Gamified Crowd-sourcing for Word Sense Disambiguation of Turkish

DILARA TORUNOĞLU SELAMET, Computer Engineering, Istanbul Technical University, Istanbul, Turkey
ALI ŞENTAŞ, Computer Engineering, Istanbul Technical University, Istanbul, Turkey
GÜLŞEN ERYIĞIT, AI & Data Engineering, Istanbul Technical University, Istanbul, Turkey

Word sense disambiguation (WSD) is the process of determining the correct meaning of a word based on its context in a sentence, a task that remains one of the core challenges in natural language processing (NLP) despite the advancements made by LLMs. Although LLMs have improved WSD performance, challenges remain–especially in nuanced contexts, low-resource languages, and domain-specific terminology. These gaps can hinder the accuracy of LLMs in high-stakes applications and multilingual settings. In response to the need for WSD, our study introduces a novel approach to WSD data collection through gamified crowdsourcing, which, to our knowledge, has not been previously applied in this field. A messaging bot method has been used to engage a wide, diverse audience to generate high-quality WSD data. Using a multiplayer game format, native speakers provide examples for different senses of ambiguous words and rate others' contributions. Together with the suggested enhancements, this platform attracted a diverse range of participants–spanning ages, backgrounds, and genders–20 times more than a similar crowdsourcing method applied to a different NLP task, and sustained their engagement over an extended period. Unlike conventional academic crowdsourcing pools, this approach focuses on drawing individuals from varied backgrounds beyond academia or AI-focused communities. It not only gathered data but also encouraged participants to evaluate each other's contributions, creating a rich and reliable dataset. Our findings suggest that gamified crowdsourcing can be a powerful tool for creating WSD corpora. Our approach not only supports future WSD research but also contributes valuable training data for developing more precise and resilient language models.

CCS Concepts: • **Computing methodologies → Natural language processing**; • **Human-centered computing → Collaborative and social computing;**

Additional Key Words and Phrases: Crowdsourcing, gamification, game with a purpose (GWAP), word sense disambiguation, language resources

ACM Reference Format:
Dilara Torunoğlu Selamet, Ali Şentaş, and Gülşen Eryiğit. 2025. Gamified Crowd-sourcing for Word Sense Disambiguation of Turkish. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 24, 11, Article 130 (November 2025), 23 pages. https://doi.org/10.1145/3767160

Authors' Contact Information: Dilara Torunoğlu Selamet (corresponding author), Computer Engineering, Istanbul Technical University, Istanbul, Turkey; e-mail: torunoglud@itu.edu.tr; Ali Şentaş, Computer Engineering, Istanbul Technical University, Istanbul, Turkey; e-mail: sentas@itu.edu.tr; Gülşen Eryiğit, AI & Data Engineering, Istanbul Technical University, Istanbul, Turkey; e-mail: gulsenc@itu.edu.tr.

ACM 2375-4699/2025/11-ART130
https://doi.org/10.1145/3767160

1 Introduction

WSDe is the process of determining the correct meaning of a word based on its context in a sentence. Due to the numerous ambiguities in natural languages, researchers suggest addressing the problem across various languages for better disambiguation results [18, 56, 58].

WSD enhances several NLP tasks, such as sentiment analysis, abstract meaning representation, multilingual information extraction, and the semantic web [1, 18, 40]. These tasks require the ability to distinguish subtle differences in word meanings, often relying on implicit context. Humans are particularly skilled at this disambiguation, even with limited information, which suggests that understanding this skill could improve NLP performance. Grasping a word's meaning within a sentence is a core challenge in WSD tasks–not only for AI but even for humans at times–and large language models like ChatGPT still struggle to achieve complete accuracy [9, 37]. The lack of abundant, high-quality labeled data makes this task challenging for both humans and AI [25]. For example, the Turkish word "ocak" has 10 to 11 meanings listed by the **Turkish Language Association** (**TDK**) and WordNet, yet a model like ChatGPT typically only identifies 4 or 5. This reveals the model's limitations in capturing fine-grained distinctions, raising the question of whether such precise interpretation is desirable in AI–or, if so, how confidently we can rely on its interpretations.

While LLMs have made strides in addressing this issue, WSD is still essential for several key reasons; LLMs often struggle in nuanced or ambiguous contexts, particularly in specialized domains and low-resource languages. WSD provides targeted solutions for such challenges by ensuring accurate word interpretation across various applications. It is indispensable in fields like medicine and law, where precise terminology is critical, and in multilingual tasks such as translation and summarization, where cross-linguistic clarity is essential [12, 33, 40]. Additionally, WSD addresses rare word senses that are underrepresented in large datasets, enhancing performance in low-resource scenarios [6, 36]. By reducing misinterpretations in sensitive applications and boosting efficiency in tasks like information retrieval and text simplification, WSD complements LLMs in addressing ambiguities they may not fully resolve, thereby improving overall NLP robustness [5, 46, 55].

This study aims to tackle several challenges by proposing a novel approach to collecting **Word Sense Disambiguation** (**WSD**) data, with a particular focus on low-resource languages like Turkish, where labeled data is both scarce and difficult to produce at scale. One of the primary obstacles in this area is the development of diverse, high-quality, and sense-rich corpora that accurately represent real-world language use. Traditional annotation methods are time-consuming, resource-intensive, and costly, which restricts the size and coverage of datasets needed for supervised learning. Additionally, the crowdsourcing pools available on existing online platforms often feature a limited and homogeneous participant base, which hampers the linguistic and contextual diversity essential for training robust and generalizable models.

The novelty of this work lies in its dual mechanism of crowd-driven creation and evaluation within a gamified environment, advancing beyond conventional labeling methods to produce richly contextualized examples and corresponding quality assessments at scale. Recent advancements have demonstrated that gamified crowdsourcing offers an innovative and effective method for gathering supervised training data [11, 15–17, 21, 30, 33, 35, 48, 59]. In this study, we apply this novel approach to the collection of data for WSD–a core NLP task for which, to our knowledge, this method has not previously been utilized in the literature. Our findings indicate that gamified crowdsourcing is indeed a valuable tool for WSD data collection, showing promising efficacy in this new application. This article introduces a gamified crowdsourcing approach for collecting WSD examples for the words with multiple meanings; A messaging bot is structured as an asynchronous multiplayer game where native speakers compete by providing various examples for different meanings of a given ambiguous word and rating other players' submissions. Our work follows

traditional crowd-processing annotation efforts in the field; however, for the first time in the literature, a crowd-creating and crowd-rating approach [24, 43, 53] is implemented and tested specifically for WSD data collection. A similar approach has been undertaken by [17] for collecting idiomatic expressions in corpora. This article extends the previous work with a re-design and adaptation for the WSD task together with modifications for the game to be used in a wider crowd sourcing environment (i.e., Instagram). In this study, our approach involved actively engaging the public in WSD data generation through a gamified platform promoted on social media (Instagram), reaching 20 times more participants over a longer period than traditional crowdsourcing methods. Unlike conventional academic crowdsourcing pools, our participants spanned diverse ages and backgrounds, with a balanced mix of genders, providing data for 30 days. Participants not only generated data but also rated each other's contributions, creating a rich and reliable dataset. The approach is tested on Turkish language with 30 different ambiguous words (15 of them nouns, 15 of them verbs) and evaluated in comparison to traditional data preparation techniques in the field. With rewards as an incentive, this gamified approach proved both engaging and effective, as users found it entertaining and useful.

This study introduces gamified crowdsourcing as a novel approach to constructing WSD corpora, addressing key challenges such as data scarcity and the imbalance between ambiguous and unambiguous word samples in traditional methods. Our contributions to the field are threefold: (1) the first crowd-creating and crowd-rating application of gamified crowdsourcing for WSD in the literature, (2) the creation of a new WSD dataset for Turkish, and (3) the development of a diverse dataset with potential applicability for multilingual WSD research. This method not only accelerates the creation of WSD corpora but also provides high-quality supervised training data, supporting future WSD studies and the advancement of robust and precise language models. The article is structured as follows: Section 2 discusses related work, Section 3 details the game design, Section 4 presents analyses, and Section 5 concludes.

2 Background and Related Work

WSD is one of the most significant challenges in the field of NLP [4, 46]. It involves determining which meaning of a word is being used in a given context, a task that is inherently complex due to the polysemous nature of language. Words often have multiple meanings, and discerning the correct one requires a deep understanding of the surrounding text, as well as the nuances of language use. Creating high-quality, manually annotated corpora with accurate sense labels is often both time-consuming and resource-intensive [18, 40]. As a result, nearly all widely used sense-tagged corpora are produced by trained annotators [26, 52], creating a knowledge acquisition bottleneck for training systems that rely on sense labels [23]. Despite advances in machine learning and AI, achieving high accuracy in WSD remains difficult, as it necessitates large, manually annotated corpora for training, which are costly and laborious to produce [25]. These foundational studies underline the limitations of traditional supervised approaches, particularly in terms of scalability and cost.

In the context of low-resource languages, WSD becomes even more challenging due to the scarcity of annotated corpora and tools. Recent studies have explored various methods to improve WSD performance in these contexts. In the study [60], researchers focused on RegNet with efficient channel attention and dilated convolution. In another study [14], researchers applied a specific WSD method for the Portuguese language, while in [22], researchers introduced a hybrid approach tailored for Marathi. Other approaches include [27], where researchers developed a supervised method for clinical abbreviation normalization, and [34], where researchers focused on WSD for large documents using neural networks. Additionally, Ref. [8] presented a novel method for morphologically rich, low-resource languages. These approaches demonstrate varied strategies such

as hybrid modeling, domain-specific adaptation, and neural architectures, each with advantages like flexibility or domain-fit, but also with limitations stemming from data scarcity or computational complexity.

To overcome the data bottleneck, crowdsourcing has emerged as an alternative means of collecting linguistic annotations. It has been increasingly applied in various NLP tasks to gather sense annotations. Recent studies have proposed crowdsourcing as an effective method for collecting sense annotations, highlighting its potential to mitigate the data scarcity problem in WSD [7, 13, 31, 51, 54, 57]. However, the sense labels generated through these methods often differ from those found in widely used sense inventories such as WordNet [19]. While crowdsourcing enables broader participation and scalable annotation, it also introduces challenges in annotation consistency and sense inventory alignment. Due to these issues, there are relatively few studies introducing word sense corpora.

Building upon crowdsourcing, gamified approaches have recently gained traction for data collection tasks in NLP and beyond. Recent advancements show that gamified crowdsourcing is an innovative and effective way to collect supervised training data across diverse fields [42, 43, 45]. Researchers in [48] used gamified close-range sensing for developing a Digital Twin of the Earth, while in [15], researchers applied this approach to label lung ultrasound datasets. In [33], researchers explored lexical diversity through crowdsourcing, and [30] constructed a large English-Croatian parallel corpus via a gamified platform. Other notable studies include [16], where researchers examined gamification's role in crowdsourcing; [21], where researchers surveyed perspectivist approaches in NLP; and [11], where researchers addressed lexical ambiguity in natural language requirements using Gamify4LexAmb. In [59], researchers leveraged gamified crowdsourcing for high-quality visual fine-tuning data, while [35] supported the semantic transformation of digital libraries. In [17], researchers specifically applied gamified crowdsourcing for constructing idiom corpora, further illustrating the potential of this approach across language and data collection tasks. While these studies emphasize the scalability, engagement, and creativity of gamified methods, none have targeted the creation of word sense corpora through this paradigm.

In this context, our study contributes a novel direction by applying gamified crowdsourcing for word sense annotation in Turkish. Unlike prior work that either used traditional crowdsourcing or targeted different annotation goals (e.g., idioms, visual data), our platform leverages both crowd-creating and crowd-rating mechanisms tailored to word sense tasks. This work belongs to the intersection of low-resource WSD and gamified crowdsourcing, aiming to fill the gap where no prior studies have focused on gamified WSD corpus creation in Turkish or other low-resource settings. Our approach proposes a scalable, engaging, and linguistically aligned methodology to construct high-quality sense-annotated corpora.

In addition to the studies mentioned above, several others have focused specifically on Turkish WSD. Ref. [28] constructed a Turkish lexical sample dataset by identifying highly ambiguous words and manually annotating them with expert taggers. Ref. [41] proposed adaptations of Lesk and simplified Lesk algorithms for Turkish, incorporating stemming and polysemy-aware mechanisms. Ref. [2] translated English Penn Treebank sentences into Turkish and applied various feature-based classifiers. Ref. [49] explored ways to induce sense distinctions from scratch, such as hierarchical binary splits and pseudo-word generation. Ref. [3] introduced an all-words WSD corpus by aligning a Turkish Treebank with TDK dictionary senses and tested increasingly complex features for disambiguation. Ref. [32] improved the Lesk algorithm by taking advantage of the hierarchical structure of Turkish WordNet. While previous studies have contributed valuable resources and techniques for Turkish WSD, they present key limitations in terms of accessibility, representativeness, and design.

Table 1. Categorized Overview of Prior Work on Word Sense Disambiguation

Category	Representative Studies
General WSD Methods and Challenges	[4, 18, 23, 25, 26, 40, 46, 52]
WSD in Low-Resource Languages	[8, 14, 22, 27, 34, 60]
Crowdsourcing for WSD	[7, 13, 31, 51, 54, 57]
Gamified Crowdsourcing (Non-WSD)	[11, 15–17, 21, 30, 33, 35, 42, 43, 45, 48, 59]
WSD Studies in Turkish	[2, 3, 28, 32, 41, 49]

Among the four that offer linguistic datasets, Ref. [28] is closed-source, limiting its usability by the broader community. Another dataset [2] is not originally Turkish, which limits its applicability for natural Turkish ambiguity. The remaining two [3, 32] theses are not specifically designed to represent distinct senses of ambiguous words, making them suboptimal for training or evaluation. Overall, the lack of high-quality, open, and purpose-built datasets limits progress in Turkish WSD. Our work provides a freely available, native, and sense-focused dataset created via gamified crowdsourcing, offering both scalability and user engagement. A summary of the categorized literature discussed above is presented in Table 1 for a clearer comparison of methodologies and domains.

3 Game Design

The name of the original game in [17] was "Dodiom" for expressing that it is designed for idiom corpora collection; the merging of the words Dodo (the game's bird character representing the bot's persona) and idiom. We changed the telegram bots name to "DodoME" which reflects the word MEanings for WSD. According to researchers [44], the users increasingly expect software to be not only functional but also enjoyable to use, and creating gamified software necessitates an in-depth understanding of motivational psychology and multidisciplinary knowledge. In our case, these disciplines include computational linguistics, NLP, and gamification. To illuminate the successful design of gamified software, the aforementioned study divides the engineering process into seven main phases and outlines thirteen design principles adopted by experts. The complete list of design principles used in Dodiom is provided in [17].

The software was designed to create an enjoyable and cooperative environment that would motivate volunteers to assist with research studies. The game is re-designed to collect usage samples for the WSD task. A messaging bot is designed as an asynchronous multiplayer game[1] for native speakers who compete with each other while providing word senses for given ambiguous words and rating other players' entries. The game is an explicit crowdsourcing game, and players are informed from the outset that by playing, they are contributing to the creation of a public data source.

The story of the game is based on Dodo trying to learn a foreign language and having difficulty learning ambiguous words in that language. Players try to teach Dodo the word senses in that language by providing examples. Dodiom has been developed as an opensource project (available on Github [2]) with the focus on being easily adapted to different languages or different tasks such as in our case WSD. Similarly, we also share the DodoME as an open project,[3] which will open the way for further research in WSD studies in other languages.

[1]Asynchronous multiplayer games enable players to take their turns at a time that suits them; i.e., the users do not need to be in the game simultaneously.
[2]https://github.com/Dodiom/dodiom
[3]to be shared upon acceptence

Table 2. Chosen Ambiguous Words and English Translations with the Most Common Sense

	Word	UNIQUE MEANING COUNT WordNet	TDK	Mapped		Word	UNIQUE MEANING COUNT WordNet	TDK	Mapped
	Açık (Open)	16	16	4		Aç (Open)	28	27	6
	Baskı (Pressure)	8	8	3		Al (Take)	38	33	4
	Baş (Head)	15	15	4		At (Throw)	38	33	4
	Derece (Degree)	8	7	3		Bak (Look)	20	17	3
	Dünya (World)	5	7	4		Çevir (Turn)	15	15	3
	El (Hand)	11	10	3		Çık (Exit)	59	56	3
	Göz (Eye)	14	13	4		Geç (Pass)	39	38	3
NOUN	Hat (Line)	8	9	3	**VERB**	Gel (Come)	39	36	3
	Hava (Air)	14	14	3		Gir (Enter)	21	19	3
	Kaynak (Resource)	8	8	4		Gör (See)	20	20	2
	Kök (Root)	10	12	4		Kal (Stay)	22	21	3
	Kör (Blindness)	7	7	3		Ol (Be)	24	25	3
	Ocak (Stove)	10	13	4		Sür (Drive)	17	16	4
	Yaş (Age)	8	9	5		Ver (Give)	23	22	3
	Yüz (Face)	14	15	3		Yap (Do)	21	20	3

3.1 Main Interactions, Gameplay

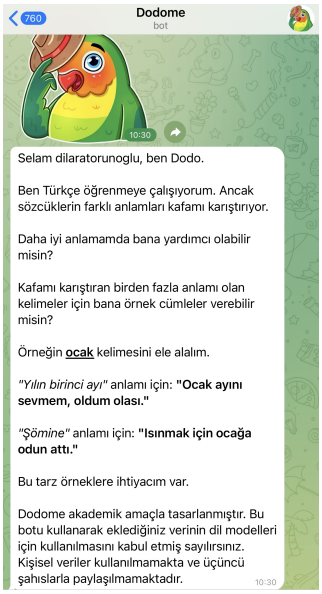
The words to be played each day are selected according to the study [29] according to their tendency suitability to be referred as ambiguous word. The list of played nouns and verbs and their numbers of different meanings are given in Table 2. The table lists ambiguous words, with the most common meanings provided in parentheses based on definitions by the TDK. However, these meanings do not encompass all possible senses of the words. However, we downsized these word meanings with limiting each word to a maximum of four or five meanings. The downsizing process and its reasons are thoroughly explained in Section 3.4.

Dodo learns a new ambiguous word every day. For each chosen word, players have a predetermined time frame to submit their samples and reviews so that they can play at their own pace. DodoMe's time frame is determined as between 9 a.m. and 11 p.m. based on the working hours of our crowd.

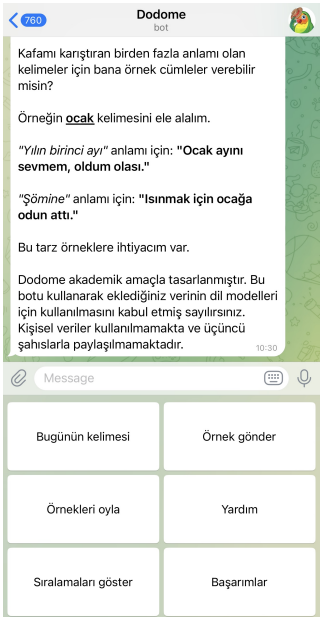
When users connect to the bot for the first time, they are greeted by Dodo, who explains the game and teaches them how to play step-by-step (Figure 1(a)). This pregame tutorial and the simplicity of the game have proven effective, as most players were able to start playing within minutes and provided high-quality examples. Players are then presented with the word of the day along with its related meanings (see Figure 1). Additionally, random tips for the game are shared with players right after they submit their examples. Figure 1(b) shows the main menu, from where the players can access various modes of the game.

Bugünün Kelimesi (Today's Word) tells the player what that day's chosen ambiguous word is (from Table 2). Later players can then submit different meaning examples for the word to get more points.

Örnek Gönder (Submit) allows players to submit new usage examples for different meanings of a word. When clicked, Dodo presents the first meaning of the word with an example to help the user understand it. The player is then asked to provide their own example sentence for that meaning. The submitted sentence is checked to ensure it contains the relevant words, including lemmas, Turkish characters, and correct diacritics. Players are rewarded when others like their examples, incentivizing high-quality submissions. After submitting an example, Dodo asks if the player wants to continue providing examples for the same meaning or move on to the next. Meanings are presented dynamically in different orders; for instance, while the first player is asked for an example of the first meaning, the second player is asked for the second meaning. If there are multiple meanings and active players, each is initially assigned a different meaning to ensure balanced coverage. When a new player joins, Dodo assigns the meaning with the fewest examples first. Additionally, the system limits each player to five examples per meaning before encouraging them



(a) Dodo greeting the player, describing the game, and showing the next steps



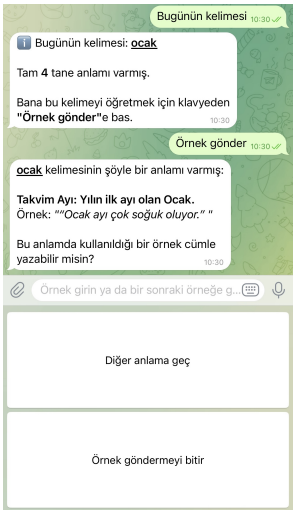
(b) Main Menu showing the currently available options

Fig. 1. Dodiom welcome and menu screens.

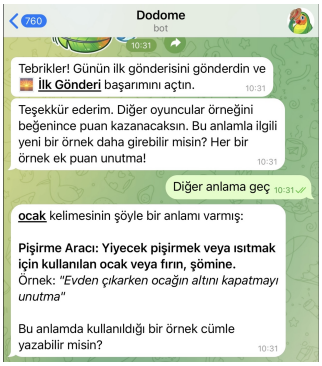
to continue with a new meaning. This approach ensures that all meanings are evenly represented in terms of example sentence count.

In Figure 2, the word "ocak" is used as an example in gameplay. While "ocak" has 10 meanings in WordNet and 11 in TDK, as shown in Table 2, these have been reduced to four primary meanings, a process detailed in Section 3.4. The first player is presented with the first meaning, "the first month of the year" The player can then choose to provide an example for this meaning by selecting "Örnek Gönder" (Submit), or move on to the next meaning by selecting "Diğer anlama geç" (Next Meaning), or stop submitting examples altogether with "Örnek göndermeyi bitir" (Stop). The second meaning presented is "a cooking tool, such as a stove, oven, or fireplace." Third meaning is "aboveground, underground place where minerals are extracted, mine" and forth meaning is "community, family or lineage".

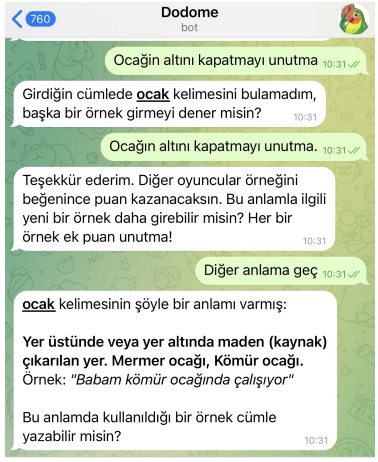
Örnekleri Oyla (Review) lets players evaluate submissions from others. Dodo displays examples along with their annotations (e.g., whether the meaning is correct) and asks for approval. Players earn points for each review, encouraging active participation. An example dialog is given in Figure 3. In Figure 3(a) when the user selects "Örnekleri Oyla" (Review), Dodo prompts them to confirm whether the given example matches the meaning. The user can agree ("Katılıyorum doğru anlam"), disagree ("Katılıyorum yanlış anlam"), report ("Bu örneği şikayet et"), or end the review ("İncelemeyi bitir"). Based on their choice, the user earns points and Dodo thanks them for their contribution as shown in Figure 3(b). Users can also report examples that violate guidelines (e.g., vulgar language, improper use of the platform) for later review by moderators. Moderators can flag inappropriate submissions and ban users if necessary. Submissions with fewer reviews are prioritized, ensuring each example receives a similar number of evaluations. This Review section



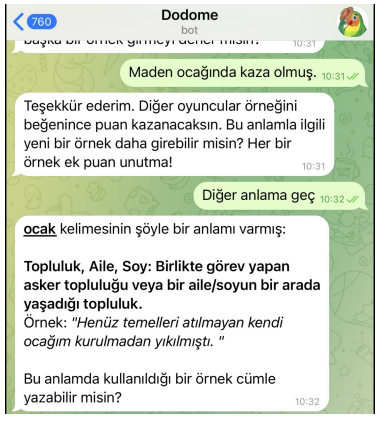
(a) Today's Word and First Meaning



(b) Second Meaning



(c) Third Meaning



(d) Forth Meaning

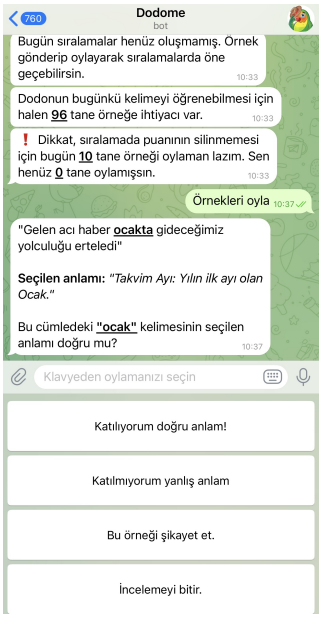
Fig. 2. Meanings of the word "Ocak".

introduces some differences and additions compared with the previous study in [17], which are detailed in Section 3.3.

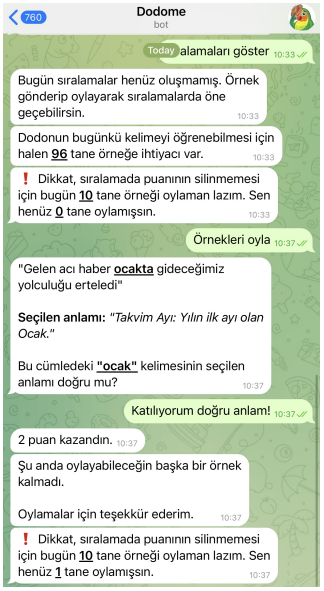
Yardım (Help) option displays a condensed version of the pregame tutorial.

Sıralamaları Göster (Show Scoreboard) shows the current leaderboard, updated whenever a player earns points. As seen in Figure 4(b), the scoreboard shows the scores of the top five players along with the current player's score. It resets daily for each word. For a player's points to count, they must review at least 10 sentences before the game ends at 11 p.m.; otherwise, their score will be reset. Additionally, a soft target of 100 submissions is set, with a message below the scoreboard indicating how many submissions are needed to reach this goal. Dodo gives warning messages to user as seen in Figure 3(a) saying Dodo still needs 96 more examples and user needs to review at least 10 sentences for the points to be counted. Once the target is met, the message disappears. This warnings are also a different approach compared with study in [17] which is explained in detain in Section 3.3.

Başarımlar (Achievements) option shows the score, level, and locked/unlocked achievements of the user. An example can be seen in Figure 4(a) where many achievements are designed to



(a) Dodo asks the player if the given meaning matches the provided example sentence



(b) Player answers Dodo and earns points

Fig. 3. Dodiom review screen.

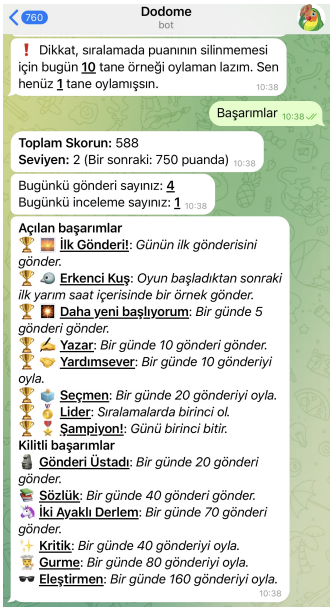
gamify the process and reward players for specific actions such as *Early Bird* achievement for early submissions and *Author* for sending 10 submissions in a given day. Whenever an achievement is obtained, the user is notified with a message and an exciting figure (similar to the ones in Figure 5).

DodoMe employs both gamification affordances and additional incentives [43] to motivate its users. We adapted the work done in study [17] and before they finalized the design, they tested various scoring systems, both with and without additional incentives. In their study, the authors used push notifications to enhance player engagement. Multiple notifications are sent through the game. We used the same push notifications as in the previous study [17]. To give an example, when a player loses his/her top position on the leader board 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 5(b)).

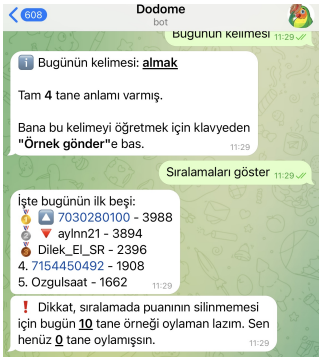
3.2 Game Implementation

The Dodiom [4] was initially designed by [17] as a Telegram bot to leverage Telegram's advanced features, such as multi-platform support. This approach enabled them to concentrate on the NLP back-end instead of developing web-based or mobile versions of the game. We downloaded Dodiom online and modified its main goal to focus on word senses instead of idioms. In Dodiom, Python-telegram-bot [5] library was used to communicate with the Telegram servers and to implement the main messaging interface. A PostgreSQL [6] database was used as the data back-end. The "Love Bird"

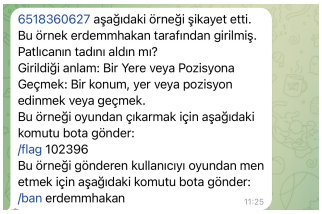
[4] https://github.com/Dodiom/dodiom
[5] https://github.com/python-telegram-bot/python-telegram-bot/
[6] https://www.postgresql.org/



(a) Achievements



(b) Show Scoreboard



(c) Report

Fig. 4. Some interaction screens.

(a) Score increase

(b) Falling back from the top 1

Fig. 5. Notification samples.

Telegram sticker package have 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 [38] was used for tokenization. For Dodiom they used NLTK for idiom submissions but for our resarch we used them for word tokenization. Zeyrek [7] is also used for the lemmatization of Turkish. If the words' surface forms are missing from the submission due to typos or incorrect entries, the player is prompted to make a new submission.

3.3 Modifications and Enhancements

Section 3.1 provided our proposed design for a gamified crowdsourcing approach for the WSD task. As explained before, the approach has been inspired from an earlier study [17] focusing on

[7]An NLTK based lemmatizer, customized for the Turkish language, https://zeyrek.readthedocs.io

collecting literal or figurative sentence samples for specific idioms. In addition to our redesign of the game for the WSD task, we also needed to make enhancements over [17] to be able to apply it within a larger community. Below we list these modifications and enhancements:

(1) **Dynamic Approach**: For each ambiguous word, there are multiple meanings to explore in the game. If all meanings were presented in the same order (e.g., the game always starting by asking for an example of the first meaning, followed by the second), players would consistently start with the same meaning, resulting in less engagement with subsequent meanings. To ensure a balanced distribution of example sentences across all meanings, we implemented the DodoMe game with a dynamic approach. The first meaning is presented to the initial player, and subsequent players are prompted with the meaning that currently has the fewest examples. Ref. [17] has used different scoring to collect different usage examples for idioms (e.g., to encourage adjacent vs separated usage of idiom components). With modified this behavior to automatically direct the players to less frequently exemplified senses.
(2) **Limited Number of Examples**: To maintain a balanced distribution of example sentences across all meanings, we added a rule that limits players to a maximum of five examples[8] for the same meaning. After reaching this limit, players are prompted to proceed to the next meaning with less frequent samples, ensuring a balanced number of examples for each meaning.
(3) **Fishing Questions**: To ensure the accuracy of player reviews during the review process, we enhanced the DodoMe game with fishing questions. These pre-answered examples included both true positives and false negatives, presented in the review section to the users. The purpose of these fishing questions was to covertly evaluate whether players were participating honestly or merely attempting to earn points. For each word, we incorporated ten fishing questions, with two or three examples per meaning. Examples were presented randomly, alternating between actual examples and fishing questions, and players were asked to review if the given examples matched the intended meanings. The system already knew the correct answers and automatically calculated players' performance on these fishing questions, which contributed to determining the players' reliability scores.
(4) **Banning Players**: Players with a reliability score below 30% were banned for the remainder of the day, though they could return to play the next day with a new word. However, if a player was banned more than three times, they were permanently removed from the game. This strategy effectively filtered out cheaters and enhanced the overall quality of examples and reviews.
(5) **Promoting Reviews**: In Dodiom, the review process was not widely adopted among players, as most preferred writing example sentences for idioms rather than reviewing others' submissions. To address this and make the review process more mandatory in DodoME, we implemented a rule requiring players to review at least 10 examples submitted by others before the end of the day. Failure to meet this requirement would result in the deletion of their earned points, preventing them from winning the game even with the highest score. This approach ensured that players reviewed at least 10 examples, including fishing questions, allowing us to measure and calculate their reliability scores.
(6) **Today's Winner**: As we began playing DodoMe, some users started winning continuously, causing a drop in motivation among other players. Based on feedback from participants and to keep everyone motivated, we introduced a new rule: the current day's winner could continue playing the next day but would be ineligible to win, even with the highest score. This meant that the day's winner could not win again the following day but would be eligible

[8]The maximum threshold can be set to any number greater than one to avoid slowing down the game.

the day after. Additionally, we started tracking players' weekly performances. Players who appeared in the top five more than four days a week without ever securing first place were awarded a weekly prize. Although this prize was smaller than the daily winner's prize, it provided an extra incentive for continued participation.

(7) **Prizes**: The prize for the daily winner was 250 **Turkish Liras** (**TL**), equivalent to approximately $7.8 in April 2024. Weekly winners received 200 TL, or about $6.2. In total, we spent 7500 TL on daily winners and 800 TL on weekly winners, amounting to 8300 TL or approximately $256.

(8) **LLM Controls**: We also monitored the game for unauthorized use of LLMs, such as ChatGPT, by tracking the timing of user entries to ensure no one was using or connecting to a generated AI API. We identified a user who connected ChatGPT to generate example sentences and subsequently banned them from the game. One should note that we did not forbid the users from using LLMs or search engines to crawl high quality samples. However in that case we expect that they read and validate the samples before submitting the system. Although we are aware that tracking the timings is not a perfect solution to validate this, it is an initial precaution which that could be enhanced in further studies.

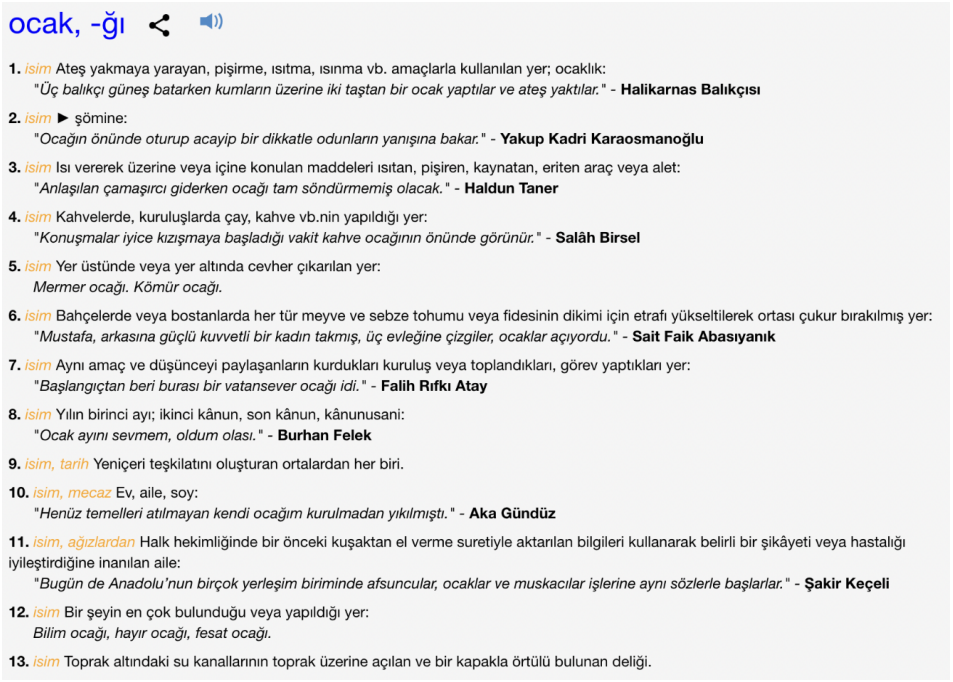
(9) **Focused Users**: Another difference between DodoMe and Dodiom is our shift in target users, as detailed in Section 3.4. While Dodiom was primarily played by academic users such as university students, lecturers, and professors, DodoMe was promoted through Instagram and attracted a broader audience, particularly random users aged 25–34. More statistics on the users can be found in Figure 7.

3.4 Reduction of Word Meanings and Sense Bags

WSD is a challenging task in NLP because many words have multiple meanings. For instance, in resources like TDK or WordNet, a single word can have as many as 59 different meanings (Table 2). The list of played nouns and verbs and their meaning frequencies are given in Table 2. However, not all these meanings are commonly used, which complicates the process of accurately determining the intended sense in a given context. To illustrate the challenge, let us examine the ambiguous Turkish word "ocak", which has 13 meanings listed in the TDK, as shown in Figure 6. Consider the first four meanings of this Figure 6:

— 1. noun: A place used for lighting fires, cooking, heating, or warming; a hearth: "Three fishermen built a hearth out of two stones on the sand and lit a fire as the sun was setting." - Halikarnas Balıkçısı
— 2. noun: A fireplace: "He sits in front of the fireplace, staring intently at the burning logs." - Yakup Kadri Karaosmanoğlu
— 3. noun: A device or tool that provides heat to warm, cook, boil, or melt materials placed on or in it: "Apparently, the laundress had not completely put out the stove before leaving." - Haldun Taner
— 4. noun: A place in cafés or establishments where tea, coffee, and so on, is prepared: "When the discussions started to heat up, he appeared in front of the café's kitchen area." - Salâh Birsel.

Although these meanings differ in context, they are conceptually close and often overlap. For example, in the second sense, while the word "fireplace" is used in the example, it is not explicitly clear whether it refers to a traditional fireplace or another type of wood-burning stove. Without considering the full context, determining the exact sense is challenging. Similarly, distinguishing between the first, second, and third meanings–"hearth," "fireplace," and "stove"–requires careful analysis of the context and additional clues, which can be difficult to ascertain from the sentence



ocak, -ğı

1. *isim* Ateş yakmaya yarayan, pişirme, ısıtma, ısınma vb. amaçlarla kullanılan yer; ocaklık:
"*Üç balıkçı güneş batarken kumların üzerine iki taştan bir ocak yaptılar ve ateş yaktılar.*" - **Halikarnas Balıkçısı**
2. *isim* ► şömine:
"*Ocağın önünde oturup acayip bir dikkatle odunların yanışına bakar.*" - **Yakup Kadri Karaosmanoğlu**
3. *isim* Isı vererek üzerine veya içine konulan maddeleri ısıtan, pişiren, kaynatan, eriten araç veya alet:
"*Anlaşılan çamaşırcı giderken ocağı tam söndürmemiş olacak.*" - **Haldun Taner**
4. *isim* Kahvelerde, kuruluşlarda çay, kahve vb.nin yapıldığı yer:
"*Konuşmalar iyice kızışmaya başladığı vakit kahve ocağının önünde görünür.*" - **Salâh Birsel**
5. *isim* Yer üstünde veya yer altında cevher çıkarılan yer:
Mermer ocağı. Kömür ocağı.
6. *isim* Bahçelerde veya bostanlarda her tür meyve ve sebze tohumu veya fidesinin dikimi için etrafı yükseltilerek ortası çukur bırakılmış yer:
"*Mustafa, arkasına güçlü kuvvetli bir kadın takmış, üç evleğine çizgiler, ocaklar açıyordu.*" - **Sait Faik Abasıyanık**
7. *isim* Aynı amaç ve düşünceyi paylaşanların kurdukları kuruluş veya toplandıkları, görev yaptıkları yer:
"*Başlangıçtan beri burası bir vatansever ocağı idi.*" - **Falih Rıfkı Atay**
8. *isim* Yılın birinci ayı; ikinci kânun, son kânun, kânunusani:
"*Ocak ayını sevmem, oldum olası.*" - **Burhan Felek**
9. *isim, tarih* Yeniçeri teşkilatını oluşturan ortalardan her biri.
10. *isim, mecaz* Ev, aile, soy:
"*Henüz temelleri atılmayan kendi ocağım kurulmadan yıkılmıştı.*" - **Aka Gündüz**
11. *isim, ağızlardan* Halk hekimliğinde bir önceki kuşaktan el verme suretiyle aktarılan bilgileri kullanarak belirli bir şikâyeti veya hastalığı iyileştirdiğine inanılan aile:
"*Bugün de Anadolu'nun birçok yerleşim biriminde afsuncular, ocaklar ve muskacılar işlerine aynı sözlerle başlarlar.*" - **Şakir Keçeli**
12. *isim* Bir şeyin en çok bulunduğu veya yapıldığı yer:
Bilim ocağı, hayır ocağı, fesat ocağı.
13. *isim* Toprak altındaki su kanallarının toprak üzerine açılan ve bir kapakla örtülü bulunan deliği.

Fig. 6. Word meanings in Turkish language association (TDK) for the ambiguous word "ocak".

alone. This difficulty highlights the inherent challenges in both human interpretation and WSD systems. Even for humans, the subtle distinctions between these senses can be ambiguous, making it even more demanding for WSD algorithms to reliably resolve such cases.

WSD systems are trained on examples from natural texts and require a substantial amount of data. For Turkish, WSD studies often rely on resources such as the TDK. However, a common issue in dictionaries is the presence of closely related or overlapping definitions, a challenge observed in the dictionaries of many languages. In the literature, this issue is addressed by grouping similar meanings into word sense bags [20, 39, 46]. A word sense bag clusters semantically similar meanings of a word under a single category, streamlining their representation and making it easier to focus on the most relevant senses for WSD tasks. This method prioritizes commonly used meanings while filtering out less significant or overly specific ones, enhancing the efficiency and accuracy of WSD systems. Constructing sense bags often relies on language dictionaries, supplemented by example sentences to clarify each meaning. However, when example sentences are insufficient or absent, it becomes challenging for the system to fully comprehend the intended sense within the bag. Ref. [50] states that "well-defined sense groups can be of value in improving WSD by both humans and machines".

To tackle the complexity of semantic ambiguity in Turkish, our study employs a hybrid approach that combines ChatGPT and human annotators to create and refine word sense bags. Our annotation team consists of two experts, native speakers of the language in focus: the first is the lead author of this article and a PhD candidate in NLP, and the second is a PhD holder in the same field. Both annotators have extensive experience in computational linguistics and linguistic data annotation. To quantify the consistency between annotators, we calculated Cohen's kappa coefficient[10], which measures agreement corrected for chance overlap. The resulting kappa value of **0.897** indicates very strong agreement according to standard benchmarks. While the concept of reducing meanings

into sense bags is not new, our approach leverages LLMs like ChatGPT in combination with human annotators, which adds a novel layer to this process. ChatGPT is used to generate initial suggestions for grouping related meanings of ambiguous words, drawing on large-scale language data to provide a broad perspective. These suggestions are then reviewed by human annotators, who use their linguistic intuition and expertise to adjust and validate the sense bags. The annotators assess the accuracy of the groupings, determining whether additional meanings should be considered or if the suggested meanings sufficiently cover all possible interpretations. Through this iterative process, the list of meanings is refined and downsized, with cross-referencing examples helping to consolidate meanings into a more manageable set, typically reducing each word to a maximum of four or five primary senses. This hybrid approach ensures a balance between automation and human insight, enabling the creation of robust and contextually accurate word sense bags. These refined sense groupings serve as a foundation for building more effective WSD models tailored to the Turkish language. By integrating sense bagging into the WSD process, we streamline the handling of ambiguous words and improve the overall precision of language models. For example, the verb "çık" (to go out) has 59 meanings in WordNet and 56 in the TDK. By applying our hybrid approach, we consolidated these into 4 distinct meanings that encompass all 59 interpretations.

By using the hybrid approach outlined earlier, we successfully reduced the 13 meanings of the word "ocak" to 4 primary senses:

— 1. Calendar Month: Refers to January, the first month of the year.
Example: "Ocak ayı çok soğuk oluyor." (January is very cold.)
— 2. Cooking Appliance: Refers to a stove, oven, or fireplace used for cooking or heating.
Example: "Evden çıkarken ocağın altını kapatmayı unutma." (Don't forget to turn off the stove before leaving home.)
— 3. Mine or Quarry: Refers to a site where minerals, such as coal or marble, are extracted.
Example: "Babam kömür ocağında çalışıyor." (My father works in a coal mine.)
— 4. Community, Family, Lineage: Refers to a group of soldiers serving together or a family/lineage living as a unit.
Example: "Henüz temelleri atılmayan kendi ocağım kurulmadan yıkılmıştı." (My own household was destroyed before it was even established.)

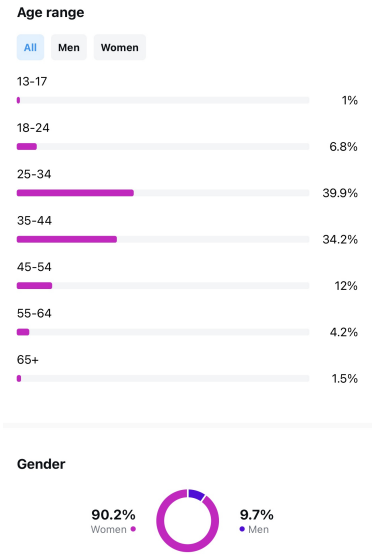
This methodology not only simplifies the disambiguation process, but also has the potential to enhance the accuracy and efficiency of WSD systems. We believe that this approach will contribute to future advancements in semantic ambiguity resolution for Turkish and support the development of more reliable natural language processing tools. A key issue in this domain has been the lack of sufficient example sentences to clarify meanings within sense bags, limiting system comprehension. Our system addresses this challenge by generating new example sentences for each sense bag, thereby improving contextual understanding and laying the groundwork for improved WSD performance.

4 Analyses

This section first introduces the methodology and participants in Section 4.1, then provides an analysis of the collected data in Section 4.2, and finally discusses error analysis in Section 4.3.

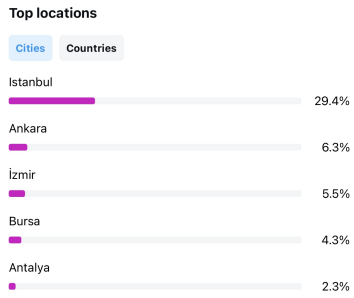
4.1 Methodology and Participants

The game was deployed in two phases: an initial 15-day period focused on 15 ambiguous nouns, followed by a second 15-day phase that introduced 15 ambiguous verbs, as detailed in Table 2. Testing these ambiguous words (15 nouns and 15 verbs) involved 2,917 participants and provided valuable insights that significantly improved the game design. The game-play spanned 30 days,



(a) Age Range and Gender of Following Users

(b) DodoMe Account on Instagram



(c) Top Locations of Following Users

Fig. 7. DodoMe on Instagram.

from March 25 to April 23, 2024. The words featured in the game were selected based on their propensity for ambiguity, as outlined in the study by [29]. A detailed list of the nouns and verbs, along with their meaning frequencies, can be found in Table 2. To streamline the analysis, the meanings for each word were reduced to a maximum of four or five, a process thoroughly described in Section 3.4.

To reach a diverse audience, we launched an Instagram account named "@dodome_bot" and used influencer networks alongside Instagram ads to promote our Telegram bot widely. This strategy helped us gather a total of 1,527 followers, 90% of whom were female and 10% male. The largest age groups were 40% aged between 25–34 and 34% aged between 35–44, with 29% residing in Istanbul, Türkiye. Detailed statistics are shown in Figure 7. While we lack comprehensive background details about our followers or players, it is evident that they are real-world users, most of whom are not academic researchers. Unlike previous crowdsourcing study, such as [17], which primarily engaged students, AI communities, or academic participants, our approach utilized social media to reach a more diverse and heterogeneous audience. This approach allowed us to include a broader mix of contributors, bringing diverse perspectives into the data collection and annotation process.

In addition to deploying the game, announcements and daily updates were actively shared on Instagram, garnering significant engagement. Over the 30-day period, the Instagram posts received approximately ~360K views, ~256 likes, ~88 shares, and ~132 comments. From the start, participants were informed that their involvement would contribute to the creation of a public data resource. This transparency encouraged many users to support the project enthusiastically, with several sharing the announcements from their own accounts, further amplifying visibility. To boost participation, monetary rewards of 250 TL per day were offered, and these announcements were also integrated into the game. The primary outreach was facilitated by a Turkish influencer with approximately ~178K Instagram followers, 90% of whom are women aged between 25 and 44.

Table 3. User Statistics

Statistic	Turkish
Total # of users who played the game	2,917
... for only 1 day	1,028 (54%)
... for 2-3 days	584 (31%)
... for 4-7 days	215 (11%)
... for >= 7 days	75 (4%)
Crowd Type	Anyone from Instagram Social Media

This network played a pivotal role in reaching a diverse audience, beyond the typical academic or AI-based communities. Overall, the game attracted 2,917 players, and user statistics reveal that nearly 13% of participants engaged with the game for more than seven days. Table 3 provides a detailed breakdown of user demographics and engagement levels. As shown in Table 3, a total of 2,917 users subscribed to the game. Of these, 1,028 users (54% of total participants) played the game for just 1 day during the 30-day period. Another 584 users (31%) engaged for 2–3 days, while 215 users (11%) played for 4–7 days. A smaller group of 75 users (4%) played the game for more than 7 days, representing the most dedicated players who consistently participated and frequently won the daily monetary prizes. Detailed user statistics are further provided in Section 4.2.

In this study, we did not assume that users would always provide accurate examples of sentences or review submissions honestly. To evaluate user performance and measure the quality of submitted sentences, a phishing approach commonly used in crowdsourcing studies was adopted [47]. This method involves presenting users with questions where the correct answer is known only to moderators. For each ambiguous word (noun or verb), 10 phishing example sentences were carefully prepared. Instead of relying on human annotators to manually assess the quality of submissions–a costly and labor-intensive process–the system automatically evaluated user performance through these phishing questions. The phishing sentences were meticulously selected by human annotators to cover various meanings of the word. Some sentences were correctly labeled, while others were intentionally mislabeled. During the review process, these phishing questions were randomly presented to users. Users were expected to mark the sentence as correct if it was accurately labeled and as incorrect if it was mislabeled. This system aimed to prevent users from intentionally mislabeling sentences to gain points themselves or to lower another user's score, ensuring a fair and honest evaluation process. To maintain score integrity, users were required to review at least 10 sentences per day, with a minimum of 5 being phishing questions. This approach allowed us to assess both the overall quality of the data and the individual performance of the user. If a user's daily performance based on phishing questions fell below 30%, they were temporarily banned from the game to preserve data quality. Users who consistently scored below 30% for three consecutive days were permanently removed from the game.

4.2 Data Analysis

During the 30-day period, we collected a total of 158,576 submissions and 465,787 reviews from users for both Turkish nouns and verbs, as shown in Table 4. The game was played for 15 days with nouns and 15 days with verbs. Of the total 158,576 submissions, 99,906 (63%) were for nouns, while 58,670 (37%) were for verbs. Of the total 465,787 reviews, 344,980 (74%) were related to nouns, and 120,807 (26%) were for verbs. This distribution aligns with feedback provided by participants through comments and messages, indicating that playing with nouns was perceived as considerably easier than playing with verbs. When we look at Table 4, the word "hava" (air) with its reduced 3 senses received a total of 11,395 example sentence submissions, making it the highest in the noun

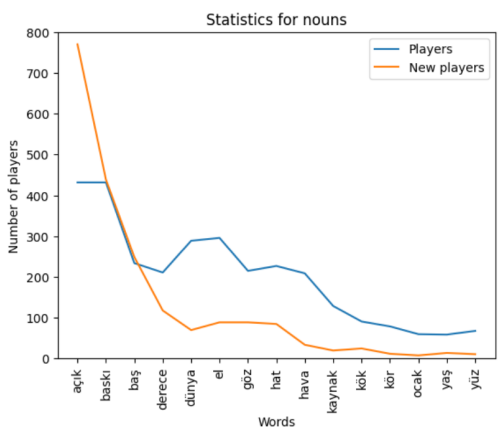
Table 4. Staticstics of the Collected WSD Data

	Day	Word	Meaning Count	Submission Count	Review Count
	1	açık (open)	4	6488	9368
	1	baskı (pressure)	3	8032	14573
	3	baş (head)	4	3834	8904
	4	derece (degree)	3	9304	58532
	5	dünya (world)	4	11325	66458
	6	el (hand)	3	10347	65095
	7	göz (eye)	4	6862	16387
NOUN	8	hat (line)	3	9062	29860
	9	hava (air)	3	11395	24025
	10	kaynak (resource)	4	5731	14334
	11	kök (root)	4	2897	9559
	12	kör (blindness)	3	5452	7545
	13	ocak (stove)	4	3431	6460
	14	yaş (age)	5	3052	6070
	15	yüz (face)	3	2694	7810
	16	açmak (to open)	6	2263	9236
	17	almak (to take)	4	4330	9261
	18	atmak (to throw)	4	4745	12947
	19	bakmak (to look)	3	2049	4619
	20	çevirmek (to turn)	3	3936	5086
	21	çıkmak (to exit)	3	4320	7357
	22	geçmek (to pass)	3	5317	13550
VERB	23	gelmek (to come)	3	5310	6632
	24	girmek (to enter)	3	2117	4967
	25	görmek (to see)	2	3737	6910
	26	kalmak (to stay)	3	3602	6553
	27	olmak (to be)	3	5719	7982
	28	sürmek (to drive)	4	3514	10920
	29	vermek (to give)	3	4265	8487
	30	yapmak (to do)	3	3446	6300
			Total	**158,576**	**465,787**

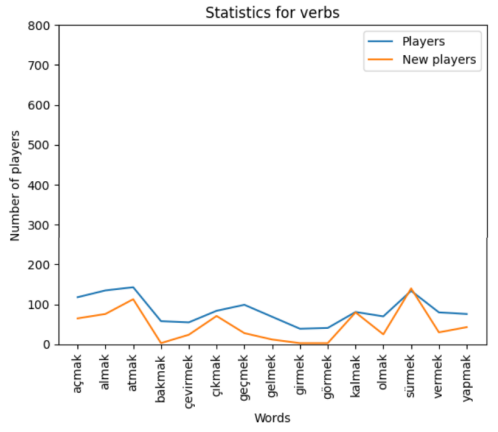
category. However, when considering review counts, the word “dünya” (world) stands out with 66,458 reviews, the highest among nouns. The maximum number of submissions for verbs is 5,719 for the verb “olmak” (to be), while the highest number of reviews is 13,550 for the verb “geçmek” (to pass).

In this section, we also present a detailed analysis of the data, including: (1) daily user engagement statistics, showcasing the number of new users visiting the bot each day (Figure 8); (2) daily submission and review trends (Figure 9); and (3) average review counts per word (Figure 10). Figures 8(a) (nouns) and 8(b) (verbs) illustrate the daily count of new users who visit the bot, regardless of whether they have actively started playing. These graphs reveal a significant influx of users during the initial days of noun gameplay, largely driven by the crowd’s curiosity. Although the number of new users gradually decreased during the verb phase, the game continued to attract new players daily. This sustained engagement was fueled by sharing updates on social media, particularly highlighting the monetary prizes won by players, which reignited interest and motivated new users to join, creating a consistent influx of participants. As a result, the daily count of new users during this phase averaged between 100 and 200 players.

Figure 9 shows the daily submission and review counts for both nouns and verbs. It also demonstrates that the target of 100 submissions per ambiguous word was consistently met and often exceeded for both tasks (nouns and verbs). On average, nouns received 6,660 submissions daily, while verbs averaged 3,911. The average review count for nouns reached 22,998, compared with 8,053 for verbs. In the first days, the number of reviews was notably high, while submissions were comparatively lower. However, on days with a high volume of samples, even a modest review

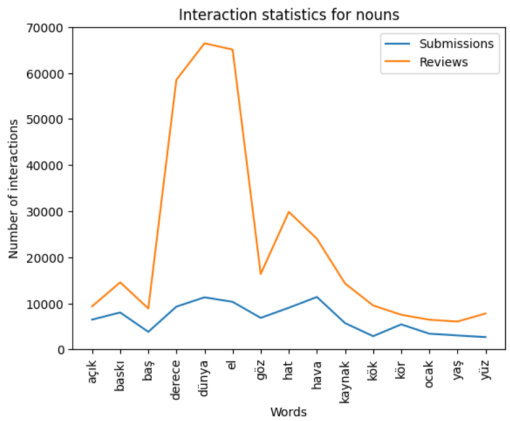


(a) Daily Player Count for Nouns

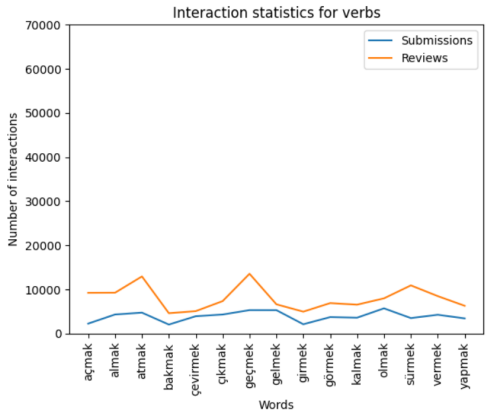


(b) Daily Player Count for Verbs

Fig. 8. Daily play statistics.



(a) Daily Submission & Review Counts for Nouns



(b) Daily Submission & Review Counts for Verbs

Fig. 9. Daily statistics for submissions and reviews.

average ensured that many samples still received more than 3,000 submissions. Figure 10 illustrates the average review counts per word for both nouns and verbs.

In calculating phishing statistics, we also evaluated user performance, as shown in Table 5. This table presents the performance based on phishing questions of the top 10 users and all users for nouns and verbs. Players who were banned and removed from the game due to low performance are excluded from this analysis. The top 10 users were determined by their consistent participation throughout the 30-day game period, and their average daily scores were used to calculate their performance. The All Users category includes the statistics of all players except those who were banned. According to Table 5, the performance of both top 10 users and all users varies between nouns and verbs. At the end of the game, the average performance for nouns among the top 10 users was 66%, while the average for all users was slightly higher at 67%. For verbs, the average performance of the top 10 users was 52%, compared with 49% for all users. These statistics

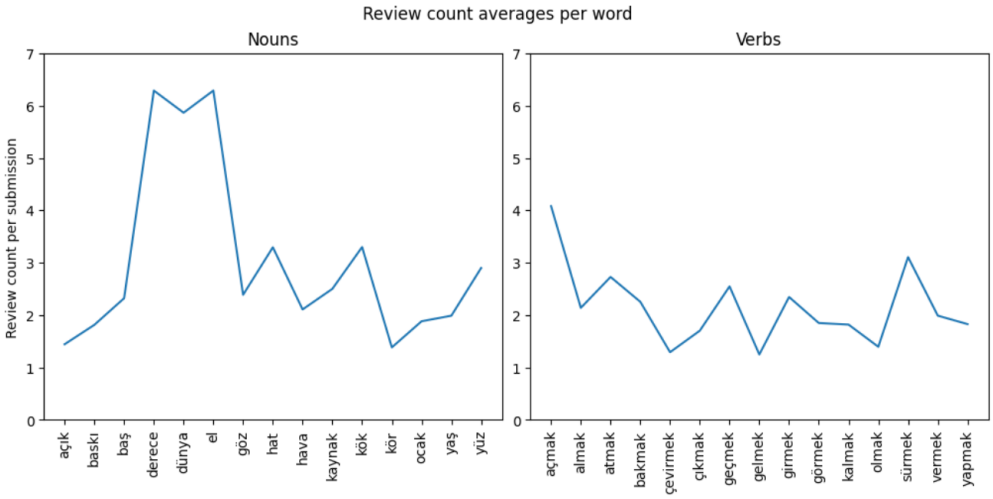


Fig. 10. Review count averages per word.

Table 5. Phishing Performances of Users

	Phishing Performances of Users				Phishing Performances of Users		
	Word	% of Top 10 Users	% of All Users		Word	% of Top 10 Users	% of All Users
	açık (open)	66	58		açmak (to open)	66	70
	baskı (pressure)	62	59		almak (to take)	58	62
	baş (head)	68	66		atmak (to throw)	62	52
	derece (degree)	64	75		bakmak (to look)	56	56
	dünya (world)	58	74		çevirmek (to turn)	60	66
	el (hand)	64	73		çıkmak (to exit)	54	58
	göz (eye)	82	70		geçmek (to pass)	54	55
NOUN	hat (line)	64	76	**VERB**	gelmek (to come)	NaN	NaN
	hava (air)	64	73		girmek (to enter)	62	55
	kaynak (resource)	72	73		görmek (to see)	16	11
	kök (toot)	66	69		kalmak (to stay)	28	16
	kör (blindness)	62	59		olmak (to be)	54	56
	ocak (stove)	72	74		sürmek (to drive)	64	67
	yaş (age)	54	38		vermek (to give)	32	17
	yüz (face)	72	73		yapmak (to do)	56	50

provide insights into the success rates of the crowdsourced contributors and align with previously presented findings, confirming the reliability and consistency of the collected data. Additionally, this study statistically derived crowdsourcing success rates through the use of phishing questions, demonstrating the effectiveness of this approach in evaluating the quality and performance of user-generated data.

To further evaluate the reliability of the collected data, we conducted an **inter-annotator agreement analysis (IAAs)** using the Cohen's Kappa statistic[10]. Separate analyses were performed for nouns and verbs. For instance, for nouns, the word "ocak" was analyzed based on its four primary meanings, and we extracted ten example sentences with the most votes (three or more) for each meaning. Similarly, for verbs, the word "Çıkmak" was analyzed across three distinct meanings, and ten example sentences with the highest votes were selected for each meaning. In total, 40 sentences were analyzed for nouns, and 30 sentences were analyzed for verbs. These sentences were re-annotated by human annotators, and Cohen's kappa statistics were calculated for nouns and verbs. According to the results, for nouns, the unweighted kappa statistic was 0.9874, with a 95% confidence interval upper limit of 1.00, indicating almost perfect agreement (kappa between

0.81 and 1.00). For verbs, the unweighted kappa statistic was 0.3651, with a 95% confidence interval upper limit of 0.7157, corresponding to substantial agreement (kappa between 0.61 and 0.80). These findings demonstrate that the data collected for both nouns and verbs are of high quality, as indicated by the inter-annotator agreement metrics.

4.3 Error Analysis: Data Quality and Misuse Scenarios

During the gamified crowd-sourcing process for our 15 noun and 15 verb target words, we observed several systematic error patterns. The situation arises not only from semantic misinterpretations of the target word senses, but also from participants' interaction behavior within the game interface. We highlight two prominent cases below:

Repetitive and Low-Value Samples: For certain target words with highly common literal senses (e.g., "baskı" meaning pressure, printing or issue), some participants repeatedly submitted essentially the same sentence with only minor alterations (e.g., "1. baskı geldi" (1st issue has arrived) / "2. baskı geldi" (2nd issue has arrived) / "3. baskı geldi," (3rd issue has arrived) with replacing "1st" with "2nd," "3rd," and so on). Although superficially different, such examples did not add semantic variation; instead, they overpopulated a single-sense category while reducing overall diversity. In some cases, this behavior appeared malicious, as participants deliberately produced low-quality repetitive examples. 1.25% of submissions for the target word "baskı" fell in this category. Although relatively rare, we identified the issue early and addressed it by applying a cosine similarity check: if the similarity with a stored example exceeded 90%, the submission was rejected and the participant warned, with repeated violations leading to temporary suspension. This mechanism effectively prevented redundancy and preserved the richness of the dataset.

As a second type of error, we observed cases where the context provided by the participant was insufficient to clearly disambiguate the sense. For instance, the sample "Baskı geldi" could be interpreted as either "The issue has arrived" or "Pressure has increased". We observed that these types of examples were grouped under the category of items disliked by crowd raters.

Automated Submissions via External Models: We also observed submission patterns where sentences were generated and sent in an unrealistically short time frame, far faster than a human could reasonably compose and submit them through the game interface. Upon backend inspection, we inferred that one participant among all players (2,917) had connected an AI model to automate responses. This posed the risk of flooding the dataset with machine-generated sentences lacking genuine human semantic interpretation. To address this, we implemented a submission timer that enforced a minimum interval between valid entries. Sentences sent faster than this threshold were automatically discarded, and the participant was banned, preventing further automated submissions. Repeated offenders would be similarly flagged for review, ensuring the integrity of the dataset.

5 Conclusion and Future Work

In this study, we introduced a novel gamified crowdsourcing approach to address long-standing challenges in WSD data collection. Using a messaging bot structured as an asynchronous multiplayer game, we engaged a diverse audience to generate and evaluate WSD examples, creating a rich and reliable dataset. Our findings demonstrate that gamified crowdsourcing can effectively overcome traditional obstacles, such as data scarcity and the imbalance between ambiguous and unambiguous word samples, while also fostering sustained engagement among participants.

Our contributions are threefold: (1) the first-ever application of a crowd-creating and crowd-rating gamified crowdsourcing approach for WSD, (2) the development of a new WSD dataset for Turkish, and (3) the creation of a diverse dataset with potential applicability for multilingual WSD research. The method not only accelerates the construction of the WSD corpus, but also ensures high-quality labeled training data, crucial for advancing robust and precise language models.

Performance evaluations further highlight this difference. Nouns achieved a higher average performance rate of 73%, with accurate and well-structured data. Verbs, in contrast, presented a more challenging task with a performance rate of 61%, likely due to their higher ambiguity and increased potential for user error. Nonetheless, the overall reliability of the collected data is evident (proven by Kappa statistics), with nouns offering stronger, high-quality results and verbs contributing a solid yet more complex dataset. These performance insights confirm that gamified crowdsourcing can successfully produce high-quality WSD datasets while balancing user engagement and data integrity across different linguistic complexities. To ensure data quality, we applied a two-step validation scheme: (1) peer rating, where participants scored each other's entries, and (2) phishing questions, where control items were used to assess participant reliability. To assess annotation consistency, we calculated inter-annotator agreement separately for nouns and verbs, yielding near-perfect and moderate agreement, respectively.

References

[1] Bekhouche Abdelaali and Yamina Tlili-Guiassa. 2022. Swarm optimization for Arabic word sense disambiguation based on English pre-trained word embeddings. In *2022 5th International Symposium on Informatics and its Applications (ISIA)*. IEEE, 1–6.

[2] Onur Açıkgöz, Ali Tunca Gürkan, Burak Ertopçu, Ozan Topsakal, Berke Özenç, Ali Buğra Kanburoğlu, İlker Çam, Begüm Avar, Gökhan Ercan, and Olcay Taner Yıldız. 2017. All-words word sense disambiguation for Turkish. In *2017 International Conference on Computer Science and Engineering (UBMK)*. IEEE, 490–495.

[3] Sinan Akçakaya. 2019. *All-Words Word Sense Disambiguation in Turkish*. Ph.D. Dissertation. ISIK UNIVERSITY.

[4] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.

[5] Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*. International Joint Conference on Artificial Intelligence, Inc, 4330–4338.

[6] Sarfraz Bibi, Sohail Asghar, and Muhammad Zubair. 2024. Sense Unveiled: Enhancing Urdu corpus for nuanced word sense disambiguation. *IEEE Access* 12 (2024), 126329–126343.

[7] Chris Biemann and Valerie Nygaard. 2010. Crowdsourcing wordNet. In *Proceedings of the 5th Global WordNet Conference*. Global WordNet Association, Mumbai, India.

[8] Puberun Boruah. 2022. A novel approach to word sense disambiguation for a low-resource morphologically rich language. In *2022 IEEE 6th Conference on Information and Communication Technology (CICT)*. IEEE, 1–6.

[9] Zhenguang Garry Cai, David A. Haslett, Xufeng Duan, Shuqi Wang, and Martin John Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*. Association for Computational Linguistics, Bangkok, Thailand, 37–56. https://doi.org/10.18653/v1/2024.cmcl-1.4

[10] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[11] Hafsa Dar, Romana Aziz, Javed Ali Khan, Muhammad IkramUllah Lali, and Nouf Abdullah Almujally. 2024. Gamify4LexAmb: A gamification-based approach to address lexical ambiguity in natural language requirements. *PeerJ Computer Science* 10 (2024), e2229.

[12] Robert David, Anna Kernerman, Ilan Kernerman, Nicolas Ferranti, and Assaf Siani. 2024. Multilingual word sense disambiguation for semantic annotations: Fusing knowledge graphs, lexical resources, and large language models. In *Proceedings of CEUR Workshop (RAGE-KG 2024)*. 16–22. https://ceur-ws.org/Vol-3950/short3.pdf

[13] Oier Lopez de Lacalle and Eneko Agirre. 2015. A methodology for word sense disambiguation at 90% based on large-scale crowdsourcing. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*. 61–70.

[14] Clovis Holanda do Nascimento, Vinicius Cardoso Garcia, and Ricardo de Andrade Araújo. 2024. A word sense disambiguation method applied to natural language processing for the portuguese language. *IEEE Open Journal of the Computer Society* 5 (2024), 268–277.

[15] Nicole M. Duggan, Mike Jin, Maria Alejandra Duran Mendicuti, Stephen Hallisey, Denie Bernier, Lauren A. Selame, Ameneh Asgari-Targhi, Chanel E. Fischetti, Ruben Lucassen, Anthony E. Samir, et al. 2024. Gamified crowdsourcing as a novel approach to lung ultrasound data set labeling: Prospective analysis. *Journal of Medical Internet Research* 26 (2024), e51397.

[16] Gesty Ernestivita, Vikas Kumar, and Tiara Nur Anisah. 2024. Role of gamification in crowdsourcing. In *Marketing and Gamification*. Routledge, 24–38.

[17] Gülşen Eryiğit, Ali Şentaş, and Johanna Monti. 2023. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering* 29, 4 (2023), 909–941.
[18] Ghada M. Farouk, Sally S. Ismail, and Mostafa M. Aref. 2023. Transformer-Based word sense disambiguation: Advancements, impact, and future directions. In *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE, 140–146.
[19] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
[20] Jorge García Flores, Laurent Gillard, Olivier Ferret, and Gaël de Chalendar. 2008. Bag of senses versus bag of words: Comparing semantic and lexical approaches on sentence extraction. In *TAC*.
[21] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation* 59 (2024), 1719–1746. https://doi.org/10.1007/s10579-024-09766-4
[22] Aparitosh Gahankari, Avinash S Kapse, Mohammad Atique, VM Thakare, and Arvind S Kapse. 2023. Hybrid approach for Word Sense Disambiguation in Marathi Language. In *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*. IEEE, 1–4.
[23] William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26, 5–6 (1992), 415–439. https://doi.org/10.1007/BF00136984
[24] David Geiger and Martin Schader. 2014. Personalized task recommendation in crowdsourcing information systems–Current state of the art. *Decision Support Systems* 65, 1 (2014), 3–16. https://doi.org/10.1016/j.dss.2014.05.007
[25] Bairu Hou, Fanchao Qi, Yuan Zang, Xurui Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Try to substitute: An unsupervised Chinese word sense disambiguation method based on hownet. In *Proceedings of the 28th International Conference on Computational Linguistics*. 1752–1757.
[26] Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. 57–60.
[27] Xiaolong Huang, Edmond Zhang, and Yun Sing Koh. 2019. Supervised clinical abbreviations detection and normalisation approach. In *PRICAI 2019: Trends in Artificial Intelligence: 16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, August 26-30, 2019, Proceedings, Part III 16*. Springer, 691–703.
[28] Bahar İlgen, Eşref Adali, and A Cüneyd Tantuğ. 2012. Building up lexical sample dataset for Turkish word sense disambiguation. In *2012 International Symposium on Innovations in Intelligent Systems and Applications*. IEEE, 1–5.
[29] Bahar Ilgen, EŞREF ADALI, and Ahmet Cüneyd Tantuğ. 2016. Exploring feature sets for Turkish word sense disambiguation. *Turkish Journal of Electrical Engineering and Computer Sciences* 24, 5 (2016), 4391–4405.
[30] Rafał Jaworski, Sanja Seljan, and Ivan Dunđer. 2023. Four million segments and counting: Building an english-croatian parallel corpus through crowdsourcing using a novel gamification-based platform. *Information* 14, 4 (2023), 226.
[31] David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 556–562.
[32] Anıl Can Kara and Oğuzhan Yetimoğlu. 2019. Turkish word sense disambiguation using wordNet. Undergraduate Senior Project. Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey. https://www.cmpe.boun.edu.tr/content/turkish-word-sense-disambiguation-usingwordnet-2. Project Advisor: Tunga Güngör, Spring 2019.
[33] Hadi Khalilia, Jahna Otterbacher, Gabor Bella, Rusma Noortyani, Shandy Darma, and Fausto Giunchiglia. 2024. Crowdsourcing lexical diversity. arXiv:2410.23133. Retrieved from https://arxiv.org/abs//2410.23133
[34] Chandrakant D. Kokane, Sachin D. Babar, and Parikshit N. Mahalle. 2021. Word sense disambiguation for large documents using neural network model. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 1–5.
[35] Ivan Kratchanov and Detelin Luchev. 2024. Gamified crowdsourcing in support of the semantic transformation of digital libraries. In *Proceedings of the International Conference on Computer Systems and Technologies 2024*. 113–118.
[36] Caterina Lacerra, Michele Bevilacqua, Tommaso Pasini, and Roberto Navigli. 2020. CSI: A coarse sense inventory for 85% word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8123–8130.
[37] Ying Liu, Genevieve B. Melton, and Rui Zhang. 2024. Exploring large language models for acronym, symbol sense disambiguation, and semantic similarity and relatedness assessment. *AMIA Summits on Translational Science Proceedings* 2024 (2024), 324.
[38] Edward Loper and Steven Bird. 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*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 63–70. https://doi.org/10.3115/1118108.1118117
[39] Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics* 47, 2 (2021), 387–443.

[40] Hlaudi Daniel Masethe, Mosima Anna Masethe, Sunday Olusegun Ojo, Fausto Giunchiglia, and Pius Adewale Owolawi. 2024. Word sense disambiguation for morphologically rich low-resourced languages: A systematic literature review and meta-analysis. *Information* 15, 9 (2024), 540. https://doi.org/10.3390/info15090540

[41] Ezgi Mert and Gokhan Dalkilic. 2009. Word sense disambiguation for Turkish. In *2009 24th International Symposium on Computer and Information Sciences*. IEEE, 205–210.

[42] Benedikt Morschheuser and Juho Hamari. 2019. The gamification of work: Lessons from crowdsourcing. *Journal of Management Inquiry* 28, 2 (2019), 145–148. DOI: https://doi.org/10.1177/1056492618790921

[43] Benedikt Morschheuser, Juho Hamari, Jonna Koivisto, and Alexander Maedche. 2017. Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies* 106 (2017), 26–43. DOI: https://doi.org/10.1016/j.ijhcs.2017.04.005

[44] 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 (2018), 219–237. DOI: https://doi.org/10.1016/j.infsof.2017.10.015

[45] Luis R. Murillo-Zamorano, José Ángel López Sánchez, and Carmen Bueno Muñoz. 2020. Gamified crowdsourcing in higher education: A theoretical framework and a case study. *Thinking Skills and Creativity* 36 (2020), 100645. DOI: https://doi.org/10.1016/j.tsc.2020.100645

[46] Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)* 41, 2 (2009), 1–69.

[47] Christopher Nguyen, Matthew L. Jensen, Alexandra Durcikova, and Ryan T. Wright. 2021. A comparison of features in a crowdsourced phishing warning system. *Information Systems Journal* 31, 3 (2021), 473–513.

[48] Timo Nummenmaa, Samuli Laato, Philip Chambers, Tuomas Yrttimaa, Mikko Vastaranta, Oğuz Turan Buruk, and Juho Hamari. 2024. Employing gamified crowdsourced close-range sensing in the pursuit of a digital twin of the earth. *International Journal of Human–Computer Interaction* 40 (2024), 1–17. https://doi.org/10.1080/10447318.2024.2352922

[49] Zeynep Orhan and Zeynep Altan. 2005. Determining senses for word sense disambiguation in Turkish. In *IEC (Prague)*. 187–192.

[50] Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering* 13, 2 (2007), 137–163.

[51] Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation* 46, 2 (2012), 219–252. https://doi.org/10.1007/s10579-011-9155-4

[52] Rebecca J. Passonneau, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *LREC*.

[53] John Prpić, Prashant P. Shukla, Jan H. Kietzmann, and Ian P. McCarthy. 2015. How to work a crowd: Developing crowd capital through crowdsourcing. *Business Horizons* 58, 1 (2015), 77–85.

[54] Anna Rumshisky, Nick Botchan, Sophie Kushkuley, and James Pustejovsky. 2012. Word sense inventories by non-experts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, 4055–4059. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/1014_Paper.pdf

[55] Martijn J. Schuemie, Jan A. Kors, and Barend Mons. 2005. Word sense disambiguation in the biomedical domain: An overview. *Journal of Computational Biology* 12, 5 (2005), 554–565.

[56] Dilara Torunoğlu Selamet and Gülşen Cebiroğlu Eryiğit. 2021. Veri Artırımı için Yarı-Denetimli Bağlamsal Anlam Belirsizliği Giderme. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi* 14, 1 (2021), 34–46. https://doi.org/10.54525/tbbmd.835744

[57] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 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*. Association for Computational Linguistics, Honolulu, Hawaii, 254–263. Retrieved from https://aclanthology.org/D08-1027

[58] Dilara Torunoğlu-Selamet, Arda İnceoğlu, and Gülşen Eryiğit. 2020. Preliminary investigation on using semi-supervised contextual word sense disambiguation for data augmentation. In *2020 5th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 337–342.

[59] Shashank Yadav, Rohan Tomar, Garvit Jain, Chirag Ahooja, Shubham Chaudhary, and Charles Elkan. 2024. Gamified crowd-sourcing of high-quality data for visual fine-tuning. arXiv:2410.04038. Retrieved from https://arxiv.org/abs//2410.04038

[60] Chun-Xiang Zhang, Ya-Li Shao, and Xue-Yao Gao. 2023. Word sense disambiguation based on RegNet with efficient channel attention and dilated convolution. *IEEE Access* 11 (2023), 130733–130742.

Received 8 January 2025; revised 18 August 2025; accepted 27 August 2025