The selected languages are Turkish and Italian, which come from different language families with different syntactic and morphological patterns and make a good pair to show that the approach works for typologically different languages. The results reveal that the approach is powerful in collecting the targeted materials, and although being an explicit crowdsourcing approach,

[1]Some examples include https://t.me/Idiomsland for English idioms with 65K subscribers, https://t.me/Deutschpersich for German idioms with 3.4K subscribers, https://t.me/deyimler for Turkish Idioms with 2.7K subscribers, https://t.me/Learn_Idioms for French idioms with 2.5 subscribers. The last three are messaging groups providing idiom examples and their translations in Arabic.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

it is found entertaining and useful by the crowd. The approach has been shown to have the potential to speed up the construction of idiom corpora for different natural languages, to be used as second-language learning material, training data for supervised idiom identification systems, or samples for lexicographic studies.

The article is structured as follows: Section 2 provides a background and the related work, Section 3 describes the game design, Section 4 provides analyses, and Section 5 gives the conclusion.

2. Background and related work

Several studies investigate idioms from a cognitive science perspective: Kaschak and Saffran (2006) constructed artificial grammars that contained idiomatic and "core" (nonidiomatic) grammatical rules and examined learners' ability to learn the rules from the two types of constructions. The findings suggested that learning was impaired by idiomaticity, counter to the conclusion of Sprenger, Levelt, and Kempen (2006) that structural generalizations from idioms and nonidioms are similar in strength. Konopka and Bock (2009) investigate idiomatic and nonidiomatic English phrasal verbs and states that despite differences in idiomaticity and structural flexibility, both types of phrasal verbs induced structural generalizations and differed little in their ability to do so.

We may examine the traditional approaches which focus on idiom annotation in two main parts: first, the studies focusing solely on idiom corpus construction and second the studies on general multiword expressions' (MWEs) annotations also including idioms. Both approaches have their own drawbacks, and exploration of different data curation strategies in this area is crucial for any natural language, but especially for morphologically rich and low resource languages (MRLs and LRLs).

The studies focusing solely on idiom corpus construction (Birke and Sarkar 2006; Cook *et al.* 2008; Hashimoto and Kawahara 2009) first retrieve sentences from a text source according to some keywords from the target group of words (i.e., target idiom's constituents) and then annotate them as idiomatic or nonidiomatic samples. The retrieval process is not as straightforward as one might think, since the keywords should cover all possible inflected forms of the words in focus (e.g., keyword "go" could not retrieve its inflected form "went"), especially for MRLs where words may appear under hundreds of different surface forms. The solution to this may be lemmatization of the corpus and searching with lemmas, but this will not work in cases where the data source is pre-indexed and only available via a search engine interface such as the internet. This first approach may also lead to unexpected results on the class distributions. For example, Hashimoto and Kawahara (2009) states that examples were annotated for each idiom, regardless of the proportion of idioms and literal phrases, until the total number of examples for each idiom reached 1000, which is sometimes not reachable due to data unavailability.

Idioms are seen as a subcategory[2] of MWEs which have been subject to many initiatives in recent years such as Parseme EU COST Action, MWE-LEX workshop series, and ACL special interest group SIGLEX-MWE. Traditional methods for creating MWE corpora (Vincze, Nagy T., and Berend 2011; Schneider *et al.* 2014; Losnegaard *et al.* 2016; Savary *et al.* 2018) generally rely on manually annotating MWEs on previously collected text corpora (news articles most of the time and sometimes books), this time without being retrieved with any specific keywords. However, the scarcity of MWEs (especially idioms) in text has presented obstacles to corpus-based studies and NLP systems addressing these (Schneider *et al.* 2014). In this approach, only idiomatic examples are annotated. One may think that all the remaining sentences containing idiom's components are nonidiomatic samples. However, in this approach, human annotators are

[2]In this article, differing from Constant *et al.* (2017), which list subcategories of MWEs, we use the term "idiom" for all types of MWEs carrying an idiomatic meaning including phrasal verbs in some languages including English (e.g., "throw up").

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

prone to overlook, especially those MWE components that are not juxtaposed within a sentence. Bontcheva, Derczynski, and Roberts (2017) state that annotating one named entity (another sub-category of MWEs) type at a time as a crowdsourcing task is a better approach than trying to annotate all entity types at the same time. Similar to Bontcheva *et al.* (2017), our approach achieves the goal of collecting quality and creative samples by focusing the crowd's attention on a single idiom at a time. Crowdsourcing MWE annotations has been rarely studied (Kato, Shindo, and Matsumoto 2018; Fort *et al.* 2018; Fort *et al.* 2020) and these were crowd-processing[3] efforts.

Crowdsourcing (Howe 2006) is a technique used in many linguistic data collection tasks (Mitrović 2013). Crowdsourcing systems are categorized under four main categories: crowd-processing, crowd-solving, crowd-rating, and crowd-creating (Geiger and Schader 2014; Prpić et al. 2015; Morschheuser *et al.* 2017). While "crowd-creating solutions seek to create comprehensive (emergent) artifacts based on a variety of heterogeneous contributions," "crowd-rating systems commonly seek to harness the so-called wisdom of crowds to perform collective assessments or predictions" (Morschheuser *et al.* 2017). The use of these two later types of crowdsourcing together has a high potential to provide solutions to the above-mentioned problems for idiom corpora construction.

One platform researchers often use for crowdsourcing tasks is *Amazon Mechanical Turk (MTurk)*.[4] Snow *et al.* (2008) used it for linguistics tasks such as word similarity, textual entailment, temporal ordering, and word-sense disambiguation. Lawson *et al.* (2010) used MTurk to build an annotated NER corpus from emails. Akkaya *et al.* (2010) used the platform to gather word-sense disambiguation data. The platform proved especially cost-efficient in the highly human labor-intensive task of word-sense disambiguation (Akkaya *et al.* 2010; Rumshisky *et al.* 2012). Growing popularity also came with criticism for the platform as well (Fort, Adda, and Cohen 2011).

MTurk platform uses monetary compensation as an incentive to complete the tasks. Another way of utilizing the crowd for microtasks is gamification, which, as an alternative to monetary compensation utilizes game elements such as points, achievements, and leaderboards. von Ahn (2006) pioneered these types of systems and called them games with a purpose (GWAP) (von Ahn 2006). ESPGame (von Ahn and Dabbish 2004) can be considered as one of the first GWAPs. It is designed as a game where users were labeling images from the web while playing a Taboo™ like game against each other. The authors later developed another GWAP, Verbosity (von Ahn, Kedia, and Blum 2006), this time for collecting common-sense facts in a similar game setting. GWAPs are popularized in the NLP field by early initiatives such as *1001 Paraphrases* (Chklovski 2005), *Phrase Detectives* (Chamberlain, Poesio, and Kruschwitz 2008), *JeuxDeMots* (Artignan *et al.* 2009), and *Dr. Detective* (Dumitrache *et al.* 2013). *RigorMortis* (Fort *et al.* 2018; Fort *et al.* 2020) gamifies the traditional MWE annotation process described above (i.e., crowd-processing).

Gamified crowdsourcing is a rapidly increasing trend, and researchers explore creative methods of use in different domains (Morschheuser *et al.* 2017; Morschheuser and Hamari 2019; Murillo-Zamorano *et al.* 2020). Morschheuser *et al.* (2017) introduce a conceptual framework of gamified crowdsourcing systems according to which the motivation of the crowd may be provided by either gamification affordances (such as leaderboards, points, and badges) or additional incentives (such as monetary rewards, prizes). In our study, we examine both of these motivation channels and report their impact. According to Morschheuser *et al.* (2017), "one major challenge in motivating people to participate is to design a crowdsourcing system that promotes and enables the formation of positive motivations toward crowdsourcing work and fits the type of the activity." Our approach to gamified crowdsourcing for idiom corpora construction relies on crowd-creating

[3]"Crowd-processing approaches rely on the crowd to perform large quantities of homogeneous tasks. Identical contributions are a quality attribute of the work's validity. The value is derived directly from each isolated contribution (non-emergent)" (Morschheuser *et al.* 2017).

[4]https://www.mturk.com.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

and crowd-rating. We both value the creativity and systematic contributions of the crowd. As explained above, since it is not easy to retrieve samples from available resources, we expect our users to be creative in providing high-quality samples.

Morschheuser *et al*. (2018) state that users increasingly expect the software to be not only useful but also enjoyable to use, and a gamified software requires an in-depth understanding of motivational psychology and requires multidisciplinary knowledge. In our case, these multidisciplines include language education, computational linguistics, NLP, and gamification. To shed light to a successful design of gamified software, the above-mentioned study divides the engineering of gamified software into 7 main phases and mentions 13 design principles (from now on depicted as DP#n, where n holds for the design principle number) adopted by experts. These seven main phases are project preparation, analysis, ideation, design, implementation, evaluation, and monitoring phases. The following sections provide the details of our gamification approach by relating the stages to these main design phases and principles. For the sake of space, we define only the crucial phases as separate sections and refer to the others between lines. The complete list of design principles is given in Table A1 in the Appendix.

3. Game design

The aim while designing the software was to create an enjoyable and cooperative environment that would motivate the volunteers to help the research studies. The game is designed to collect usage samples for idioms of which the words of the idiom may also commonly be used in their literal meanings within a sentence. An iterative design process has been adopted. After the first ideation, design, and prototype implementation phases, the prototype was shared with the stakeholders (see Acknowledgments) (as stated in DP#7) and the design has been improved accordingly.

A messaging bot (named "Dodiom"[5]) is designed as an asynchronous multiplayer game[6] for native speakers who compete with each other while providing idiomatic and nonidiomatic usage examples and rating other players' entries. The game is an explicit crowdsourcing game and players are informed from the very beginning that they are helping to create a public data source by playing this game.[7]

The story of the game is based on a bird named Dodo (the persona of the bot) trying to learn a foreign language and having difficulty learning idioms in that language. Players try to teach Dodo the idioms in that language by providing examples. Dodiom has been developed as an open-source project (available on Github[8]) with the focus on being easily adapted to different languages. All the interaction messages are localized and shown to the users in the related language; localizations are currently available for English, Italian, and Turkish languages.

3.1 Main interactions and gameplay

Dodo learns a new idiom every day. The idioms to be played each day are selected by moderators according to their tendency to be used also with their literal meaning. For each idiom, players have a predetermined time frame to submit their samples and reviews, so they can play at their own pace. Since the bot may send notifications via the messaging platform in use, the time frame is determined as between 11 a.m. and 11 p.m.[9]

[5] A language-agnostic name has been given for the game to be generalized to every language.

[6] Asynchronous multiplayer games enable players to take their turns at a time that suits them; that is, the users do not need to be in the game simultaneously.

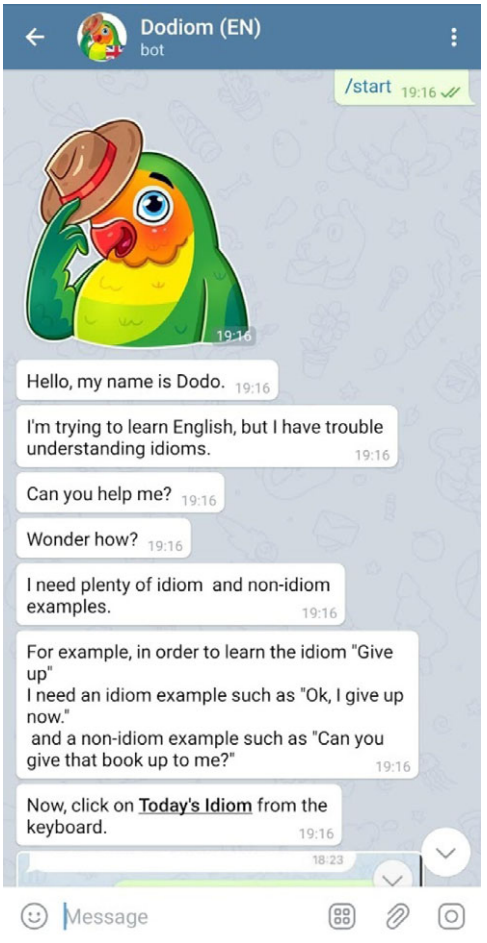
[7] In addition to the reporting and banning mechanism (DP#10), we also shared a consent agreement message in the welcome screen and the announcements in line with DP#12 related to legal and ethical constraints.

[8] https://github.com/Dodiom/dodiom.

[9] Different time frames have been tried in the iterative development cycle (DP#4) and it has been decided that this time frame is the most suitable.

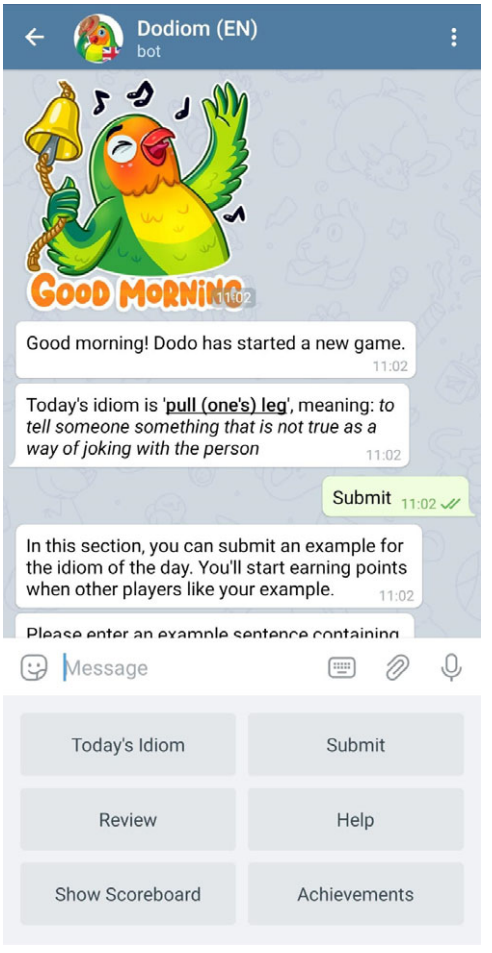
https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

(a)



Dodo greeting the player, describing the game, and showing the next steps

(b)



Main Menu showing the currently available options

Figure 1. Dodiom welcome and menu screens.

When the users connect to the bot for the first time, they are greeted with Dodo explaining to them what the game is about and teaching them how to play in a step-by-step manner (Figure 1a). This pre-game tutorial and the simplicity of the game proved useful as most of the players were able to play the game in a matter of minutes and provided high-quality examples. Players are then presented with the idiom of the day together with its intended idiomatic meaning (see Figure 1). All the game messages are studied very carefully to achieve this goal and ensure that the crowd unfamiliar with AI or linguistics understands the task easily. Random tips for the game are also shared with the players right after they submit their examples. This approach is similar to video games where tips about the game are shown to players on loading screens and/or menus.

Figure 1b shows the main menu, from where the players can access various modes of the game.

"Today's idiom" tells the player what that day's chosen idiom is, players can then submit usage examples for said idioms to get more points.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

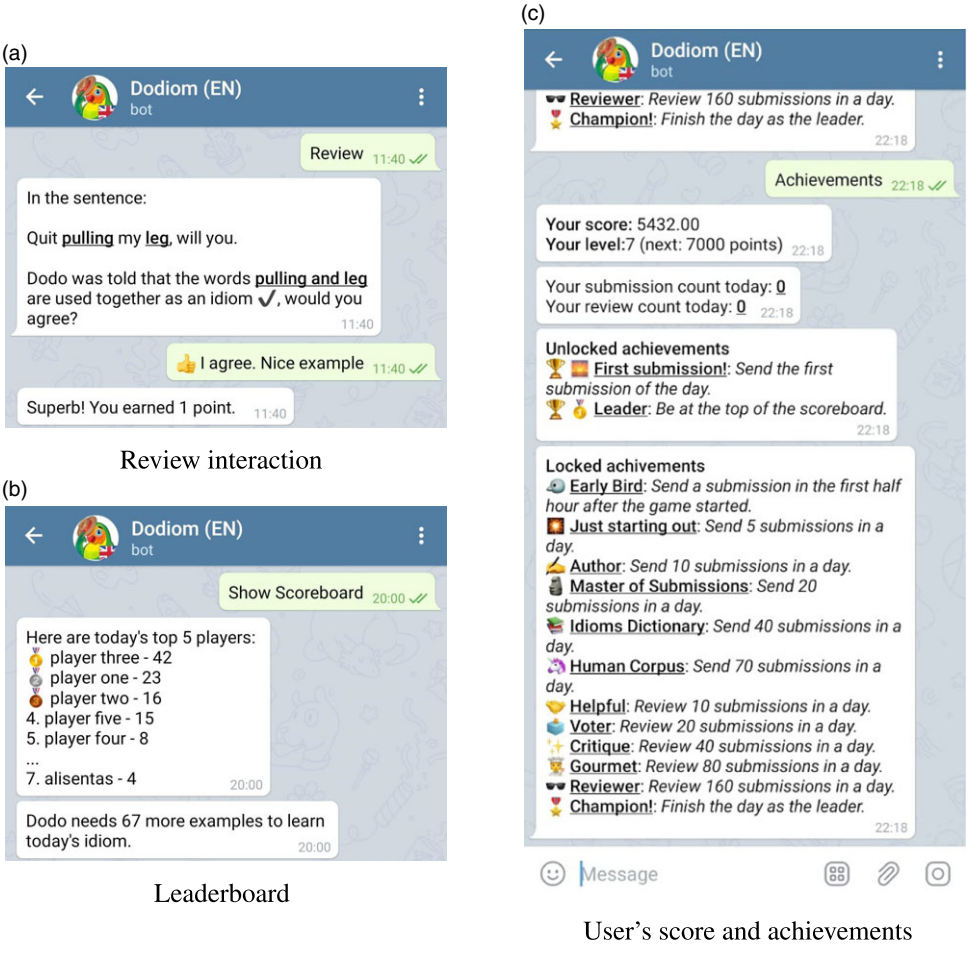


Figure 2. Some interaction screens.

"Submit" allows players to submit new usage examples. When clicked, Dodo asks the player to input the example sentence and when the player sends one, the sentence is checked if it contains the words (i.e., the lemmas of the words) that appear in the idiom. If so, Dodo then asks whether these words form an idiom in the given sentence or not. The players are awarded each time other players like their examples, so they are incentivized to enter multiple high-quality submissions.

"Review" allows players to review submissions sent by other players. Dodo shows players examples of other players one at a time together with their annotations (i.e., idiom or not) and asks its approval. Users are awarded points for each submission they review, so they are also incentivized to review. The exact scoring and incentivization system will be explained in Section 3.2. Figure 2a shows a simple interaction between Dodo and a user, where Dodo asks whether or not the words *pulling* and *leg* (automatically underlined by the system) in the sentence *"Quit pulling my leg, will you"* are used idiomatically. The user responds with acknowledgment or dislike by clicking on the corresponding button and then Dodo thanks the user for his/her contribution. Users can also report the examples which do not fit the general guidelines (e.g., vulgar language, improper usage of the platform) for the submissions to be later reviewed by moderators. The moderators can flag the submissions and ban the users from the game depending on the submission. Submissions with

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

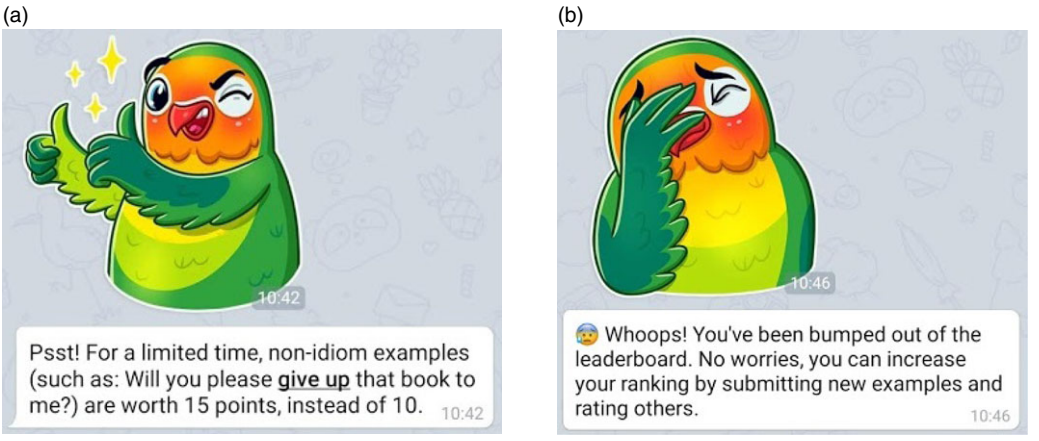


Figure 3. Notification samples.

fewer reviews are shown to the players first (i.e., before the samples that were reviewed previously) so that each submission can receive approximately the same number of reviews.

"Help" shows the help message, which is a more compact version of the pre-game tutorial.

"Show scoreboard" displays the current state of the leaderboard which is updated every time a player is awarded any points. As seen in Figure 2b, the scoreboard displays the top five players' and the current player's scores. The scoreboard is reset every day for each idiom. Additionally, 100 submissions are set as a soft target for the crowd and a message stating the number of submissions remaining to reach this goal is shown below the scoreboard. The message no longer appears when the target is reached.

"Achievements" option shows the score, level, and locked/unlocked achievements of the user. An example can be seen in Figure 2c where many achievements are designed to gamify the process and reward players for specific actions such as *Early Bird* achievement for early submissions and *Author* for sending 10 submissions in a given day. Whenever an achievement is obtained, the user is notified with a message and an exciting figure (similar to the ones in Figure 3).

3.2 Gamification affordances and additional incentives

Dodiom uses both gamification affordances and additional incentives (Morschheuser *et al*. 2017) for the motivation of its crowd. Before the decision of the final design, we have tested with several scoring systems with and without additional incentives. This section provides the detailed form of the final scoring system together with previous attempts, gamification affordances, and additional incentives.

The philosophy of the game is based on collecting valuable samples that illustrate the different ways of use and make it possible to make inferences that define how to use a specific idiom (such as the ones in the Appendix). The samples to be collected are categorized into four main types given below and referred to by the combination of the following abbreviations: Id (idiomatic), NonId (nonidiomatic), Adj (adjacent), and Sep (separated). For the sake of game simplicity, this categorization is not explicitly described to the users but is only used for background evaluations.

- Id/Adj samples: Idiomatic samples in which the constituent words are used side by side (juxtaposed) (e.g., "Please hold your tongue and wait.");

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

- Id/Sep samples: Idiomatic samples in which the constituent words are separated by some other words, which is a more common phenomenon in free word order languages[10] (e.g., "Please hold your breath and tongue and wait for the exciting announcements.");
- NonId/Adj samples: Nonidiomatic samples in which the constituent words are used side by side (e.g., "Use sterile tongue depressor to hold patient's tongue down.");
- NonId/Sep samples: Nonidiomatic samples in which the constituent words are separated by some other words (e.g., "Hold on to your mother tongue.").

Producing samples from some categories (e.g., Id/Seps and NonId/Adjs) may not be the most natural form of behavior. For Turkish, we experimented with different scorings that will motivate our users to produce samples from different categories. Before settling on the final form of the scoring system, two other systems have been experimented within the preliminary tests. These include having a fixed set of scores for each type (i.e., 30, 40, 20, and 10 for Id/adj, Id/Sep, NonId/Adj, and NonId/Sep, respectively). This scoring system caused players to enter submissions for only the Id/Sep to get the most out of a submission and resulted in very few other types of samples. To fix the major imbalance problem, in another trial, a decay parameter has been added to lower the initial type scores whenever a new submission arrives. Unfortunately, this new system had little to no effect on remedying the imbalance and made the game harder to understand for players who could not easily figure out the scoring system.[11] This latter strategy was also expected to incentivize players to enter submissions early in the game, but it did not work out as planned.

Although being one of the most challenging types for language learners and an important type that we want to collect samples from, Id/Sep utterances may not be as common as Id/Adj utterances for some idioms and be rare for some languages with fixed word order. Similarly producing NonId/Adj samples may be difficult for some idioms and overdirecting the crowd to produce more examples of this type can result in unnatural sentences. Thus, game motivations should be chosen carefully.

We used scoring, notifications, and tips to increase the type variety in the dataset in a meaningful and natural way. The final scoring system used during the evaluations (presented in the next sections) is as follows: each review is worth one point unless it is done in the happy hour during which all reviews are worth two points. As stated above, after each submission a random tip is shown to the submitter motivating him/her to either review other's entries or to submit samples from either Id/Sep or NonId/Adj. The scores for each type are set to 10 with the only difference of Id/Sep being set to 12. The system periodically checks the difference between Id/Adj and NonId/Adj samples and when this exceeds 15 samples, it increases the scores of the idiomatic or nonidiomatic classes.[12] The score increase is notified to the crowd via a message (Figure 3a stating either Dodo needs more idiomatic samples or nonidiomatic samples) and remains active until the difference falls below five samples. As stated above, although for some idioms producing NonId/Adj samples may be difficult, since the notification message is for calling nonidiomatic samples in general, the crowd is expected to provide both NonId/Adj and NonId/Sep samples in accordance with the natural balance.

Push notifications are also used to increase player engagement. There are several notifications sent through the game, which are listed below. The messages are arranged so that an inactive user

[10] In free word order languages, the syntactic information is mostly carried at word level due to affixes, thus the words may freely change their position within the sentence without affecting the meaning.

[11] User feedbacks are taken via personal communication on trial runs.

[12] That is to say, when #Id/Adj samples $\geqslant$ #NonId/Adj samples + 15, the scores of NonId/Adj and NonId/Sep samples are increased by 5, and similarly when #NonId/Adj samples $\geqslant$ #Id/Adj samples + 15, the scores of Id/Adj and Id/Sep samples are increased by 5.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

would only receive a couple of notifications from the game each day; the first three items below are sent to every user, whereas the last three are sent only to active users of that day.

(1) Every morning Dodo sends a good morning message when the game starts and tells the player that day's idiom.
(2) When a category score is changed, a notification is sent to all players (Figure 3a).
(3) A notification is sent to players when review happy hour is started. This event is triggered manually by moderators, and for 1 hour, points for reviews worth double. This notification also helps to reactivate low-speed play.
(4) When a player's submission is liked by other players, the author of the submission is notified and encouraged to check back the scoreboard. Only one message of this type is sent within a limited time to avoid causing too many messages consecutively.
(5) When a player becomes the leader of the scoreboard or enters the top five he/she is congratulated.
(6) When a player loses his/her top position on the leaderboard or loses his/her place in the top three or five he/she is notified about it and encouraged to get back and send more submissions to take his/her place back. (Figure 3b)

We've seen that player engagement increased dramatically when these types of notifications were added (this will be detailed in Section 4.4). As additional incentives, we also tested with some monetary rewards given to the best player of each day and investigated the impacts, a 5 Euro online store gift card for Italian and a 25 Turkish Lira online bookstore gift card for Turkish. These monetary rewards, which are mainly connected with intellectual rewards, that is, gift cards to buy books, do not raise in our view any particular ethical issues (Kim and Werbach 2016) as players are well informed from the very beginning that linguistic data are collected through the game and they can voluntarily choose if they want to contribute and help NLP research. In addition, the research activities are not funded by any grant, nor data collected are exploited commercially. On the contrary, the results (both the collected corpora and the platform)[8] will be made freely available to the research community under a Creative Commons (BY-NC-SA 4.0) license.

3.3 Game implementation

The game is designed as a Telegram bot to make use of Telegram's advanced features (e.g., multiplatform support) which allowed us to focus on the NLP back-end rather than building web-based or mobile versions of the game. Python-telegram-bot[13] library is used to communicate with the Telegram servers and to implement the main messaging interface. A PostgreSQL[14] database is used as the data back-end. The "Love Bird" Telegram sticker package has been used for the visualization of the selected persona, which can be changed according to the needs (e.g., with a local cultural character). For NLP-related tasks, NLTK (Loper and Bird 2002) is used for tokenization. Idioms are located in new submissions by tokenizing the submission and checking the lemma of each word whether they match that day's idiom constituents. If all idiom lemmas are found within the submission, the player is asked to choose whether the submission is an idiomatic or nonidiomatic sample. The position of the lemmas determines the type (i.e., one of the four types introduced in Section 3.2) of the submission within the system. NLTK is used for the lemmatization of English, Tint[15] (Palmero Aprosio and Moretti 2016) for the Italian and Zeyrek[16] for the

[13] https://github.com/python-telegram-bot/python-telegram-bot/.
[14] https://www.postgresql.org/.
[15] A Stanza (Qi *et al.* 2020)-based tool customized for the Italian language.
[16] An NLTK-based lemmatizer, customized for the Turkish language, https://zeyrek.readthedocs.io.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

lemmatization of Turkish.[17] If idiom lemmas are not found in the submission due to typos or incorrect entries, the player is asked to submit a new submission.

The game is designed with localization in mind. The localization files are currently available in English, Italian, and Turkish. Adaptation to other languages requires 1. translation of localization files containing game messages (currently 145 interaction messages in total), 2. a list of idioms, and 3. a lemmatizer for the target language. We also foresee that there may be need for some language-specific enhancements (such as the use of wildcard characters, or words) in the definition of idioms to be categorized under different types. The game is deployed on Docker[18] containers adjusted to each country's time zone where the game is played. In accordance with DP#4 (*following an iterative design process*) and DP#11 (*managing and monitoring to continuously optimize the gamification design*), an iterative development process has been applied. The designs (specifically the bot's messages, their timings, and frequencies) are tested and improved until they become efficient and promising to reach the goals. The system has been monitored and optimized according to the increasing workload.

4. Analysis and discussions

In accordance with DP#9 (*the definition and use of metrics for the evaluation and monitoring of the success, as well as the psychological and behavioral effects of a gamification approach*), we made a detailed analysis of the collected dataset to evaluate the success of the proposed approach for idiom corpora construction, and quantitative and qualitative analysis to evaluate its psychological and behavioral effects on users. This section first introduces the methodology and participants in Section 4.1 and then provides an analysis of the collected data in Section 4.2. It then gives intrinsic and extrinsic evaluations of the collected data in Section 4.3. The section then concludes with Section 4.4 by the analysis of the motivational and behavioral outcomes according to some constructs selected from the relevant literature.

4.1 Methodology and participants

The game was deployed three times: the first one for preliminary testing with a limited number of users, and then two consecutive 16-day periods open to crowd, for Turkish and Italian separately. The first preliminary testing of the game was accomplished on Turkish with 12 people and yielded significant improvements in the game design. The Italian preliminary tests were accomplished with around 100 people.[19] The game was played between October 13 and December 17, 2020 for Turkish, and between November 8 and December 29, 2020 for Italian. From now on, the four later periods (excluding the preliminary testing periods), for which we provide data analysis, will be referred to as TrP1 and TrP2 for Turkish, and ItP1 and ItP2 for Italian. While TrP1 and ItP1 are trials without monetary rewards, TrP2 and ItP2 are with monetary rewards.

The idioms to be played each day were selected by moderators according to their tendency to be used with their literal meaning. For ItP1 and ItP2, the selection procedure was random from an Italian idiom list,[20] where four idioms from ItP1 are replayed in ItP2 for comparison purposes. Similarly for TrP2, the idioms were selected from an online Turkish idiom list[21] again taking two idioms from TrP1 for comparison. For TrP1, the idioms were selected again with the same selection strategy but this time instead of using an idiom list, the idioms from a previous

[17] Stanza is also tested for Turkish, but outputting only a single possible lemma for each word failed in many cases in this language.

[18] https://docker.com/.

[19] Students of the third author and people contacted at EU Researchers Night at Italy.

[20] http://www.impariamoitaliano.com/frasi.htm.

[21] https://www.dilbilgisi.net/deyimler-sozlugu/.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

Table 1. User statistics

Statistic	Turkish	Italian
Total # of users who played the game	255	205
. . . for only 1 day	113 (44%)	93 (45%)
. . . for 2–3 days	87 (34%)	61 (30%)
. . . for 4–7 days	31 (12%)	32 (16%)
. . . for >= 7 days	24 (9%)	19 (9%)
Total # of users who filled in the survey:	25 (10%)	31 (15%)
# of days the survey was open:	Last 3 days of TrP2	Last 10 days of ItP2
Crowd type	AI-related people	Students, translators

annotation effort (Parseme multilingual corpus of verbal MWEs (Savary *et al.* 2018; Ramisch *et al.* 2018)) are listed according to their frequencies within the corpus and given to the moderators for the selection. Tables A2 and A3, given in the Appendix section, provide the idioms played each day together with daily submission, review statistics, and some extra information to be detailed later.

For the actual play, the game was announced on LinkedIn and Twitter for both languages at the beginning of each play (viz., TrP1, TrP2, ItP1, and ItP2). For Italian, announcements and daily posts were also shared via Facebook and Instagram. In total, there were ∼25K views and ∼400 likes/reshares for Turkish, and ∼12K views and ∼400 likes/reshares for Italian. As mentioned in the previous sections, players are informed from the very beginning that they are helping to create a public data source by playing this game. It should be noted that many people wanted to join this cooperative effort and shared the announcements from their accounts, which improved the view counts. For both languages, the announcements of the second 16-day period with monetary reward were also shared within the game itself. The Turkish crowd influencer (the first author of this article) is from NLP and AI community, and the announcements mostly reached her NLP-focused network. On the other hand, the Italian crowd influencer (the last author of this article) is from the computational linguistics community, and the announcements mostly reached students and educators. In total, there were 255 and 205 players who played the game for periods 1 and 2, respectively. Table 1 provides the detailed user statistics. As may be seen from Table 1, almost 10% of the players played the game for more than 7 days. A survey has been shared with the users at the end of TrP2 and ItP2. About 10% of the players filled in this survey.

Figure 4a shows the new player counts for each day. This graphic shows the users visiting the bot, whether they start playing or not. It can be seen that the player counts in the initial days are very high for almost all periods due to the social media announcements. The new player counts in the following days are relatively low compared to the initial days, which is understandable. Still, it may be seen that the game continues to spread except for ItP1. It should be noted that the spread also applies to Turkish (due to likes/reshares), although there had been no daily announcements contrary to Italian.

Figure 4b provides the daily player counts who either submitted or reviewed. It should be noted that the initial values between Figure 4a and b differ from each other since some players, although entering the game (contributed to the new player counts in Figure 4a), did not play it, or the old players from previous periods continued to play the game. As Figure 4b shows, for TrP1, TrP2 and ItP2 there are more than 10 players playing the game each day (except the last day of TrP1). For ItP1, the number of daily players is under 10 for 9 days out of 16. Figure 4b shows a general decline in daily player counts for TrP1 and ItP1, whereas each day, nearly 20 players played the game for TrP2 and ItP2.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

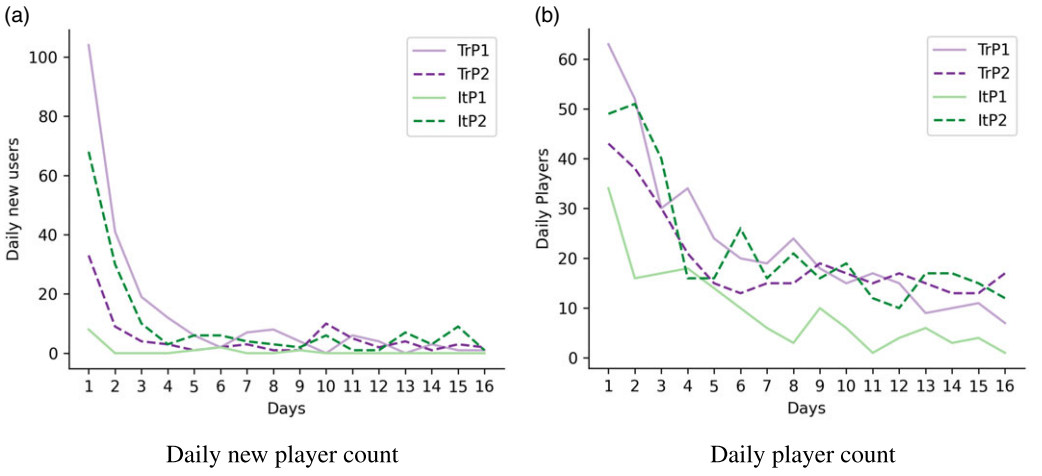


Figure 4. Daily play statistics.

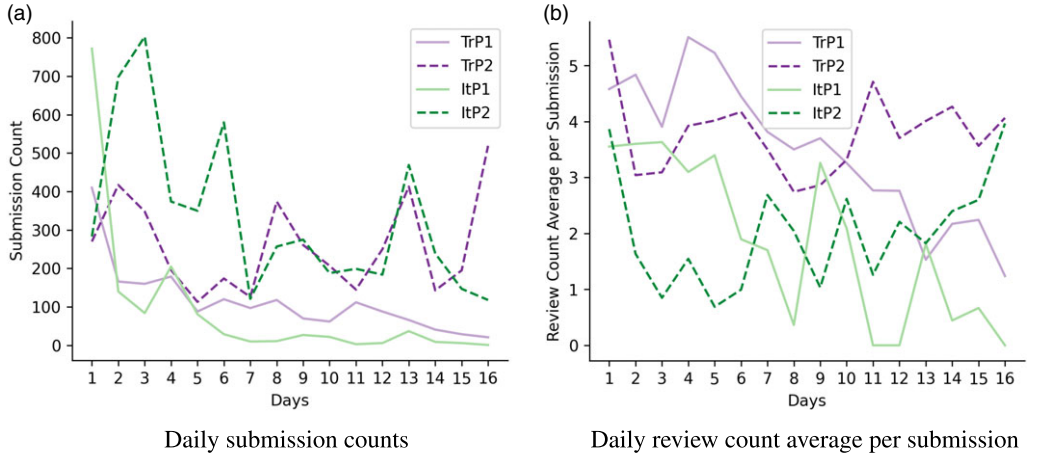


Figure 5. Daily statistics for submissions and reviews.

The following constructs are selected for the analysis of the motivational and behavioral outcomes of the proposed gamification approach: *system usage*, *engagement*, *loyalty*, *ease of use*, *enjoyment*, *attitude*, *motivation*, and *willingness to recommend* (Morschheuser *et al.* 2017; Morschheuser, Hamari, and Maedche 2019). These constructs are evaluated quantitatively and qualitatively via different operational means, that is, survey results, bot usage statistics, and social media interactions.

4.2 Data analysis

During the four 16-day periods, we collected 5978 submissions and 22,712 reviews for Turkish, and 6728 submissions and 13,620 reviews for Italian in total. In this section, we make a data analysis by providing (1) submission and average review statistics in Figure 5, (2) daily review frequencies per submission in Figure 6, and (3) collected sample distributions in Figure 7 according to the sample categories provided in Section 3.2. The impact of the monetary reward can be observed on all figures, but the comparisons between periods with and without monetary reward

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

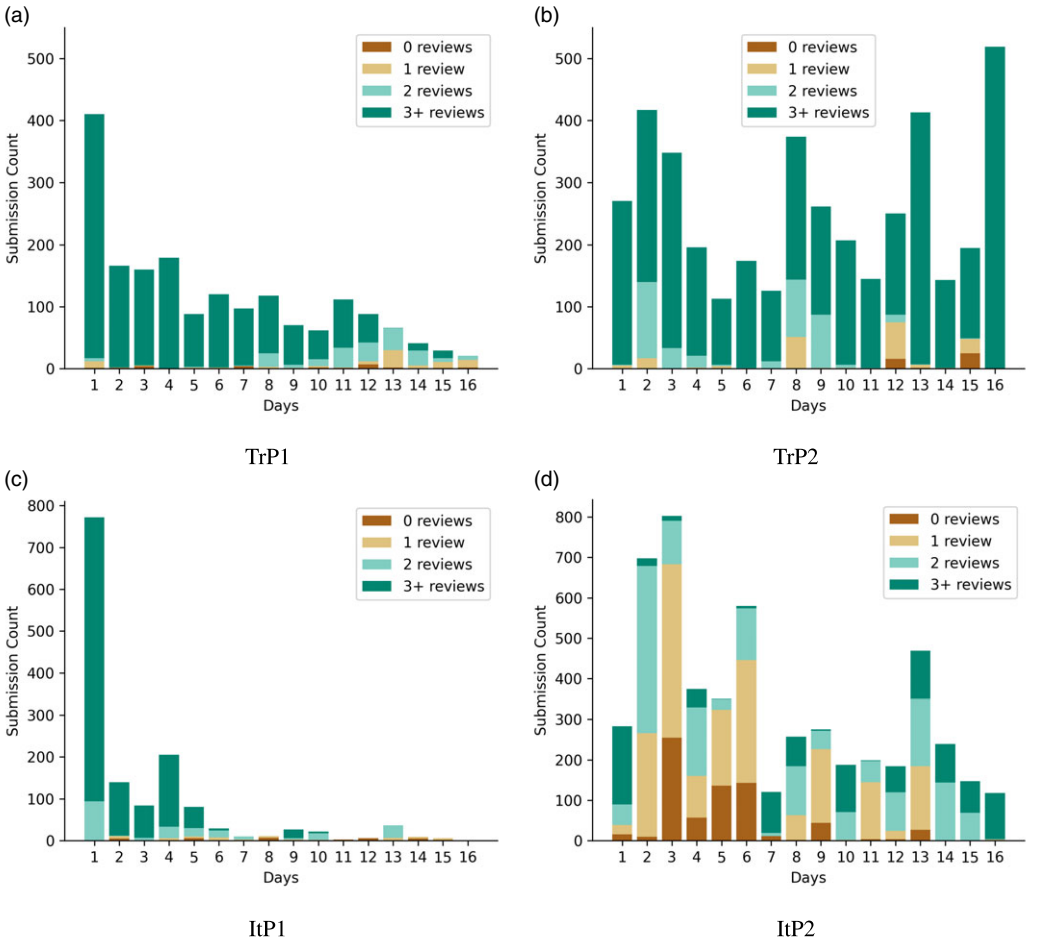


Figure 6. Daily review frequencies per submission.

are left to be discussed in Section 4.4 under the related constructs. In this section, although the analyses are provided for all the four periods, the discussions are mostly carried out on TrP2 and ItP2, which yielded a more systematic data collection (see Figure 5a—daily submission counts).

Figure 5a shows that the soft target of 100 submissions per idiom is reached for both of the languages, most of the time by a large margin: 258 submissions on daily average for Turkish and 330 submissions for Italian. The average review counts are most of the time above 3 for Turkish idioms with a mean and its standard error of 3.7±0.2, whereas for Italian this is 2.0±0.2. The difference between averages may also be attributed to the crowd type (mostly AI-related people for Turkish and students for Italian), which is again going to be discussed in the next section. But in here, we may say that in ItP2, especially in the first days, the submission counts were quite high and review averages remained relatively lower when compared to this. However, since we have many samples on some days, although the review average is low, we still have many samples that have more than two reviews. Figure 6 shows the review count distributions per submission. As an example, when we look at Figure 6d, the third day of ItP2 (which received 803 samples with 0.8 reviews in average Table A3), we may see that we still have more than 100 hundred samples (specified with green colors) which received more than two reviews. On the other hand, TrP2 results (Figure 6b) show that there are quite a lot of submissions that are reviewed by at least three

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

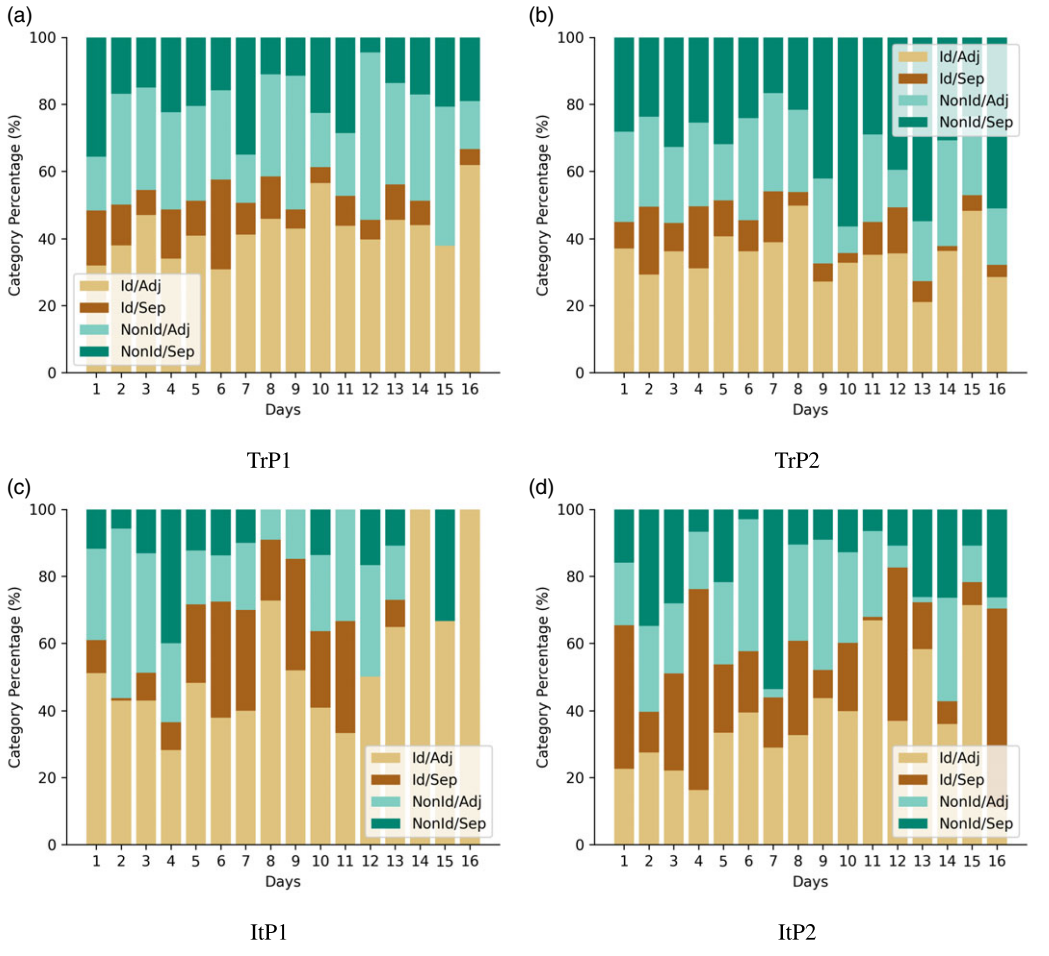


Figure 7. Daily sample type distributions.

people. Similarly for TrP1 (Figure 6a) and ItP1 (Figure 6c), although the submissions counts are lower, most of them are reviewed by at least two people.

The Appendix tables (Tables A2 and A3) also provide the dislikes percentages for each idiom in their last column. The daily averages are 15.5±2.7% for TrP2 and 24.1±3.2% for ItP2. It should be noted that two days (6th and 11th) in ItP2 were exceptional, and the dislike ratios were very high. In those days, there were players who entered very similar sentences with slight differences, and reviewers caught those and reported. It was also found that these reported players repeatedly sent dislikes to other players' entries. The moderators had to ban them, and their submissions and reviews were excluded from the statistics. No such situation had been encountered in TrP2 where the idiom with the highest dislike ratio appears in the eighth day with 36%. Although the data aggregation stage[22] is out of the scope of this study, it should be mentioned that despite this ratio, we still obtained many fully liked examples (87 out of 374 submissions, liked by at least 2 people).

[22] "One of the biggest challenges of crowdsourcing is aggregating the answers collected from the crowd, since the workers might have wide-ranging levels of expertise. In order to tackle this challenge, many aggregation techniques have been proposed" (Quoc Viet Hung *et al*. 2013).

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

Figure 7 shows the type distributions (introduced in Section 3.2) of the collected samples. We see that the scoring, notifications, and tips helped to achieve our goal of collecting samples from various types. The type ratios change from idiom to idiom according to their flexibility. When we investigate Id/Sep samples for Italian, we observe pronouns, nouns, and mostly adverbs intervening between the idiom components. Italian Id/Sep samples seem to be more prevalent than Turkish. This is due to the fact that possession is generally represented with possessive suffixes in Turkish and we do not see any Id/Sep occurrences due to this. The possessive pronouns, if present, occur before the first component of the idiom within the sentence. For Turkish, we see that generally interrogative markers (enclitics), adverbs, and question words intervene between the idiom components. We see that some idioms can only take some specific question words, whereas others are more flexible.

As explained at the beginning, collecting samples with different morphological varieties was also one of our objectives for the aforementioned reasons. When we investigate the samples, we observe that the crowd produced many morphological variations. For example, for the two Turkish idioms "#4 – karşı çıkmak" ("in front of – to climb/to step up" → *to oppose*) and "#16 defterden silmek" ("from notebook – to erase" → *to forget someone*), we observe 65 and 57 different surface forms in 167 and 97 idiomatic samples, respectively. For these idioms, the inflections were mostly on the verbs, but still, we observe that the first constituents of the idioms were also seen under different surface forms ("karşı" (*opposite*) in four different surface forms inflected with different possessive suffixes and the dative case marker, and "defterden" (*from the notebook*) in five different surface forms inflected with different possessive suffixes with no change on the ablative case marker). We also encounter some idioms where the first constituent only occurs under a single surface form (e.g., "#8 sıkı durmak" (*to stay strong or to be ready*)). The observations are in line with the initial expectations, and the data to be collected with the proposed gamification approach are undeniably valuable for building knowledge bases for idioms.

4.3 Intrinsic and extrinsic evaluation of the collected dataset

This section presents intrinsic and extrinsic evaluations of the dataset collected through Dodiom. Section 4.3.1 first provides the evaluation of the collected datasets by linguists and then an evaluation compared to a traditional data annotation approach. Section 4.3.2 evaluates the collected dataset within an automatic idiom identification system and shows that the introduced crowdsourcing approach provides outputs on a par with the classical data annotation approach for training such systems.

4.3.1 Intrinsic evaluations

Expert evaluations:

In order to evaluate the collected dataset in terms of linguistic quality, three linguists for each language, who are native speakers of the related language and had not played the game, manually annotated a subset of the dataset, that is, 300 submissions for each language. The subsets were randomly selected from the idioms receiving at least three ratings for both languages. In order to guide the annotators about the quality assessment, we first provided a set of criteria and asked them to make binary evaluations on these: viz., "wrong category," "undecidable," "low context," "vulgar," "incorrect grammar," "incorrect spelling," "meaningless," "negative sentiment," and "restricted readers" with the label "0" as the default value when the submission may not be treated under the relevant criterion and the label "1" when the submission matches the relevant criterion. We then asked our linguists to give a quality score of either "0" (for bad quality), "1" (for good quality), or "2" (for excellent quality—very good examples which can be included in dictionaries and language learning resources) to the submissions. The assessments of the experts

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

differed from each other, as expected. We used Fleiss Multirater Kappa[23] to measure the inter-annotator agreements (IAAs). For the "wrong category" evaluations, Kappa was measured as 0.5 for Italian and 0.4 for Turkish. For "quality" evaluations, they were measured as 0.4 and 0.2 for Italian and Turkish, respectively.[24]

To explore the correlation between the crowd and expert assessments, we aggregated the expert scores, giving an overall expert quality score between 0 and 6 for each submission and an expert criterion score between 0 and 3 for the above-listed nine criteria. For crowd evaluation, we calculated a like ratio for each submission: the ratio of the number of likes for the related submission to the number of its total reviews. 145 over 300 submissions for Turkish and 87 for Italian were rated as excellent quality by the experts with full agreement. An overall expert quality score between 3 to 6 was calculated for 266 and 245 samples for Turkish and Italian, respectively, which shows that more than 50% of the collected dataset was found valuable by our experts. For the submissions on which the three experts fully agreed that the category (idiomatic/nonidiom) was correct, the crowd like ratio was measured as 0.9 for Turkish (268 submissions in total) and 0.6 for Italian (287 submissions in total). For the submissions on which the three experts fully agreed that the category was wrong, the crowd like ratio was measured as 0.4 for Turkish (six submissions in total) and 0.3 for Italian (two submissions in total). This analysis showed that the crowd liked the incorrectly marked samples less than the correctly marked ones.

We made a correlation analysis (i.e., Pearson's correlation) between the crowd and linguists' assessments for the "wrong category" and "quality" criteria. We observed a statistically significant correlation ($p < 0.01$) between the experts' quality scores and crowd like ratios for Turkish with Pearson's $r = 0.354$. We observed a statistically significant correlation ($p < 0.01$) between the "wrong category" evaluations by the experts and the crowd like ratios for both languages: with $r = 0.302$ for Turkish and $r = 0.198$ for Italian.

Comparison with Parseme Annotations:

Parseme multilingual corpus of verbal MWEs (Savary *et al*. 2018; Ramisch *et al*. 2018) contains 280K sentences from 20 different languages including Italian and Turkish. As the name implies, this dataset includes many different types of verbal MWEs including verbal idioms. It is therefore very convenient to use as an example for the output of a classical annotation approach. During the preparation of this corpus, datasets, retrieved mostly from newspaper text and Wikipedia articles, were read (i.e., scanned) and annotated by human annotators according to well-defined guidelines. Since MWEs are random in these texts, only the surrounding text fragments (longer than a single sentence) around the annotated MWEs were included in the corpus, instead of the entire scanned material (Savary *et al*. 2018). Due to the selected genre of the datasets, it is obvious that many idioms, especially the ones used in colloquial language, do not appear in this corpus. Additionally, annotations are error-prone as stated in the previous sections. Table 2 provides the statistics for Turkish and Italian parts of this corpus.

In order to compare the outputs of the classical annotation approach and the gamified construction approach, we select four idioms for each language (from Tables A2 and A3) and manually check their annotations in the Parseme corpus. For Turkish, we select one idiom which is annotated the most in the Parseme corpus ("yer almak"—*to occur* 123 times[25]), one which appears very few ("zaman öldürmek"— *to waste time* 1 time) and two which appear in between. The selected idioms are given in Table 3. For Italian, since the idioms were selected from an idiom

[23] The statistical analysis was made using IBM SPSS Statistics for Windows, version 27.

[24] For the sake of brevity, we only provide the analysis for "wrong category" and "quality" criteria which are more meaningful to compare with crowd assessments.

[25] "yer" and "al" are also the lemmas of another idiom "yerini almak"—*take (someone's or something's) place*. There is no distinction for different idioms in Parseme annotations (132 idiom annotations with these lemmas). The numbers in Table 3 refers to the mentioned idiom's counts ("yer almak"—*to occur*). We also observe 12 false negatives for the idiom "yerini almak."

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

Table 2. Parseme Turkish and Italian datasets (Savary *et al.* 2018)

	Turkish	Italian
# of annotated sentences	18,036	17,000
# of MWE annotations	6670	2454
VID (Verbal Idioms)	3160	1163
LVC (Light-Verb Constructions)	2823	482
VPC (Verb-Particle Constructions)	0	73

Table 3. Comparison with classical data annotation. (Id.: # of idiomatic samples, NonId.: # of nonidiomatic samples, Rev.: average review count, Unnon.: # of unannotated sentences containing lemmas of the idiom components, Fn.: # of false negatives, please see Tables A2 and A3 for the meanings of idioms)

		Dodiom			Parseme		
Lang.	Idiom	Id.	NonId.	Rev.	Id.	Unann.	Fn.
Turkish	yer almak	69	49	3.5	123	83	18
	meydana gelmek	103	92	3.6	29	25	16
	karşı çıkmak	167	352	4.1	27	46	14
	zaman öldürmek	123	127	3.7	1	5	0
Italian	aprire gli occhi	143	132	1.0	2	0	0
	prendere con le pinze	409	394	0.8	1	0	0
	essere tra i piedi	152	32	2.2	0	3	0
	mandare a casa	102	137	2.4	2	2	0

list (as opposed to Turkish (Section 4.1)), their occurrence in the Parseme corpus is very rare as may be seen from Table A3. Thus, we selected the four idioms with the highest counts.

As stated before, only the idiomatic samples were annotated in the Parseme corpus. To further analyze, we retrieved all the unannotated sentences containing the lemmas of the idioms' constituents and checked to see whether they are truly nonidiomatic usages or are mistakenly omitted by human annotators (i.e., false negatives (Fn)).[26] 3 human annotators[27] worked on this manual control (Fleiss' kappa of 0.97). As may be seen from Table 3, the mistakenly omitted idiomatic samples (the last column) are quite high, although this dataset is reported to be annotated by two independent research groups in two consecutive years: for example, 16 idiomatic usage samples for the idiom "meydana gelmek" (*to happen*) were mistakenly omitted out of 25 unannotated sentences. Similar to the findings of Bontcheva *et al.* (2017) on named entity annotations, these results support our claim about the quality of the produced datasets when the crowd focuses on a

[26]Selecting certain types of examples for additional annotation is relatively common (e.g., in crowd-rating situations, when the first two judges disagree by a certain amount, a third judge is brought in, whether another crowd worker or an expert). Here, we use the conflicting samples between a human and a basic automated computer program that only matches the lemmas for identification.

[27]The three human annotators are native Turkish speakers and NLP researchers acting in the ITU NLP team, the same group that had made the original annotations on the Parseme data Turkish section. The annotators focused only on sentences probably overlooked, as they did not contradict existing idiomatic annotations.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

single phenomenon at a time. Additionally, the proposed gamified approach (with a crowd-rating mechanism) also provides multiple reviews on the crowd-created dataset.

When the idiomatic annotations in Parseme are investigated, it is seen that they are mostly Id/Adj samples and Id/Sep samples very rarely appear within the corpus (95% of Turkish verbal idioms and 74% of Italian verbal idioms), which could be another side effect of the selected input text genres.

4.3.2 Extrinsic evaluations

For the extrinsic evaluation of the collected dataset, we use an idiom identification architecture, namely, a BiLSTM-CRF model used in many recent studies for idiom and MWE identification (Boros and Burtica 2018; Berk, Erden, and Güngör 2018; Yirmibeşoğlu and Güngör 2020; Ehren *et al.* 2020; Saxena and Paul 2020) within the literature. Idiom identification tasks are generally modeled as assigning two or more labels to each token within a sentence. In our first set of experiments, we use three labels: Idiom (I), Literal (L), and Other (O). For representing the words within our deep learning architecture, we use pretrained Fasttext (Bojanowski *et al.* 2017) embeddings both for Turkish and Italian. Each sentence is fed to the BiLSTM-CRF model (dimension size 2×100 with 0.5 dropout) as a word sequence and the label sequence is generated as the output. The initial weights are kept constant across all our experiments.

Figure 8 provides token-based macro average F1 scores at the end of 5-fold cross-validation[28] over the collected datasets for Turkish in the first row and Italian in the second row. As explained in previous sections, the submitted sentences receive different numbers of reviews. We investigate the impact of the data quality from the crowd's point of view on the idiom identification system. With this purpose, we evaluate the idiom identification system with different data subsets split according to the received review counts and like ratios (i.e., the ratio of the number of likes for the related submission to the number of its total reviews). Each row in Figure 8 has four subfigures showing performances on subsets with review counts bigger than 0, 1, 2, and 3, respectively: the first subfigure of the Turkish row shows the tests on the entire dataset (with review count ≥ 0), whereas the last subfigure of the same row shows the tests on a smaller subset (85% of the collected Turkish corpus). Similarly, for Italian, the last subfigure shows the tests on 44% of the collected Italian corpus with submissions that received at least three reviews. Additionally, each subfigure provides the performances according to like ratios. Each column bar in the figure provides the results for the data splits with like ratios bigger than some thresholds between 0 and 1. For example, when the threshold is 0 the entire data within that subset is used, and when it is 1, this means that only submissions that are fully liked by the crowd are used during the experiments. This filtering results in a drastic data size decrease in the last experiments for both rows: for Turkish, only 53% of the collected data is used for review count ≥ 3 and like ratio ≥ 1 (last column bar in first row fourth subfigure). For Italian, this ratio is only 12% due to low review counts reported in Section 4.2.

We see that the corpus size has an important impact on the identification performances. We observe a noticeable drop in performances (from 0.86 to 0.65 F1 score) for Italian when 12% of the dataset is used in its last experiment (review count ≥ 3 and like ratio ≥ 1) compared to 100% in its first experiment (review count ≥ 0 and like ratio ≥ 0). However, data quality also seems to be a very important factor during training. For Turkish experiments, although the data size is almost halved, the performances remain within the same ranges: 82.6% versus 81.4%. This shows that the same idiom identification system may be trained with a lower amount of high-quality data, and the crowd performed well on selecting these examples. As a result of these experiments, we may say that the crowd-rating approach seems successful, and players are good at choosing high-quality samples. We see that although the data size decreases with our review count and

[28]The datasets are randomly shuffled before splitting the cross-validation folds.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

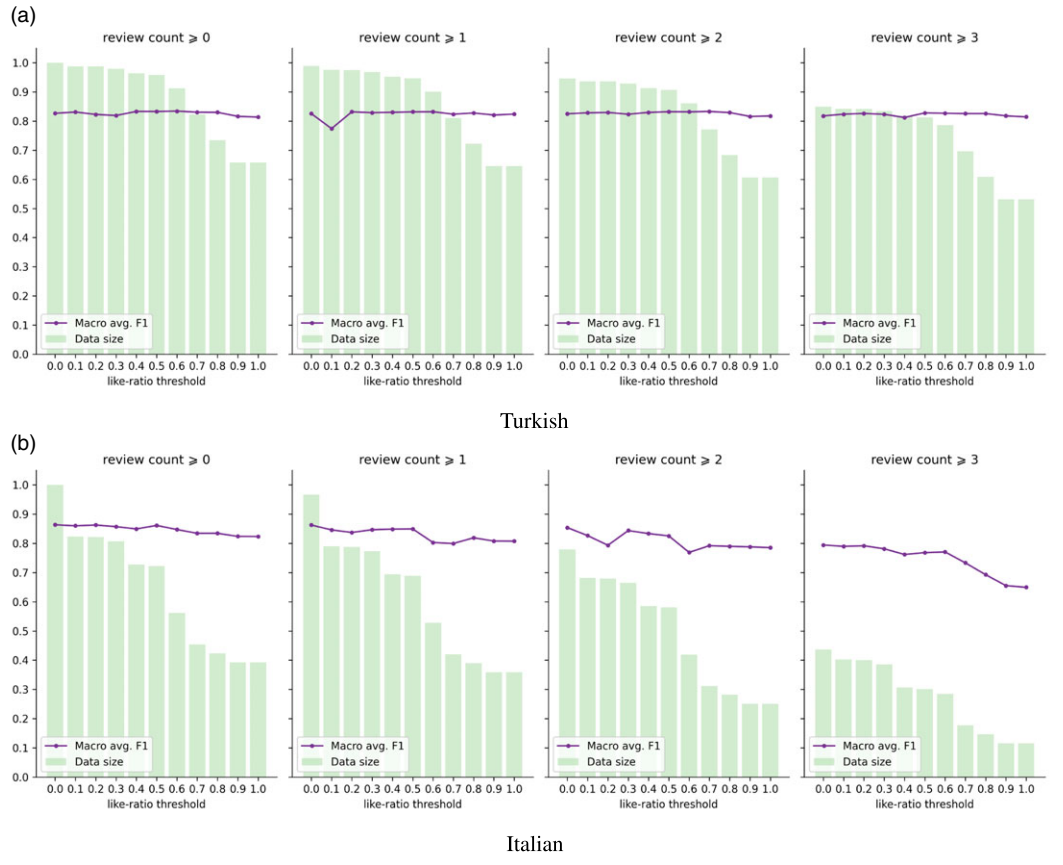


Figure 8. Performances of the idiom identification model with respect to crowd ratings.

threshold selections, the obtained performances remain almost the same across different experiments. On the other hand, the use of the entire collected data without any crowd evaluation did not harm the system performances showing that although there could be erroneous submissions, the introduced noise due to this does not affect the identification performances showing that the crowd-creating approach is also successful for collecting such datasets.

Parallel to the previous section (Section 4.3.1), we use the Parseme dataset for the second extrinsic evaluation. This experiment investigates the impact of adding our collected dataset to the Parseme dataset on a similar idiom identification task. However, since Parseme dataset does not contain annotations for literal cases, in this set of experiments, we use only two labels: Idiom (I) and Other (O). In order to prepare the dataset, we label all verbal idioms (VIDs) in the Parseme dataset with label I and the remaining tokens with label O. For the Dodiom dataset, we make a similar preprocessing and replace all the L labels with the label O. We again use 5-fold cross-validation during our experiments. We first run the experiments with the Parseme dataset alone, and then in the second step, we add the Dodiom dataset to the training folds (in addition to the Parseme data), keeping the test folds fixed. Since the Italian section of the Parseme dataset and the Dodiom dataset share very few idioms in common (Section 4.1), this experiment is conducted only on Turkish datasets.

Figure 9 shows the performances of the two idiom identification experiments with respect to training epochs. The figure shows that the augmentation of the Parseme dataset with the Dodiom dataset generally provides better performances at all epochs. The performance improvement is very slight after the 150th epoch. However, we see that the augmentation of the collected dataset

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

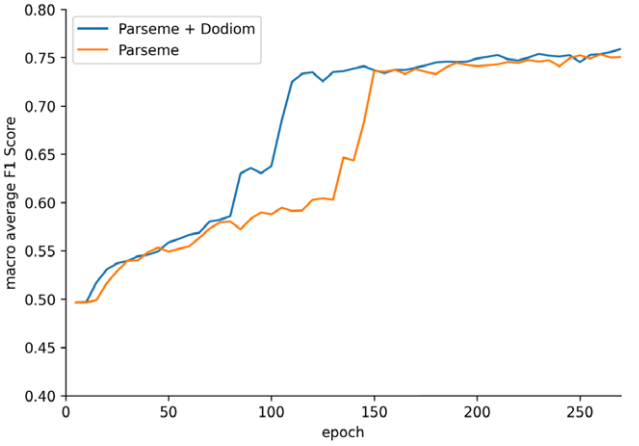


Figure 9. Impact of augmenting the Parseme dataset with the Dodiom dataset on idiom identification performances.

allows the idiom identification system to converge at earlier epochs: that is, 80th instead of 150th epoch. We should note that in this experiment, we used the entire collected data without considering their like ratio or review counts. This shows that the introduced crowd-creating approach provides outputs on a par with a classical data annotation approach for training such systems.

4.4 Motivational and behavioral outcomes

In this section, we provide our analysis on motivational and behavioral outcomes of the proposed gamification approach for idiom corpora construction. The survey results (provided in Table 4), bot usage statistics (provided in Section 4.2), and social media interactions are used during the evaluations. The investigated constructs are *system usage*, *engagement*, *loyalty*, *ease of use*, *enjoyment*, *attitude*, *motivation*, and *willingness to recommend*.

Table 4 summarizes the survey results in terms of response counts provided in the last two columns for Turkish and Italian games, respectively. In questions with 5-point Likert scale answers, the options go from 1: strongly disagree or disliked to 5: strongly agree or liked. The first four questions of the survey are related to demographic information. The answers to question 2 (Q2 of Table 4) reveal that the respondents for Turkish play are mostly AI- and computer technology-related people (21 out of 25 participants selected the related options and 2 stated NLP under the *other* option), whereas for Italian play they are from different backgrounds; 21 people out of 31 selected the *other* option, where only 2 of them stated NLP and computational linguistics, and the others gave answers like translation, student, administration, tourism, and sales. The difference between crowd types seems to also affect their behavior. In TrP2, we observe that the review ratios are higher than ItP2 as stated in the previous section. On the other hand, ItP2 participants made more submissions. There were more young people in Italian plays (Q3) than Turkish plays. This may be attributed to their eagerness to earn more points. We had many free text comments (to be discussed below) related to the low scoring of the review process from both communities.

The overall *system usage* of the participants is provided in Section 4.2. Figures 4b and 5 shows player counts and their play rates. Although, only 50% of survey Q7 answers, about the gift card *motivation*, says agree (4) or strongly agree (5), Figures 4b and 5 reveal that the periods with additional incentives (i.e., gift card rewards) (TrP2 and ItP2) are more successful at fulfilling the expectations about *loyalty* than the periods without (TrP1 and ItP1). Again related to the *loyalty* construct (Q18 and Q19), we see that more than half of the Turkish survey participants were playing the game for more than 1 week at the time of filling out the survey (which was open

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

Table 4. Survey constructs, questions and results. (Answer types: 5-point Likert scale (5PLS), predefined answer list (PL), PL including the "other" option with free text area (PLwO))

Q	Constructs	Survey questions	Answer type	Turkish	Italian
1	demographic	-What is your educational background? PL:{from 1:primary school to 5:PhD}	PL	0 0 9 12 4	0 2 6 20 3
2	demographic	-What field do you work in? PLwO:{education, AI, computer tech., other}	PLwO	0 12 8 5	6 2 2 21
3	demographic	-How old are you? PL:{<18, 18-25, 25-30, >30}	PL	0 9 9 7	0 14 12 5
4	demographic	-How did you hear about Dodiom? PLwO:{Linkedin, Twitter, a friend, other}	PLwO	7 4 10 4	6 7 9 9
5	attitude	-What's your opinion about Dodiom?	5PLS	0 0 0 10 15	0 2 1 15 12
6	motivation	-Why did you play Dodiom, what was the main motivation for you to play? PLwO:{help dodo, daily achievements, fun, help NLP studies, other}	PLwO	0 4 0 20 1	1 4 4 21 1
7	motivation	-The Gift Certificate was an important motivation for me to play the game	5PLS	4 2 7 4 8	6 1 6 8 10
8	enjoyment	-The leaderboard and racing components made the game more fun	5PLS	0 1 2 5 17	3 2 5 7 14
9	engagement	-Dodo's messages about my place in the rankings increased my participation in the game	5PLS	1 0 4 9 11	4 3 3 8 13
10	attitude	-I liked the interface of the game and the ease of play, it kept me playing the game	5PLS	0 1 0 5 19	0 0 9 10 12
11	ease of use	-I was able to learn the gameplay of the game without much effort	5PLS	0 0 1 2 22	0 0 1 7 23
12	engagement	-The frequency of Dodo's notifications was not disturbing	5PLS	4 2 8 3 8	4 4 10 7 6
13	enjoyment	-The theme and gameplay was fun, I enjoyed playing	5PLS	0 0 1 8 16	0 1 4 11 15
14	loyalty	-Dodo will take a break from learning soon. Do you want to continue helping when it starts again? PLwO:{yes, no, other}	PLwO	24 0 1	28 2 1
15	attitude	-Which aspect of the game did you like the most?	free-text	–	–
16	attitude	-Was there anything you did not like in the game, and if so, what?	free-text	–	–
17	loyalty	-How many days did you play Dodiom? PL:{1, 2-3, <1week, >1 week}	PL	2 2 4 16	7 10 7 7
18	loyalty	-How many samples did you send to Dodiom per day on average? PL:{2-3, <10, 10-20, >20}	PL	3 7 6 9	12 6 7 5
19	–	-Can you share any suggestions about the game?	free-text	–	–

for the last three days of TrP2) and they were providing more than 10 samples each day. Since the Italian survey was open for a longer period of time (see Table 1), we see a more diverse distribution on the answers. Most of the participants also stated that they would like to continue playing the game (Q14).

A very high number of participants (20 out of 25 Turkish participants, and 21 out of 31 Italian participants) stated that their *motivation* to play the game was to help NLP studies (Q6). This is

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

a very encouraging outcome for further research on such gamification approaches for collecting NLP-related data, which we hope the platform we provide will help. Four of them answered Q15 as: *"I felt that I'm helping a study," "The scientific goal," "The ultimate aim," "I liked it being the first application designed to provide input for Turkish NLP as far as I know. Apart from that, we are forcing our brains while entering in a sweet competition with the friends working in the field and contributing at the same time."* We see that the gamification elements and the additional incentive helped the players to stay on the game with this motivation (Q8, Q13 *enjoyment*). In TrP2, we also observed that some game winners shared their achievements on social media (*willingness to recommend*) and found each other on the same channel. Setting more moral goals than monetary rewards, they combined distributed bookstore gift cards and sent book gifts to poor children by using these. Around 800 social media likes and shares were made in total (for both languages). More than half of the respondents chose the answer "from a friend" or "other" to Q4 ("How did you hear about Dodiom?") instead of the first two options covering Linkedin and Twitter. The "other" answers were covering either the name of the influencers, or Facebook and Instagram for Italian. We may say that the spread of the game (introduced in Section 4.1) is not due to the social media influences alone but people let each other now about it, which could also be seen as an impact of their *willingness to recommend* the game.

Almost all of the users found the game easy to understand and play (Q11 *ease of use*). Almost all of them liked the game; only 3 out of 31 Italian participants scored under 4 (liked) to Q5 (*attitude*) and 9 of them scored neutral (3) to Q10. Only one Turkish participant was negative to this later question about the interface. When we analyze the free-text answers to Q15, we see that 8 people out of 56 stated that they liked the game elements. Some answers are as follows: *"I loved Dodo gifts," "Gamification and story was good," "It triggers a sense of competition," "The icon of the application is very sympathetic, the messages are remarkable," "I liked the competition content and the ranking," "Gift voucher," "Interaction with other players."* Three participants stated that they liked the game being a bot with no need to download an application. Three of them mentioned that they liked playing the game: *"I liked that the game increases my creativeness. It makes me think. I'm having fun myself," "To see original sentences," ". . . Besides these, a fun opportunity for mental gymnastics," "Learn new idioms," "Linguistic aspect."* Eight participants stated that they liked the uniqueness of the idea: *"The originality of the idea," "The creativity," "Efficiency and immediacy," "The chosen procedure," "The idea is very useful for increasing the resources for the identification of idiomatic expressions," "The idea of being interacting with someone else," "Undoubtedly, as a Ph.D. student in the field of NLP, I liked that it reduces the difficulty of labeling data, makes it fun, and is capable of enabling other people to contribute whether they are from the field or not."*

More than half of the participants were okay with the frequency of the Dodo's instant messages and most of them agreed about their usefulness in keeping them in the game (Q9 and Q12). Four people out of 56 participants in total complained about the frequency of the messages as an answer to Q16 (*"Slightly frequent notifications," "Notifications can be sent less often," "Too many notifications"*). As opposed to this, one participant said *"It is nice that when you put it aside, the reminders and notifications that encourage you to earn more points make me re-enter words during the day"* as an answer to Q15.

Other answers to Q16 are as follows: *"I do not think it should allow the possibility of repeating the same sentences," "It can get repetitive, a mixed-mode where automatically alternating between suggestions and evaluations with multiple expressions per day would have been more engaging," "Error occurrence during voting has increased recently. Maybe it could be related to increased participation. However, there is no very critical issue," "Sometimes it froze."* Regarding the last two comments, we have stated in the previous sections the need for optimization towards the end of the play with the increased workload and the action taken. On the other hand, the first two comments are also very good indicators for future directions.

For Q19, we received three suggestions for the scoring system, one suggestion for automatic spelling correction, two suggestions for detailing dislikes, and one suggestion for the need to

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

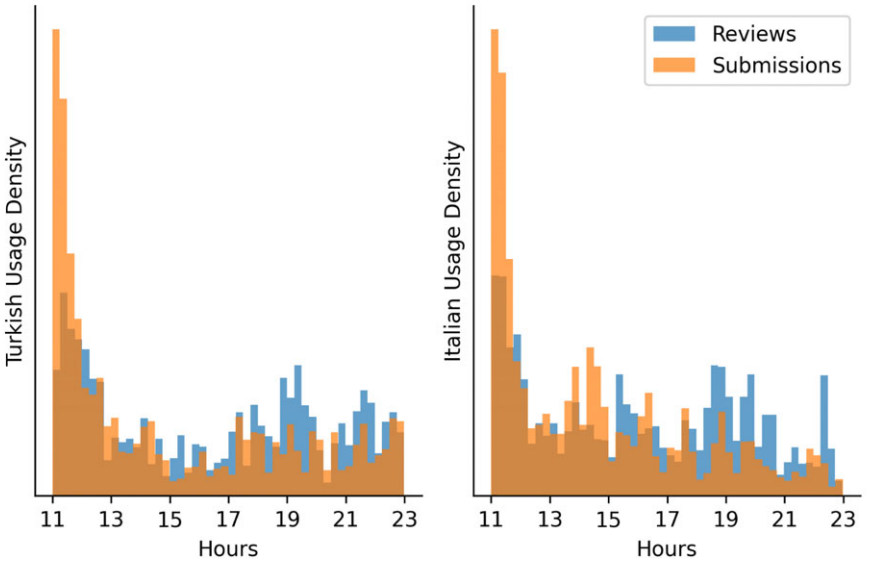


Figure 10. Histogram of interaction times in TrP2 and ItP2.

cancel/change an erroneous submission or review. Obviously, the users wanted to warn about spelling mistakes in the input sentences but hesitated to send a dislike due to this. That is why they suggested differentiating dislikes according to their reasons. Suggestions for scoring are as follows: *"More points can be given to the reviews," "The low score for reviews causes the reviewing to lose importance, especially for those who play for leadership. Because while voting, you both get fewer points and in a sense, you make your opponents earn points.", "I would suggest that the score was assigned differently, that is, that the 10/15 points can be obtained when sending a suggestion (and not when others evaluate it positively). In this way, those who evaluate will have more incentives to positively evaluate the suggestions of others (without the fear of giving more points to others) (thus giving a point to those who evaluate and one to those who have been evaluated)."* We see that in the last two comments, the players are afraid of making other players earn points.

As explained in the game design section above, the reviews worth 1 point and sometimes 2 in happy hours, triggered by the moderators to attract the attention of the players. Although open for discussions and changes in future trials, in the original design, we did not want to give high points to reviews since we believe that people should review with the responsibility of a cooperative effort to create a public dataset. Giving very high scores to reviews may lead to unexpected results. Other scenarios together with cheating mechanisms (such as consecutive rapid likes/dislikes detection) may be considered in future works. As stated before, we had some reporting and banning mechanisms added to control cheating/gaming the system in line with DP#10. The literature recommends that this is necessary since it can reverse the effects of gamification and discourage users. "However, some experts reported that cheating could also help to better understand the users and to optimize gamification designs accordingly" (Morschheuser *et al.* 2018). As future work, automatic cheating detection for detecting rephrases and malicious reviews may be studied.

"Tailoring the game elements according to the users' profile is a way to improve their experience while interacting with a gamified system, and has been noted as a current trend in gamification research" (Klock *et al.* 2020). We tested the game in an asynchronous multiplayer game scenario where the players are free to choose the time they want to contribute according to their schedule. Figure 10 shows the interaction times of the users, where the submissions are high at the beginning of the day and the reviews surpass the submissions towards the end of the day. Also, the individual

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

peaks and increased density near the end of the days correspond to the happy hour notifications sent by moderators (generally around 5 p.m. Istanbul Time for both languages, and observed as peaks around 7 p.m. in Figure 10 Italian graphic on the right). However, other more condensed timings may also be considered depending on the crowd in focus.

5. Conclusion

Idiom corpora are valuable resources for foreign language learning, NLP, and lexicographic studies. Unfortunately, they are rare and hard to construct. For the first time in the literature, this article introduced a gamified approach that uses crowd-creating and crowd-rating techniques to speed up idiom corpora construction for different languages. The approach has been evaluated under different motivational strategies on two languages, which produced the first idiom corpora for Turkish and Italian. The implementation developed as a Telegram messaging bot and the collected data for the two languages in a time span of 30 days are shared with the researchers. Our detailed qualitative and quantitative analyses revealed that the outcomes of the research are appreciated by the crowd, found useful and enjoyable, and yielded to the collection and assessment of valuable samples that illustrate the different ways of use (i.e., idiomatic/non-idiomatic and adjacent/separated usages under different inflections), which is not easily achievable with traditional data annotation techniques. Gift cards were found to be very effective in incentivizing the users to continue playing the game in addition to gamification affordances. Linguists manually evaluated subsets of the collected dataset, and analysis showed a significant correlation between the crowd and their assessments. The collected data were used within an automatic idiom identification system and shown to be successful for training such systems.

Our first short-term goal is to extend and play the game for languages other than the ones in this article, especially for languages with few lexical resources. We hope that the game introduced as an open-source project will speed up the development of idiom corpora and the research in the field. The game currently targets adult native speakers. During the initial sharing of the first prototype with the stakeholders, the initial reaction of language teachers was to use the game within classrooms as a teaching aid as well. For future direction, one may consider developing a different mode of the game for within classroom settings for both native speaker students and foreign language learners. The game may be enhanced to be played under the moderation of the teachers.

Acknowledgments. The authors would like to offer their special thanks to Cihat Eryiğit for the discussions during the initial design of the game, Fatih Bektaş, Branislava Šandrih, Josip Mihaljević, Martin Benjamin, Daler Rahimjonov, and Doruk Eryiğit for fruitful discussions during its implementation, Federico Sangati for providing the codes of a telegram bot example (Plagio which is used in another game (Grace Araneta *et al*. 2020) for foreign language learners practicing phrasal verbs), Martin Benjamin for helping on the English localization messages, Inge Maria Cipriano for discussion concerning the game interactions and possible deceptive behaviors by players, Bihter Dereli, Selin Karlı, Esin Ezgi Yıldız, Adriana Capasso, Giovanna Carandente, and Giuseppina Morza for manually evaluating a subset of the collected dataset in terms of linguistic quality, and all the anonymous volunteer players. The study has been proposed as a task by the first author and took place in a crowdfest event (in February 2020 in Coimbra, Portugal) of EU COST Action (CA16105) Enetcollect, where the prototype has been introduced and discussed with stakeholders. The authors would like to thank Enetcollect for this opportunity which initiated new collaborations and ideas. Finally, the authors would like to express their special thanks to all reviewers for their valuable comments.

References

Akkaya C., Conrad A., Wiebe J. and Mihalcea R. (2010). Amazon Mechanical Turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles. Association for Computational Linguistics, pp. 195–203.

Artignan G., Hascoët M. and Lafourcade M. (2009). Multiscale visual analysis of lexical networks. In *2009 13th International Conference Information Visualisation*, pp. 685–690.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

Berk G., **Erden B. and Güngör T.** (2018). Deep-BGT at PARSEME shared task 2018: Bidirectional LSTM-CRF model for verbal multiword expression identification. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 248–253.

Birke J. and Sarkar A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics, pp. 329–336.

Bojanowski P., **Grave E.**, **Joulin A. and Mikolov T.** (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **5**, 135–146.

Bontcheva K., **Derczynski L. and Roberts I.** (2017). Crowdsourcing named entity recognition and entity linking corpora. In **Ide**, **N. and Pustejovsky**, **J.** (eds), *Handbook of Linguistic Annotation*. Dordrecht, Netherlands: Springer, pp. 875–892.

Boros T. and Burtica R. (2018). GBD-NER at PARSEME shared task 2018: Multi-word expression detection using bidirectional long-short-term memory networks and graph-based decoding. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 254–260.

Caruso V., **Barbara B.**, **Monti J. and Roberta P.** (2019). How can app design improve lexicographic outcomes? examples from an italian idiom dictionary. In *ELEX 2019: SMART LEXICOGRAPHY*. Lexical Computing CZ SRO, pp. 374–396.

Chamberlain J., **Poesio M. and Kruschwitz U.** (2008). Addressing the resource bottleneck to create large-scale annotated texts. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*. College Publications, pp. 375–380.

Chklovski T. (2005). Collecting paraphrase corpora from volunteer contributors. In *Proceedings of the 3rd International Conference on Knowledge Capture*, K-CAP'05, New York, NY, USA. Association for Computing Machinery, pp. 115–120.

Constant M., **EryiĞit G.**, **Monti J.**, **van der Plas L.**, **Ramisch C.**, **Rosner M. and Todirascu A.** (2017). Survey: Multiword expression processing: A Survey. *Computational Linguistics* **43**(4), 837–892.

Cook P., **Fazly A. and Stevenson S.** (2008). The vnc-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 19–22.

Dumitrache A., **Aroyo L.**, **Welty C.**, **Sips R.-J. and Levas A.** (2013). "dr. detective": Combining gamication techniques and crowdsourcing to create a gold standard in medical text. In *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web - Volume 1030*, CrowdSem'13, Aachen, DEU. CEUR-WS.org, pp. 16–31.

Ehren R., **Lichte T.**, **Kallmeyer L. and Waszczuk J.** (2020). Supervised disambiguation of German verbal idioms with a BiLSTM architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, Online. Association for Computational Linguistics, pp. 211–220.

Fort K., **Adda G. and Cohen K.B.** (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics* **37**(2), 413–420.

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)*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 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*, Marseille, France. European Language Resources Association, pp. 4395–4401.

Geiger D. and Schader M. (2014). Personalized task recommendation in crowdsourcing information systems – current state of the art. *Decision Support Systems* **65**, 3–16. Crowdsourcing and Social Networks Analysis.

Grace Araneta M., **Eryigit G.**, **König A.**, **Lee J.-U.**, **Luis A.R.**, **Lyding V.**, **Nicolas L.**, **Rodosthenous C. and Sangati F.** (2020). Substituto - A synchronous educational language game for simultaneous teaching and crowdsourcing. In *9th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2020)*, Gothenburg, Sweden, pp. 1–9.

Hashimoto C. and Kawahara D. (2009). Compilation of an idiom example database for supervised idiom identification. *Language Resources and Evaluation* **43**(4), 355.

Howe J. (2006). The rise of crowdsourcing. *Wired Magazine* **14**(6), 1–4.

Kaschak M.P. and Saffran J.R. (2006). Idiomatic syntactic constructions and language learning. *Cognitive Science* **30**(1), 43–63.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kim T.W. and Werbach K. (2016). More than just a game: Ethical issues in gamification. *Ethics and Information Technology* **18**(2), 157–173.

Klock A.C.T., **Gasparini I.**, **Pimenta M.S. and Hamari J.** (2020). Tailored gamification: A review of literature. *International Journal of Human-Computer Studies* **144**, 102495.

Konopka A.E. and Bock K. (2009). Lexical or syntactic control of sentence formulation? structural generalizations from idiom production. *Cognitive Psychology* **58**(1), 68–101.

Lawson N., **Eustice K.**, **Perkowitz M. and Yetisgen-Yildiz M.** (2010). Annotating large email datasets for named entity recognition with cechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles. Association for Computational Linguistics, pp. 71–79.

Loper E. and Bird S. (2002). NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics, pp. 63–70.

Losnegaard G.S., **Sangati F.**, **Escartn 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)*, PortoroŽ, Slovenia. European Language Resources Association (ELRA), pp. 2299–2306.

MitroviĆ J. (2013). Crowdsourcing and its application. *INFOtheca* **14**(1), 37–46.

Moon R. (2015). Idioms: A view from the web. *International Journal of Lexicography* **28**(3), 318–337.

Morschheuser B. and Hamari J. (2019). The gamification of work: Lessons from crowdsourcing. *Journal of Management Inquiry* **28**(2), 145–148.

Morschheuser B., **Hamari J.**, **Koivisto J. and Maedche A.** (2017). Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies* **106**, 26–43.

Morschheuser B., **Hamari J. and Maedche A.** (2019). Cooperation or competition – when do people contribute more? a field experiment on gamification of crowdsourcing. *International Journal of Human-Computer Studies* 127, 7–24. Strengthening gamification studies: critical challenges and new opportunities.

Morschheuser B., **Hassan L.**, **Werder K. and Hamari J.** (2018). How to design gamification? a method for engineering gamified software. *Information and Software Technology* **95**, 219–237.

Murillo-Zamorano L.R., **ángel López Sánchez J. and Bueno Muñoz C.** (2020). Gamified crowdsourcing in higher education: A theoretical framework and a case study. *Thinking Skills and Creativity* **36**, 100645.

Palmero Aprosio A. and Moretti G. (2016). Italy goes to Stanford: a collection of CoreNLP modules for Italian. arXiv e-prints, arXiv:1609.06204.

PrpiĆ J., **Shukla P.P.**, **Kietzmann J.H. and McCarthy**, **I.P.** (2015). How to work a crowd: Developing crowd capital through crowdsourcing. *Business Horizons* **58**(1), 77–85.

Qi P., **Zhang Y.**, **Zhang Y.**, **Bolton J. and Manning C.D.** (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online. Association for Computational Linguistics, pp. 101–108.

Quoc Viet Hung N., **Tam N.T.**, **Tran L.N. and Aberer K.** (2013). An evaluation of aggregation techniques in crowdsourcing. In **Lin X.**, **Manolopoulos Y.**, **Srivastava D. and Huang G.** (eds), *Web Information Systems Engineering – WISE 2013*, Berlin, Heidelberg: Springer, pp. 1–15.

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 Escartn 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)*, Santa Fe, New Mexico, USA. Association for Computational Linguistics, pp. 222–240.

Rumshisky A., **Botchan N.**, **Kushkuley S. and Pustejovsky J.** (2012). Word sense inventories by non-experts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA), pp. 4055–4059.

Savary A., **Candito M.**, **Mititelu V.B.**, **BejČek E.**, **Cap F.**, **Čéplö**, **S.**, **Cordeiro S.R.**, **EryiĞit G.**, **Giouli V.**, **van Gompel**, **M.**, **HaCohen-Kerner Y.**, **KovalevskaitĖ J.**, **Krek S.**, **Liebeskind C.**, **Monti**, **J.**, **Escartín C.P.**, **van der Plas L.**, **QasemiZadeh B.**, **Ramisch C.**, **Sangati F.**, **Stoyanova I. and Vincze V.** (2018). PARSEME multilingual corpus of verbal multiword expressions. In **Markantonatou S.**, Ramisch C., **Savary A.** and Vincze V. (eds), *Multiword Expressions at Length and in Depth: Extended Papers from the MWE 2017 Workshop*. Berlin: Language Science Press, pp. 87–147.

Saxena P. and Paul S. (2020). Epie dataset: A corpus for possible idiomatic expressions. In **Sojka**, **P.**, **KopeČek**, **I.**, **Pala**, **K. and Horák**, **A.** (eds), *Text, Speech, and Dialogue*, Cham: Springer International Publishing, pp. 87–94.

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)*, Reykjavik, Iceland. European Language Resources Association (ELRA), pp. 455–461.

Siyanova-Chanturia A. (2017). Researching the teaching and learning of multi-word expressions. *Language Teaching Research* **21**(3), 289–297.

Snow R., **O'Connor B.**, **Jurafsky D. and Ng A.** (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*, Honolulu, Hawaii. Association for Computational Linguistics, pp. 254–263.

Sprenger S.A., **Levelt W.J. and Kempen G.** (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language* **54**(2), 161–184.

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

Vasiljevic Z. (2015). Teaching and learning idioms in l2: From theory to practice. *Mextesol Journal* **39**(4), 1–24.
Vincze V., Nagy T. I. and Berend G. (2011). Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, Hissar, Bulgaria. Association for Computational Linguistics, pp. 289–295.
von Ahn L. (2006). Games with a purpose. *Computer* **39**(6), 92–94.
von Ahn L. and Dabbish L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'04, New York, NY, USA. Association for Computing Machinery, pp. 319–326.
von Ahn L., Kedia M. and Blum M. (2006). Verbosity: A game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'06, New York, NY, USA. Association for Computing Machinery, pp. 75–78.
YirmibeŞoĞlu Z. and Güngör T. (2020). ERMI at PARSEME shared task 2020: Embedding-rich multiword expression identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, online. Association for Computational Linguistics, pp. 130–135.

A. Appendix

As stated in the introduction, deducting well-defined rules to express an idiom is usually a challenging task. Below we provide an example idiom from Turkish, with an English explanation of its meaning and usage patterns (for an English-speaking second-language learner of Turkish).

Example idiom: "(birinin) baŞının etini yemek"

Table A1. Design principles for engineering gamified software (Morschheuser *et al.* 2018)

Design principle	Meaning
DP#1	Understand the user needs, motivation and behavior, as well as the characteristics of the context
DP#2	Identify project objectives and define them clearly
DP#3	Test gamification design ideas as early as possible
DP#4	Follow an iterative design process
DP#5	Profound knowledge in game-design and human psychology
DP#6	Assess if gamification is the right choice to achieve the objectives
DP#7	Stakeholders and organizations must understand and support gamification
DP#8	Focus on user needs during the ideation phase
DP#9	Define and use metrics for the evaluation and monitoring of the success, as well as the psychological and behavioral effects of a gamification approach
DP#10	Control for cheating/gaming-the-system
DP#11	Manage and monitor to continuously optimize the gamification design
DP#12	Consider legal and ethical constraints in the design phase
DP#13	Involve users in the ideation and design phase

https://doi.org/10.1017/S1351324921000401 Published online by Cambridge University Press

Table A2. Idioms of the TrP1 (first 16 rows) & TrP2 (last 16 rows) – (Id.:idiomatic samples, NonId.:nonidiomatic samples, :dislikes)

Day	Idiom	Literal meaning	Idiomatic meaning	# of collected samples Id.	NonId.	Total	# of Parseme Id.	Avg. # of Rewiews	% of
1	hesap vermek	bill - to give	to explain the reason for any behavior	198	212	410	5	4.6	7
2	altını çizmek	to underline	to emphasize	83	83	166	5	4.8	8
3	yer vermek	place - to give	to emphasize the importance of sth	87	73	160	23	3.9	5
4	ayvayı yemek	to eat a quince	to get in a bad situation	87	92	179	0	5.5	4
5	rol oynamak	to act	to play an important role in sth	45	43	88	10	5.2	6
6	üzerinde durmak	on top - to stand	to emphasize	69	51	120	10	4.5	4
7	ağırlık vermek	weight - to give	to emphasize	49	48	97	8	3.8	5
8	yer almak	place - to take/buy	to occur	69	49	118	132	3.5	8
9	kolları sıvamak	arms - to roll up	to get ready to do sth difficult	34	36	70	7	3.7	8
10	öne sürmek	to put/drive to the front	to suggest	38	24	62	48	3.3	8
11	ortaya koymak	to the middle - to put	to introduce/to put forward	59	53	112	43	2.8	25
12	iz bırakmak	a mark - to leave	to place in one's mind	40	48	88	2	2.8	18
13	ele almak	to the hand - to take	to handle	37	29	66	39	1.5	3
14	yol açmak	a road - to open	to cause	21	20	41	38	2.2	6
15	meydana gelmek	to the center - to come	to happen	11	18	29	29	2.2	2
16	karşı çıkmak	in front of - to climb/to step up	to oppose	14	7	21	27	1.2	8
1	içi erimek	its inside - to melt	to worry/to be upset	121	149	270	0	5.5	19
2	el açmak	hand - to open	to beg	206	211	417	0	3.0	17

Table A2. Continued

Day	Idiom	Literal meaning	Idiomatic meaning	# of collected samples Id.	NonId.	Total	# of Parseme Id.	Avg. # of Rewiews	% of
3	zaman kazanmak	time - to earn	to save time	155	193	348	0	3.1	5
4	defterden silmek	to erase from notebook	to forget someone	97	99	196	1	3.9	28
5	nabzını tutmak	to hold pulse	to measure intension	58	55	113	2	4.0	7
6	basamak yapmak	step - to make	to exploit someone	79	95	174	0	4.2	26
7	baŞa geçmek	to the head - to pass	to govern	68	58	126	1	3.5	2
8	sıkı durmak	to stay/to look tight	to stay strong or to be ready	201	173	374	0	2.7	36
9	üste çıkmak	to the top - to climb/to step up	to blame others even though being guilty	85	176	261	0	2.9	26
10	sayıp dökmek	to count and to pour	to tell everything	74	133	207	0	3.3	10
11	üstünden atmak	from over to throw	to get rid of	65	80	145	0	4.7	18
12	zaman öldürmek	time - to kill	to waste time	123	127	250	1	3.7	14
13	üstüne almak	onto - to take	to undertake	113	300	413	1	4.0	3
14	parmak basmak	finger - to press	to attract attention on sth	54	89	143	1	4.3	6
15	meydana gelmek	to come to the center	to happen	103	92	195	29	3.6	29
16	karşı çıkmak	to climb/step up opposite	to oppose	167	352	519	27	4.1	6

Table A3. Idioms of the ItP1 (first 16 rows) & ItP2 (last 16 rows) – (Id.:idiomatic samples, NonId.:nonidiomatic samples, :dislikes)

Day	Idiom	Literal meaning	Idiomatic meaning	# of collected samples Id.	NonId.	Total	# of Parseme Id.	Avg. # of Rewiews	% of
1	gettare la spugna	to throw the sponge	to throw in the towel	470	302	772	0	3.6	29
2	coltivare il proprio orto	to cultivate one's vegetable garden	to care only about one's problems	61	79	140	0	3.6	42
3	buttare giu	to throw down	to swallow, to overthrow, to push over	43	41	84	0	3.6	32
4	mettere dentro	to put inside	to put in jail	75	130	205	0	3.1	40
5	abbaiare alla luna	to bark to the moon	to bark at the moon, to swear	58	23	81	0	3.4	33
6	acchiappare farfalle	to catch butterflies	to do useless things	21	8	29	0	1.9	35
7	ingoiare una pillola	to swallow a pill	to subject oneself to something unpleasant	7	3	10	0	1.7	6
8	ammainare le vele	to furl the sails	to abandon, to surrender	10	1	11	0	0.4	25
9	andare a gonfie vele	to go with inflated sails	to be successful	23	4	27	0	3.3	6
10	andare in barca	to go in boat	to break down	14	8	22	0	2.1	2
11	aprire gli occhi	to open the eyes	to awaken, to realize	2	1	3	2	0.0	0
12	attaccare bottone	to attach button	to chat up, to talk endlessly	3	3	6	0	0.0	0
13	avere la coda di paglia	to have the tail of straw	to feel guilty	27	10	37	0	1.8	1
14	avere la corda al collo	to have the rope at the neck	to not have control	9	0	9	0	0.4	25
15	avere le mani lunghe	to have the hands long	to steal	4	2	6	0	0.7	0
16	avere birra in corpo	to have beer in body	to have strength	1	0	1	0	0.0	0
1	avere il becco lungo	to have the beak long	to speak outright	185	98	283	0	3.9	19
2	avere il mestolo in mano	to have the ladle in hand	to rule despotically	277	421	698	0	1.6	24

Table A3. Continued.

Day	Idiom	Literal meaning	Idiomatic meaning	# of collected samples Id.	NonId.	Total	# of Parseme Id.	Avg. # of Rewiews	% of
3	prendere con le pinze	to take with the pincers	to take it with a pinch of salt	409	394	803	1	0.8	32
4	raggiungere il bersaglio	to reachthe target	to reach the objective	285	89	374	0	1.5	23
5	buttare al vento	to throw to the wind	to fritter away	188	162	350	0	0.7	35
6	brancolare nel buio	to grope in the dark	to grope in the dark	334	246	580	0	1.0	49
7	attaccare bottone	to attach button	to talk endlessly	53	68	121	0	2.7	8
8	avere le batterie scariche	to have the batteries dead	to be exhausted	156	101	257	0	2.0	34
9	aprire gli occhi	to open the eyes	to awaken, to realize	143	132	275	2	1.0	22
10	portare a casa	to take home	to earn	113	75	188	2	2.6	17
11	tirare su	to pull up	to raise	135	64	199	0	1.3	44
12	essere tra i piedi	to be among the feet	to get in sb's way	152	32	184	2	2.2	17
13	dare corda	to give cord	to give sb a free hand	339	130	469	0	1.8	35
14	mandare a casa	to send at house	to send away, to dispatch, to kick out	102	137	239	2	2.4	6
15	avere le mani lunghe	to have the hand long	to steal	115	32	147	0	2.6	15
16	avere birra in corpo	to have beer in body	to have strength	83	35	118	0	4.0	6

Literal meaning: ≪*eat (someone's) head's meat*≫
Idiomatic meaning: "*annoying someone by talking too much*" as "*to nag at*"

Rule#1: *someone's* may be replaced with one of the possessive pronouns (e.g., my, your, his) or any noun taking a possessive suffix -'s (i.e., the genitive suffix in the target language).

Rule#2: *someone's* may be omitted since the target language is a pro-drop language and the word *head* also takes possessive suffixes which also carry the person agreement information thus *someone* is pragmatically or grammatically inferable.

Rule#3: the verb *eat* may be inflected

Rule#4: since this language is an MRL and pro-drop language, the inflected verb will also carry the person agreement information thus the subject information coming with the verb (or either separately) should be different than *someone*, that is reflexive usage is generally not welcome; "eating own's head's meat".

As one may notice, although it could be possible to define rules, they are both hard to deduct (e.g., for teachers or lexicographers) and hard to understand (for language learners: humans or computers). Language learners will still need usage examples both to understand the usage patterns and to practice. Additionally, for being able to define such rules even teachers or lexicographers should investigate many usage samples or come up with new ones.

Cite this article: Eryiğit G, Şentaş A and Monti J. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering* https://doi.org/10.1017/S1351324921000401