Dodiom: a Gamified Bot supporting Diversity and Multilinguality for Idiom Corpora Construction

Gülşen Eryiğit and **Ali Şentaş**
ITU NLP - Dep. of
Artificial Intelligence & Data Eng.,
Istanbul Technical University
gulsen.cebiroglu@itu.edu.tr
sentasa19@itu.edu.tr

Johanna Monti
UNIOR NLP - Dep. of
Literary, Linguistic and Comparative
Studies, University of Naples L'Orientale
jmonti@unior.it

Idiomatic expressions are one the linguistic phenomena with few or rare occurrences within text. That is why traditional methods focusing on their annotations within actual text or their retrieval from existing resources suffer from the data scarcity problem. Recently, a new approach called "gamified crowdsourcing for idiom corpora construction" has been published and reported to provide a solution to this problem by collecting the data via a crowd-creating & crowd-rating approach. The approach was implemented as a Telegram bot and tested on Italian, Turkish, and Russian so far. The tool, being easily adaptable to new languages and new phenomena, is seen as a good means of supporting inter- and intra-language diversity.

Relevant UniDive working groups: WG1

1 Introduction

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", "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. 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. Recently, Eryiğit et al. (2022) introduced a gamified crowdsourcing approach for collecting language learning materials for idiomatic expressions; a messaging bot 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.

This paper describes a recently published work and aims to introduce the Dodiom bot to the Unidive Community as a means of supporting inter- and intra-language diversity.

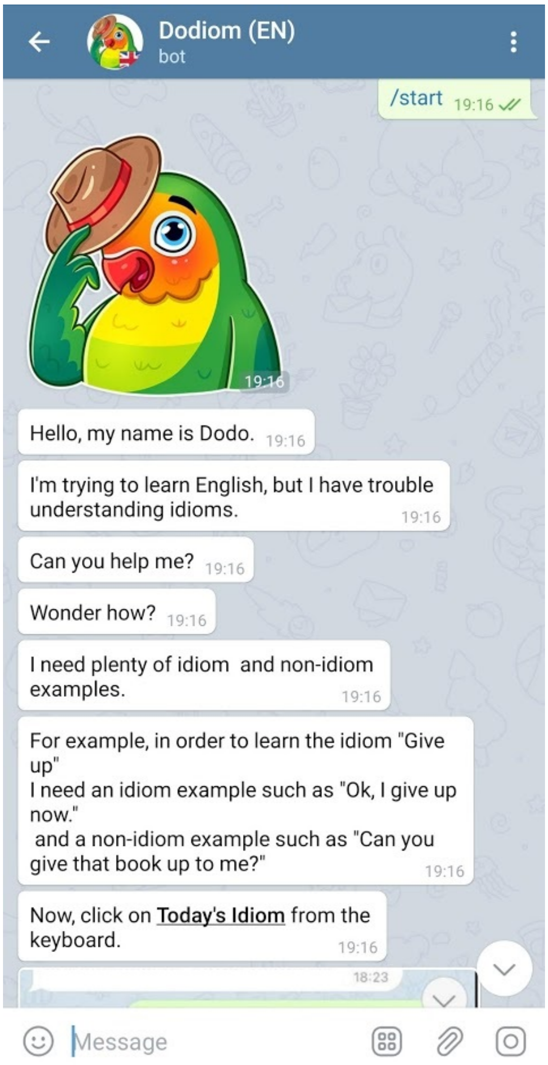
2 Dodiom

The approach is language independent and Eryiğit et al. (2022) evaluated it on two languages (i.e., Turkish and Italian) 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 revealed 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. The source codes of the developped Telegram bot is publicly available via a Github page[1]

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

An online presentation of the study is also available from YouTube[2].

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. Some screenshots from this application is provided in Figure 1 and Figure 2.



/start 19:16

Hello, my name is Dodo. 19:16

I'm trying to learn English, but I have trouble understanding idioms. 19:16

Can you help me? 19:16

Wonder how? 19:16

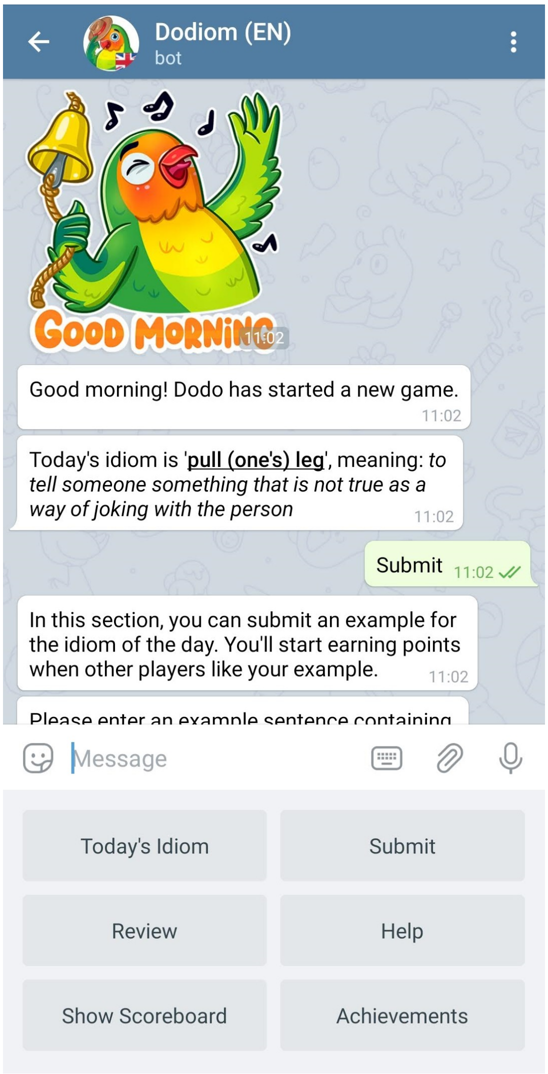
I need plenty of idiom and non-idiom examples. 19:16

For example, in order to learn the idiom "Give up"
I need an idiom example such as "Ok, I give up now."
and a non-idiom example such as "Can you give that book up to me?" 19:16

Now, click on **Today's Idiom** from the keyboard. 19:16

18:23

Figure 1: Dodo greeting the player, describing the game, and showing the next steps



Good morning! Dodo has started a new game. 11:02

Today's idiom is '**pull (one's) leg**', meaning: *to tell someone something that is not true as a way of joking with the person* 11:02

Submit 11:02

In this section, you can submit an example for the idiom of the day. You'll start earning points when other players like your example. 11:02

Please enter an example sentence containing

Figure 2: Main Menu showing the currently available options

Design principles from Morschheuser et al. (2018) were applied during the design phase of Dodiom. The game is designed with localization in mind. All the interaction messages are localized and shown to the users in the related language; localizations are currently available for English, Italian, Turkish, and Russian languages. 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. It is also foreseen 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[3] containers adjusted to each country's time-zone where the game is played.

[2] https://www.youtube.com/watch?v=4z3PV6EAWqQ

[3] https://docker.com/

References

Gülşen Eryiğit, Ali Şentaş, and Johanna Monti. 2022. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, page 1–33.

Benedikt Morschheuser, Lobna Hassan, Karl Werder, and Juho Hamari. 2018. How to design gamification? a method for engineering gamified software. *Information and Software Technology*, 95:219–237.