A Gamified Approach to Frame Semantic Role Labeling

Emily Amspoker
Carnegie Mellon University
eamspoke@andrew.cmu.edu

Miriam R. L. Petruck
International Computer Science Institute
miriamp@icsi.berkeley.edu

Abstract

Much research has investigated the possibility of creating **games with a purpose (GWAPs)**, i.e., online games whose purpose is gathering information to address the insufficient amount of data for training and testing of large language models (Von Ahn and Dabbish, 2008). Based on such work, this paper reports on the development of a game for **frame semantic role labeling**, where players have **fun** while using **semantic frames as prompts for short story writing**. This game will generate additional annotation for FrameNet and original content for annotation, **supporting FrameNet's goal of characterizing the English language in terms of Frame Semantics.**

1 Introduction

To create large-scale linguistic resources, linguistic database development projects have turned to crowd-sourcing. **Games with a purpose** (GWAPs), games whose purpose is to gather information, are a common approach to crowd-sourcing. GWAPs have been used for various tasks in computational linguistics, from anaphoric co-reference identification (Poesio et al., 2013) to word sense disambiguation (Lafourcade and Brun, 2017) to ontology population (Lafourcade et al., 2018). Informed by both the successes and shortcomings of previous games, this paper reports on the development of a game for frame semantic role labeling, ultimately for the FrameNet project, where players use semantic frames as prompts for short story writing. Upon completion of their stories, they also must provide semantic role annotation of their stories.

2 Background

This project focuses on the crowd-sourcing of frame semantic role labeling in FrameNet. This section provides background information about Frame Semantics (Fillmore, 1985) FrameNet, and crowd-sourcing to understand designing the game.

2.1 Frame Semantics

Frame Semantics (Fillmore, 1985), holds that each word or phrase in a text evokes a **semantic frame**, or structured background knowledge, that helps language users understand the text based on their experience (as a human, of a nationality, culture, etc). FrameNet calls a word or phrase that evokes a frame a **lexical unit** (**LU**), a pairing of a lemma and a frame. FrameNet treats each sense of a word or phrase with multiple meanings as different LUs based on their meaning in context.

For each LU, FrameNet records information about its dependents, the words or phrases that supply additional information about the participants in the frame, i.e., **frame elements** (**FE**s).

She	WALKED	along the road	for a while.
Self_mover	*Target*	Path	Duration

Figure 1: Semantic Role Annotation: `Self_Motion`

Consider the sentence *She walked along the road for a while.* in Figure 1. The LU **walk** evokes the `Self_Motion` frame, while the other parts of the sentence fill roles that give details about self-motion, including who is moving, how they are moving, and for how long they are moving.[1]

2.2 FrameNet

FrameNet (Ruppenhofer et al., 2016) is a research and resource development project based on the principles of Frame Semantics that provides information about the mapping between form and meaning for English, and documents its findings with corpus-based research. The FrameNet database holds frames, their descriptions, FEs, LUs, lexical entries with valence descriptions, and annotations of sentences that illustrate the use of each LU.

[1] By convention, FrameNet frames appear in `teletype font`, frame element names appear in SMALL CAPS; and in the prose, *italicized text* are example sentences.

Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances), pages 37 - 42
December 8, 2022 ©2022 Association for Computational Linguistics

FrameNet analysts create annotations in a two-step process. The first step is **frame disambiguation**. Polysemous words can exist in multiple frames; determining which frame a word evokes is critical. For example, the word **about** exists in the `Topic` frame, as in *This book is mostly **about** particle physics*, and the `Proportional_quantity` frame, as in *It took **about** three hours*.

FrameNet analysts determine which frame a word evokes, then label the FEs of that frame on the parts of the sentence (i.e., syntactic constituents) to which they correspond. Labeling is annotating; doing so automatically is **semantic role labeling** (**SRL**). The example in Figure 1 requires labeling *she* as the SELF_MOVER FE, *along the road* as the PATH FE, and *for a while* as the DURATION FE.

To date, FrameNet lexicographers have annotated example sentences **manually**, a resource-intensive activity that necessarily limits the amount of training data that the project has produced. Gamifying frame semantic SRL is but one effort to incorporate automatically produced annotation for the project. See (Pancholy et al., 2021) for another potential approach to automate annotation in FrameNet.

2.3 Crowdsourcing

Since large language models require massive amounts of data, which do not exist for FrameNet, the project has sought a variety of different ways to bolster the number of gold-standard annotated example sentences.

For example, Hong and Baker (2011) used crowdsourcing for frame disambiguation. On Amazon Mechanical Turk, a platform where *Turkers*, crowd-workers, perform small tasks (Human Intelligence Tasks) for a small amount of money. Workers had to choose the frame for a given target word based on its use in context. After filtering based on agreement, this method of collecting data yielded results that were approximately 86%-96% accurate.

While the crowd work approach works well for a multiple-choice task like frame disambiguation, semantic role labeling is more complex and requires a different approach, for example, that of **GWAPs**. Crowdsourcing efforts using GWAPs began in the early 2000s and mostly included simple labeling and image recognition tasks. For example, the ESP Game attracted over 200K players and created over 50 million labels four years after its release (Von Ahn and Dabbish, 2008).

Other GWAPs have completed linguistic annotation tasks: PackPlay, developed for semantic annotation, focused on Named Entity Recognition (Green et al., 2010). Similarly, Phrase Detectives, another successful GWAP used a reading comprehension game to identify anaphoric co-reference (Poesio et al., 2013). These games were successful in creating resources of a similar quality to traditionally generated data. However, many of these examples and others such as JeuxDesMots (Lafourcade et al., 2018), OntoGalaxy (Krause et al., 2010), and Zombilingo (Fort et al., 2014), focused on tasks less challenging than frame-semantic role labeling.

QANom (Klein et al., 2020) illustrates crowdsourcing for SRL, which gathered data for NomBank (Meyers et al., 2004) as microworkers answered questions about filling semantic roles via a question-and-answer format. Similarly, VerbCorner (Hartshorne et al., 2013), a game to crowdsource SRL for VerbNet (Kipper et al., 2000) had users read sentences with a sci-fi backstory and answer multiple-choice questions about the sentence. This approach required choosing specific verbs and crafting stories around them. If applied to SRL for FrameNet, the question-and-answer format would limit the ability to crowdsource data for numerous frames because rephrasing FrameNet data into comprehensible questions for the average player is terribly time-consuming.

Other fields boast examples of more complex tasks through gamification, such as *Foldit*, where players fold three-dimensional protein chains. This game generated enough high-quality data that players received credit as authors on several papers about the structure of various proteins (Khatib et al., 2011). Since online micro-working services can perform mechanical tasks, researchers have called for the creation of a new generation of GWAPs, where players complete complex tasks for a meaningful cause (Tuite, 2014).

3 The FrameGame

3.1 Principles

A literature review on designing GWAPs helped to establish four principles to guide our design.

1. The game's purpose must be **transparent** so players connect to its cause (Krause, 2013).
2. The game must have **skilled** tasks that highlight player creativity (Tuite, 2014).
3. The game must be **social** with an active community of members (Lafourcade et al., 2018).

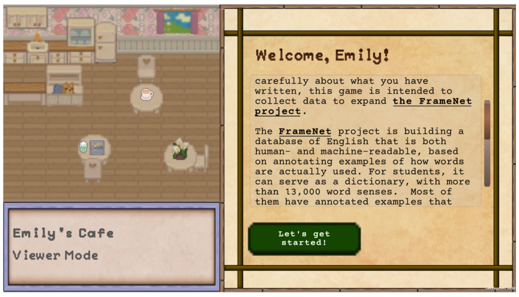
4. If the game has orthogonal game elements, they must be **specifically aligned** with a stated goal (Bonetti and Tonelli, 2021).

3.2 Game Format

Given these core ideas, we created the FrameGame, where players use semantic frames as writing prompts. Along with creating annotated data for FrameNet, the game makes players brainstorm, think critically, and receive feedback on stories inspired by the frames that serve as prompts.

The game begins with a **start screen**. Players log in using the Facebook Gaming API, as FrameNet collects annotations through Facebook's unique player identifier. Once logged in, players can choose to read information about the game and FrameNet. Upon agreeing, players must agree to the terms and conditions, which state that they retain the rights to their creative work and allow FrameNet to use their text and annotation.

Next, players can navigate away from this page or read more about the FrameNet project. Also, the description of the project includes links to FrameNet's website,[2] as shown in Figure 2. The start screen is key for transparency about the purpose of the game and gives players the opportunity to engage with FrameNet further.



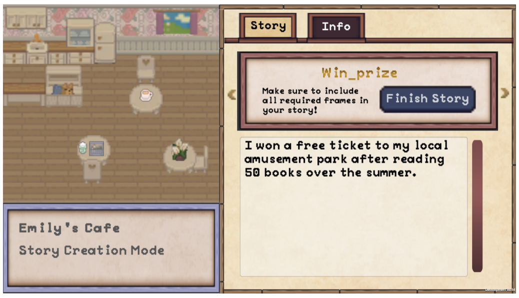
carefully about what you have written, this game is intended to collect data to expand **the FrameNet project**.

The **FrameNet** project is building a database of English that is both human- and machine-readable, based on annotating examples of how words are actually used. For students, it can serve as a dictionary, with more than 13,000 word senses. Most of them have annotated examples that

Figure 2: FrameNet Information Screen

If players choose to click on the coffee cup icon in the cafe in the left corner, they see the **story creation screen**, which makes up most of the gameplay. Likely players are already familiar with the concept of practicing their writing via social platforms that provided prompts for writing, e.g. Reddit's Writing Prompts. The FrameGame uses a single textbox where users enter one or more sentences to create a story, based on the LUs of the frame under consideration.

While writing, players view a list of frames using LUs from those frames (Figure 3). Also, they read further information about each frame by clicking on the **Info** tab; doing so displays annotated example sentences, lists of LUs that evoke the frame, and definitions of both the frame itself and FEs.

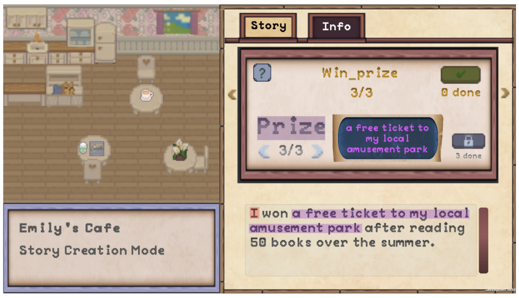


I won a free ticket to my local amusement park after reading 50 books over the summer.

Figure 3: Writing Screen

These annotated example sentences, LUs. etc., exemplify how annotation for a given frame actually works. Similarly, players can take advantage of the opportunity to study and internalize the frame definition, including its FEs and their definitions.

After successfully completing their stories, players press a button to lock the text and begin the annotation part of the gameplay (Figure 4). Players must highlight the frame-evoking LU and the FEs of the given frame.



I won a free ticket to my local amusement park after reading 50 books over the summer.

Figure 4: Annotation Screen

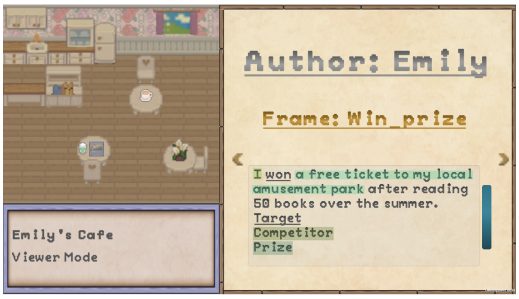
To ensure consistent annotation, the game places some restrictions on players during the annotation process: players must select a lexical unit from the list via the Info tab (available for reference during annotation), and they must label the FEs in the phrase or sentence under consideration.

Players can also access a screen for **viewing other players' work** (Figure 5) by clicking on the computer icon in the café. At present, this screen shows individual annotated sentences, their author, the frame, and its FEs.

3.3 Unexpected Findings

The FrameGame has not yet been deployed; the collection of player stories and sentence annota-

[2] https://framenet.icsi.berkeley.edu/fndrupal/



Author: Emily

Frame: Win_prize

I won a free ticket to my local amusement park after reading 50 books over the summer.

Target

Competitor

Prize

Emily's Cafe

Viewer Mode

Figure 5: Viewing Others' Work Screen

Issue	Frame(s)
Incorrect Exemplars	`Transition_to_a_state`
Incorrect FE Labels	`Arraignment` `Expensiveness`
Missing all LUs	`Influencing_potential`

Table 1: Problematic Issues and Frames

tion can only begin after the release of the game. However, even in the development phase, the game contributed to improving the FrameNet database. During the testing of gameplay (in the development phase), the game parsed the XML for a given frame to JSON format for display. Whenever the code threw an error, it provided a link to FrameNet's frame index, and sent FrameNet an error report.

Parse errors originating from the game become an issue in the game's open access Github repository[3] and later corrected. Such parsing also detects and facilitates the correction of errors in FrameNet data. Importantly, the game will parse FrameNet's XML when anyone plays the game, i.e., not only during the development of the game.

The restrictions on gameplay (Section 3.2) also highlight issues with missing targets or confusing examples. Table 1 displays these issues and their respective frames. The subsections that follow here discuss the errors.

3.3.1 Incorrect Exemplars

FrameNet includes exemplar sentences in the definitions of FEs in a frame.[4] Although the exemplar sentences in the FE definitions for `Transition_to_a_state` included *become*.v, that lemma is not listed as such in the frame itself. Actually, FrameNet characterizes *become.v* in the `Becoming` frame, which inherits from `Transition_to_a_state`. The investigation of the frame determined the exemplar sentence error, which the FrameNet team has since corrected.

3.3.2 Incorrect FE Labels

In both the `Arraignment` and `Expensiveness` frames, annotations for the example sentences included incorrectly placed tags or missing Frame Element names. As a result, the game could not correctly parse the examples to display to the player. FrameNet corrected these FE tags.

3.3.3 Missing all LUs

The `Influencing_potential` frame had no LUs listed, leaving no possibility for players to create annotated sentences based on the frame. While FrameNet holds valid **non-lexical** frames to maintain the integrity of the frame hierarchy, `Influencing_potential` is not one of them; the frame must have LUs in its XML file.

The unexpected exposure of errors and the need for corrections to the FrameNet database during the testing of the game showcase the potential of the game both to identify and facilitate the correction of pre-existing errors in the FrameNet database.

4 Conclusion and Future Work

4.1 Game Improvements

We must implement several features before deploying the game, including the following: (1) adding an interactive tutorial to ensure that players fully understand how to write stories and annotate sentences before beginning to play the game, (2) building a separate database to store player data and these crowd-sourced annotations, (3) implementing a points system, and (4) creating a system for verification and correction of annotations. We will detail these steps in the following paragraphs.

In its current unfinished state, the game only includes written instructions. We plan to add an interactive tutorial where users annotate an example sentence and receive feedback on the accuracy of their annotations.

Before declaring the game available to the general public, we also must ensure that the player-produced data remains separate from FrameNet's existing data. Game players do not possess the same expertise as highly-trained FrameNet annotators; mixing the two types of data before checking the quality of player-produced annotations is not desirable. We must store player-annotated sentences and player points to provide an accurate record of players' achievements, as well as other data too, such as average annotation time and quality, descriptions of which appear below.

[3] `https://github.com/eamspoker/FrameGameAssets`

[4] See, for example, the FE definitions for `Becoming`.

We will implement a points system based on the number of frames for which players provide annotation. The goal of this points system is to motivate players to produce more annotations. As a result, players will receive a small number of points each time they annotate a sentence. Based on previous game theoretic approaches to GWAP design (Ghosh, 2013), this choice might cause players to create rushed or incorrect annotations simply to earn points. To prevent this scenario from occurring, players will receive a greater point reward than their initial reward once their annotation is deemed correct.[5]

Finally, numerous GWAPs, such as PackPlay (Green et al., 2010), Phrase Detectives (Poesio et al., 2013), and Jeux Des Mots (Lafourcade et al., 2018), include verification and correction measures in the game itself, or in the form of another game. We would filter out user annotations by drawing from the user data collected in the FrameGame database. Additionally, we want players to read each others' stories and to suggest revisions or corrections for others' annotations. Combining these recommendations and the original annotation may result in a more accurate final annotation.

4.2 User Study

After deploying the game, we will advertise its availability to several different groups, including the FrameNet mailing list, and online writing communities, like Reddit's r/WritingPrompts, and indie game enthusiasts on websites, such as Itch.io. After collecting these data, we will use a combination of agreement based on both the in-game verification methods (section 4.1) and formal analyses of the player-generated sentences for FrameNet analysts to determine the quality of the annotations.

Since the envisioned user study will only occur after the game has been deployed, we will determine the methods for filtering annotations and the strategy for evaluating the quality of annotations as the start of the study draws near.

We believe that this game has much potential to contribute to FrameNet. By crowd-sourcing both the creation of new example sentences and their annotation, the game will help FrameNet to capture language as it exists "in the wild" through the lens of frame semantics.

Acknowledgements

The authors acknowledge the contributions of Collin F. Baker to this work; he advised the lead author during the development of the game. Additionally, the International Computer Science Institute (**ICSI**) hosted Amspoker during her stay in Berkeley. The authors also thank the NSF and the Summer Undergraduate Program in Engineering Research at Berkeley (SUPERB) for their generous support of this research.

References

Federico Bonetti and Sara Tonelli. 2021. Measuring orthogonal mechanics in linguistic annotation games. *Proc. ACM Hum.-Comput. Interact.*, 5(CHI PLAY).

Charles J Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di semantica*, 6(2):222–254.

Karën Fort, Bruno Guillaume, and Hadrien Chastant. 2014. Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, page 2–6, New York, NY, USA. Association for Computing Machinery.

Arpita Ghosh. 2013. Game theory and incentives in human computation systems. In Pietro Michelucci and Peng Dai, editors, *Handbook of Human Computation*, pages 725–742. Springer.

Nathan Green, Paul Breimyer, Vinay Kumar, and Nagiza Samatova. 2010. PackPlay: Mining semantic data in collaborative games. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 227–234, Uppsala, Sweden. Association for Computational Linguistics.

Joshua K. Hartshorne, Claire Bonial, and Martha Palmer. 2013. The VerbCorner project: Toward an empirically-based semantic decomposition of verbs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1438–1442, Seattle, Washington, USA. Association for Computational Linguistics.

Jisup Hong and Collin F. Baker. 2011. How good is the crowd at "real" WSD? In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 30–37, Portland, Oregon, USA. Association for Computational Linguistics.

Firas Khatib, Seth Cooper, Michael D. Tyka, Kefan Xu, Ilya Makedon, Zoran Popović, David Baker, and Foldit Players. 2011. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, 108(47):18949–18953.

[5] We envision involving a highly trained member of the FrameNet team to make such decisions.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. AAAI Press.

Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Markus Krause. 2013. Designing systems with homo ludens in the loop. In *Handbook of Human Computation*, pages 393–409. Springer.

Markus Krause, Aneta Takhtamysheva, Marion Wittstock, and Rainer Malaka. 2010. Frontiers of a paradigm: Exploring human computation with digital games. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, page 22–25, New York, NY, USA. Association for Computing Machinery.

Mathieu Lafourcade and Nathalie Le Brun. 2017. Ambiguss, a game for building a sense annotated corpus for French. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Mathieu Lafourcade, Alain Joubert, and Nathalie Brun. 2018. The jeuxdemots project is 10 years old: What we have learned. In *Proceedings of the LREC 2018 Workshop "Games and Gamification for Natural Language Processing (Games4NLP)*, Miyazaki (Japan).

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 24–31, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ayush Pancholy, Miriam R L Petruck, and Swabha Swayamdipta. 2021. Sister help: Data augmentation for frame-semantic role labeling. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 78–84, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1).

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. ICSI: Berkeley.

Kathleen Tuite. 2014. Gwaps: Games with a problem. In *FDG*.

Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM*, 51(8):58–67.