LREC 2018 Workshop

Games4NLP
Games and Gamification for Natural Language Processing

PROCEEDINGS

Edited by

Jon Chamberlain, Udo Kruschwitz, Karën Fort, Christopher Cieri

ISBN: 979-10-95546-10-8
EAN: 9791095546108

7 May 2018

Proceedings of the LREC 2018 Workshop
"Games and Gamification for Natural Language Processing (Games4NLP)"

7 May 2018 – Miyazaki, Japan

Edited by Jon Chamberlain, Udo Kruschwitz, Karën Fort, Christopher Cieri

http://anawiki.essex.ac.uk/dali/games4nlp/

Organising Committee

- Jon Chamberlain, University of Essex, UK
- Udo Kruschwitz, University of Essex, UK
- Karën Fort, Université Paris-Sorbonne, France
- Christopher Cieri, Linguistic Data Consortium, USA

Programme Committee

- Richard Bartle, University of Essex, UK
- Johan Bos, University of Groningen, Netherlands
- Eric de la Clergerie, INRIA, France
- James Fiumara, Linguistic Data Consortium, US
- Ivan Habernal, Technische Universität Darmstadt, Germany
- Frank Hopfgartner, University of Glasgow, UK
- Michael Meder, TU Berlin, Germany
- Mathieu Lafourcade, LIRMM, France
- Verena Lyding, EURAC, Italy
- Lionel Nicolas, EURAC, Italy
- Massimo Poesio, University of Essex, UK
- Pontus Stenetorp, University College London, UK

Preface

The creation of Language Resources required for research in Natural Language Processing often relies on tedious manual labor, for example, the wide variety of annotations of raw human language data required to train and evaluate automatic machine learning algorithms. A recent trend to perform these tasks is the use of crowdsourcing techniques, i.e., obtaining annotations from anonymous crowd workers via an open call (Howe, 2008). Although research indicates that such techniques can be useful (Snow et al., 2008), they motivate users through micropayments thus may not be suitable for large-scale efforts (Poesio et al., 2013; Fort et al., 2011). A promising approach to overcome this challenge is by employing games and gamification methods to motivate users.

Games can be used to make tasks traditionally performed by paid workers more enjoyable and entertaining. The first, and perhaps most successful, Game-With-A-Purpose (GWAP) was The ESP Game which attracted over 200,000 players who produced over 50 million labels (von Ahn, 2006). Since then GWAPs have been developed for numerous tasks, including image and video annotation, natural language processing, biomedical research and search refinement (Lafourcade et al., 2015). Several GWAPs have attempted annotation of human language such as Phrase Detectives (Poesio et al., 2013), a game developed to collect data about anaphoric reference, Zombilingo (Guillaume et al., 2016), a GWAP for dependency syntax annotation, and Wordrobe(Venhuizen et al., 2013), a unified attempt to solve numerous linguistic tasks including part-of-speech tagging, named entity tagging, co-reference resolution, word sense disambiguation and compound relations. GWAPs integrated into social networking sites such as Sentiment Quiz (Rafelsberger and Scharl, 2009), to annotate sentiment in US elections, DigiTalkoot's games Mole Hunt and Mole Bridge, to digitise old Finnish documents (DigiTalkoot, 2012), and RoboCorp (Dziedzic, 2016), a machine translation game, show the potential for large-scale participation using game models where collected data is considered revenue (F2P models).

Such games leverage elements commonly found in game design; they tend to be graphically rich and give the player an experience of progression through the game by scoring points, being assigned levels and recognising their effort. NLP projects also have the potential to tap into the motivational drivers of games by using simple constructs (such as leaderboards, badges and high scores), even if the task is not presented as a game (Deterding et al., 2011).

The Games4NLP workshop aims to promote and explore the possibilities for research and practical applications of using games and gamification for the creation of language resources for Natural Language Processing. The main objective is to provide a forum for researchers and practitioners to discuss and share ideas regarding how the NLP research community can benefit from using game and gamification strategies.

References

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). From game design elements to gamefulness: Defining gamification. In Proceedings of MindTrek11, pages 9-15.

DigiTalkoot. (2012). http://www.digitalkoot.fi/index_en.html. (Last accessed 12 December 2016).

Dziedzic, D. (2016). Use of the Free to Play model in games with a purpose: the RoboCorp game case study. Bio-Algorithms and Med-Systems, 12(4):187-197.

Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon Mechanical Turk: Gold mine or coal mine? Computational Linguistics (editorial), 37:413-420.

Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax. In Proceedings of COLING16, Japan.

Howe, J. (2008). Crowdsourcing: Why the power of the crowd is driving the future of business. Crown Publishing Group.

Lafourcade, M., Joubert, A., and Le Brun, N. (2015). Games with a Purpose (GWAPS). John Wiley & Sons.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase Detectives: Utilizing collective intelligence for internet-scale language resource creation. ACM Transactions on Interactive Intelligent Systems, 3(1):1-44.

Rafelsberger, W. and Scharl, A. (2009). Games with a purpose for social networking platforms. In Proceedings of the 20th ACM Conference on Hypertext and Hypermedia.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast - but is it good?: Evaluating non- expert annotations for natural language tasks. In Proc EMNLP08.

Venhuizen, N., Basile, V., Evang, K., and Bos, J. (2013). Gamification for word sense labeling. In Proceedings of IWCS13.

von Ahn, L. (2006). Games with a purpose. Computer, 39(6):92-94.

Programme

09.00 – 09.10	Introduction by Workshop Chair
	Session 1
09.10 – 09.50	Ivan Habernal Did you also spend the last weekend playing your NLP game? (invited talk)
09.50 – 10.10	Hend Al-Khalifa, Hadil Faisal and Rawan N. Al-Matham Faseeh: A Serious Game for Arabic Synonym Acquisition
10.10 – 10.30	Chris Madge, Massimo Poesio, Udo Kruschwitz and Jon Chamberlain Testing TileAttack with Three Key Audiences
10.30 – 10.50	Break
	Session 2
10.50 – 11.10	Sarah Ita Levitan, James Shin, Ivy Chen and Julia Hirschberg LieCatcher: Game Framework for Collecting Human Judgments of Deceptive Speech
11.10 – 11.30	Karën Fort, Mathieu Lafourcade and Nathalie Le Brun Cheap, fast and good! Voting Games with a Purpose
11.30 – 11.50	Alain Joubert, Mathieu Lafourcade and Nathalie Le Brun The JeuxDeMots Project is 10 Years Old: what Assessments?
	Project Updates
11.50 – 12.15	Verena Lyding and Lionel Nicolas enetCollect COST Action
12.15 – 12.30	Christopher Cieri Lingo Boingo
12.30 – 12.50	**Panel Discussion**
12.50 – 13.00	Concluding Remarks by Workshop Chair

Table of Contents

Faseeh: A Serious Game for Arabic Synonym Acquisition

Hend S. Al-Khalifa[1], Hadil Faisal[2] and Rawan N. Al-Matham[3]
[1,2,3]Information Technology Department, College of Computer and Information sciences
King Saud University, Riyadh, Saudi Arabia
[1]hendk@ksu.edu.sa, {[2]hadeelalf|[3]r.almatham}@gmail.com

Abstract

Arabic word synonyms are highly common and valued by Arabic speakers as part of a good writing style. There have been many games that assist Arabic learners, from vocabulary and spelling acquisition to learning grammatical rules and sentence structure. However, no games were found targeting synonym acquisition. This paper presents Faseeh, a serious game that assists Arabic speakers acquire word synonyms to enrich their linguistic and expressive skills. The game employs different gamification techniques; these techniques were implemented to improve user motivation and enhance learning experience. Faseeh statistics and evaluation showed that it has delivered an educational content in a novel engaging manner.

Keywords: Synonym, Arabic Language, Serious Games, Gamification, Language Games, NLP

1 Introduction

Arabic is a Semitic language spoken by nearly 500 million people around the world and one of the official UN languages. Like any language, Arabic has its grammar, spelling, and pronunciation; yet it has its own characteristics which made it distinctive. Arabic is read and written from right to left (except numbers); its alphabet consists of 29 spoken letters, and 36 written characters. Also, Arabic is a morphologically rich language. Most Arabic words are derived from a 3-letter root that are highly generative (Diab, 2004). In addition, words in Arabic might change meaning depending on the context. Synonyms are also commonly used in Arabic, since variety in expression is valued by Arabic speakers as part of skillful writing style (Xu et al., 2002).

Some mobile tools and games have been developed for learning Arabic. These games can be referred to as serious games. A serious game defines those that positively impact users' skills. It is not limited to learning and computer games, but extends to those used for medical treatments and marketing tools (Toma et al., 2017). However, this paper focuses on a mobile game called "Faseeh فصيح" for acquiring Arabic synonyms.

Arabic learning games studied in the literature include LingoSnacks (Erradi et al., 2013), Easy Arabic (Aljunid et al.,2014), and U-Arabic (Alobaydi et al., 2016). Nevertheless, these games focused on teaching spelling, vocabulary or grammatical rules. No application has been found to teach the vast synonyms for Arabic words.

Faseeh game has the potential to assist Arabic native speakers and learners by:

- Encouraging independent learning of Arabic synonyms,
- Providing learners with an engaging game that enriches their linguistic skills, and
- Offering learning activities that allow learners to benefit from their idle times.

The rest of the paper is organized as follows: Section 2 discusses serious games. Section 3 describes Faseeh game design. Section 4 presents Faseeh statistics. Finally, the paper concludes with future directions.

2 Serious Games

Serious games have emerged due to widespread use of the internet and games, and the need to provide engaging training and educational content. Such transition in delivering education and training includes the use of metaphors of games, or 'gamification' of learning (de Freitas and Liarokapis, 2011). Gamification can be defined as using gaming techniques in non-game contexts as to improve engagement and impact behavior or attitude towards learning (Landers, 2014).

Gamification mechanisms include: progression, investment, and cascading information theory. Progression is allowing the user to reach success incrementally. Investment is allowing the user to feel pride of his/her work. Cascading information theory is by continuously unlocking information (Knewton, 2012).

Many of these mechanisms have been employed in Faseeh. Progression is applied by unlocking levels and gaining points. Investment is applied by displaying a leaderboard that gives public recognition to top users in each game mode. Cascading information theory can be applied by obliging the user to tackle problems in a limited time frame. This is implemented in Faseeh through the time limits that are imposed on each level.

3 Faseeh Design

Faseeh is a serious game that targets Arabic native speakers as well as Arabic learners to enhance and enrich their vocabulary. Basically, the game displays a word/phrase and its corresponding synonyms, from which the user should choose one or more of them. These synonyms were taken from two classical books for Arabic synonyms, namely: Fiqh Allughah/فقه اللغة by Althaalibi and Alalfaz Alkitabiyah/الألفاظ الكتابية, by Alhamazani.

The game database consists of 96 word/phrase each of which has three correct synonyms (in total 288 word/phrase). A set of words/phrases along with their synonyms are grouped into 24 levels based on the books' recommendations. Also, the game requires registration in order to record the user progress.

3.1 Game Options

Faseeh offers three game options: (1) main game, (2) challenge Faseeh and (3) network challenge, as shown in figure 1.

Figure 1: The game main page with the following buttons: (1) Main Game (2) Challenge Faseeh (3) Network Challenge (4) About the app (5) More apps by iWAN (6) Share the app (7) Rate in the store (8) Extra options (leaderboard – settings – about Faseeh) and (9) Log out.

3.1.1 Main Game

The main game option, as shown in figure 2, contains 24 levels, each level has four questions and lasts for 60 seconds. Each question displays a word/phrase, and the user is prompted to choose three synonyms from six presented choices. In this mode, the user has three helping methods (aka helpers) which are: ask a friend, get extra time or omit a choice. If the user used one of the help methods his/her collected points will be deduced.

3.1.2 Challenge Faseeh

The second option (figure 3) is a novel contribution to language learning games. It allows the user to play against a virtual character named Faseeh. Faseeh level will be assigned randomly at the beginning of the option. The level will determine the questions that this character can answer correctly, and it will answer the rest randomly. Each question displays a word/phrase and four choices. Only one of the choices is correct, so whoever chooses the correct answer before time ends gains the point.

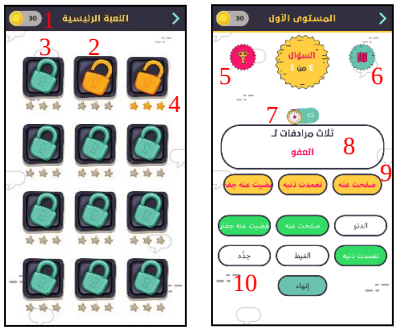


Figure 2: Main Game mode (left): (1) collected points (2) unlocked level (3) locked level (4) stars earned in the level; (right) (5) helpers (6) instructions (6) timer (8) question (9) correct answers (10) user answers.

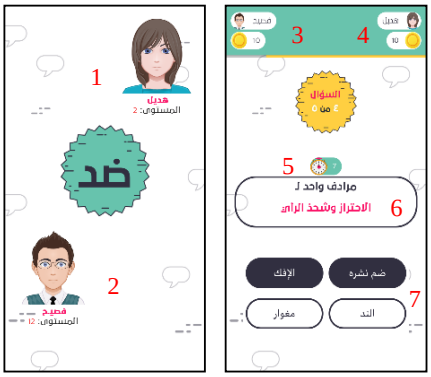


Figure 3: Challenge Faseeh option (left) (1) player's name and level (2) Faseeh character and the randomly assigned level; (right) (3) Faseeh character collected points (4) player's collected points (5) timer (6) question (7) choices.

3.1.3 Network Challenge

The last option, shown in figure 4, allows the user to play with remote users currently logged into the game. The game displays five questions. Every question shows a word/phrase with four choices, only one of which is correct. The player that answers correctly, within the time limit, gains points. The winner is the player who gains more points.

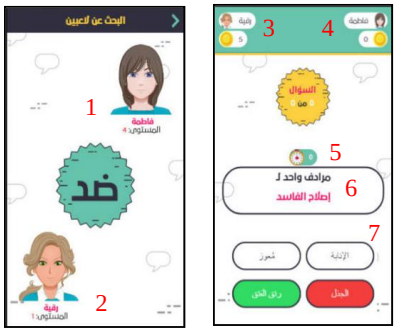


Figure 4: Network Challenge option (left) (1) player 1 name and level (2) player 2 name and level; (right) (3) player 1 collected points (4) player 2 collected points (5) timer (6) question (7) choices.

3.2 Leaderboard

Faseeh has a leaderboard that displays top users of each game option, as shown in figure 5. There is also a specific board for the player to display his/her achievements separately.

Figure 5: (Left) Game leaderboard, (Right) specific user board.

4 Faseeh Statistics

Faseeh game has gained popularity among people in Saudi Arabia after winning the grand prize of ALECSO (Arab League Educational cultural and Scientific Organization) for mobile applications development[1] in November 2017. The number of downloads exceeded 10K (since November 2017) in both Google play and Apple Store. It also got a rating of 4.8 out of 5 in Google Play and 4.9 out of 5 in Apple Store. The feedback people provided in the comments' section in the app stores were all positive and encouraging.

5 Faseeh Evaluation

Faseeh user evaluation has been conducted to measure players' satisfaction regarding the game. We developed an evaluation questionnaire based on (Göbel et al., 2013) serious games evaluation approach where they identified two main categories for the questionnaire: *user experience and game design*. The user experience has seven subcategories derived from it; we chose six subcategories that were suitable for Faseeh game. The selected categories were: *Positive Emotion, Negative Emotion, Motivation, Immersion, Flow and Arousal*. In addition, we added the challenge subcategory because it was an important aspect of the game. As for the game design, it has ten subcategories; we chose six of them, namely: *Effectance, Curiosity, Personalization, Interface, Feedback and Social Needs*.

Our final questionnaire contained 14 statements (see Table 1), one statement for each subcategory except for the game interface category which had two statements to measure the quality of the interface design and its colors. In addition, we used Likert scale to evaluate each statement. The scale ranged from 1 to 10; where 1 means Totally Disagree and 10 means Totally Agree. Moreover, the questionnaire contained some open-ended questions to elicit users' suggestions for improving the next release of Faseeh.

Table 1: Faseeh Evaluation Questionnaire

No	Cat.	Sub-Category	Question
1	User Experience	Positive Emotions	I have fun when playing Faseeh.
2		Negative Emotions	I didn't feel bored when playing Faseeh.
3		Motivation	I feel excited when playing Faseeh.
4		Immersion	I felt engaged while playing Faseeh.
5		Flow	The transition was smooth between the different levels of Faseeh.
6		Arousal	I feel excited when I win and move forward in Faseeh.
7		Challenge	I feel challenged while thinking about solutions for each level in Faseeh.
8	Game Design	Effectance	Faseeh contributed in improving my Arabic vocabulary.
9		Curiosity	I felt curious enough to complete the levels of Faseeh and learn new vocabulary.
10		Personalization	I can enter a defined name and a character that represents me when I start Faseeh.
11		Interface Design	I find the design of Faseeh interface and its icons beautiful and attractive.
12		Interface Colors	I think that the colors of Faseeh interface are comfortable.
13		Feedback	Faseeh provides me with feedback regarding my choices, whether right or wrong.
14		Social Needs	I enjoy playing with others through "the challenge network".

We received responses from 30 players (ages between 10 and 50 years old). The calculated responses' mean and Standard Deviation (S.D.) are reported in Table 2.

Table 2: Questionnaire Results

Category	Mean	S.D.
Positive Emotions	8.07	2.18
Negative Emotions	7.00	2.39
Motivation	7.53	2.19
Immersion	7.67	2.10
Flow	8.53	2.06
Arousal	8.37	2.17
Challenge	8.33	2.32
Effectance	8.13	2.03
Curiosity	7.53	2.81
Personalization	7.6	3.22
Interface Design	8.53	2.06
Interface Colors	8.57	2.01
Feedback	8.53	2.24
Social Needs	6.63	3.18

[1] http://award.alecsoapps.com/

The overall evaluation of the user experience was satisfactory; the average score for all subcategories was between 6.63 (Min) and 8.57 (Max), which is considered very good. The emotions subcategory answers (mean 8.07 S.D. 2.18) showed that most of the players enjoyed playing the game. On the other hand, the negative emotion subcategory which measured how often the players were bored while playing Faseeh showed that most players did not feel bored while interacting with Faseeh (mean 7.00 S.D. 2.39). As for the motivation and immersion subcategories, most of the players were motivated and engaged while playing Faseeh. The mean for the motivation and immersion was 7.53 (S.D. 2.19) and 7.67 (S.D. 2.10) respectively. Moreover, the mean of the flow subcategory was 8.53 (S.D. 2.06), which indicates that the transition between the different levels in Faseeh was smooth. Similarly, the arousal subcategory had a mean of 8.37 (S.D. 2.17), which is considered high. Most of the players felt excited after winning in the game and were motivated to proceed to the next level. Finally, the mean of the challenge subcategory was 8.33 (S.D. 2.32), which ensures that the game was challenging. However, the challenge level was not the same for all players based on their age. This might be attributed to the different linguistic backgrounds of the players.

As for the game design category, the first subcategory measured the Effectance of Faseeh. Specifically, how effective was it in enhancing the Arabic vocabulary of the player. This subcategory mean was 8.13 (S.D. 2.03), which showed that it affected the players' vocabulary positively. On the other hand, the curiosity subcategory mean was 7.53 (S.D. 2.81) which might indicate that some players were interested to move to next levels. As for the personalization subcategory, which measured the players' satisfaction regarding the personalization option, its mean was 7.6 with the highest S.D. equal to 3.22, this indicates that some players were somehow satisfied with the personalization options. Yet, others wanted to have more options and characters based on their answers in the open-ended questions (this will be reported later). For the interface design subcategory, we have two statements regarding their satisfaction about the interface design (mean 8.53), and colors (mean 8.57). This category gained the highest means, which showed that most players liked Faseeh's interface. Finally, the social needs subcategory represented the players' opinions regarding "the network challenge". It gained a mean of 6.63 (S.D. 3.18), which is considered the lowest mean in the questionnaire and indicates that some players were not satisfied with the network challenge. However, the high S.D. indicates that there was significant diversity between players' opinions, and we believe that it could be attributed to the availability of other players when playing the "network challenge" mode; therefore, they did not have the chance to try it.

At the end of the questionnaire, there were three open-ended questions that explored most of the features that were liked or disliked by the players. The answers indicated that, most players liked the game idea and its aim and they thought that there was a real need for it in the Arab world. In addition, they liked the quality, design and colors of Faseeh's interface. Moreover, they liked the different playing options that Faseeh has especially the network challenge. Most of the users said that playing Faseeh enhanced their Arabic vocabulary; however, there is still a need to add more levels and new vocabulary. In addition, there was a need to add different difficulty levels based on the players' background knowledge.

As for the future improvement of Faseeh, some players suggested improving the levels of Faseeh so that they should start with easy levels then increase the difficulty level gradually. Also, they suggested changing the interface color for each level to prevent the feeling of boredom. The number of the suggestions regarding network challenge indicated that the players liked the social aspect of the game and they wanted more chances to enjoy it. Finally, there was a suggestion regarding the need to add more modern Arabic vocabulary that can be used nowadays because the used vocabulary in the game was classical Arabic.

6 Conclusion

Faseeh is a serious game that aims to assist Arabic speaker in synonym acquisition. It has incorporated many gamification mechanisms to enhance learning experience. The game is available for download in both Apple Store and Google play by following this link: https://land.ly/faseeh.

7 Acknowledgements

This work has been supervised and funded by iWAN Research Group, King Saud University. We also extend our thanks to Fatimah Ahmad for programming Faseeh and Amany AlAmri for designing Faseeh interfaces.

8 References

Aljunid, S. A., Amin, M. A. M., and Sani, A. S. A. (2014, November). Scaffolded Arabic language mobile educational game. In *Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on* (pp. 1-6). IEEE.

Alobaydi, E. K., Mustaffa, N., Alkhayat, R. Y., and Arshad, M. R. H. M. (2016, November). U-Arabic: Design perspective of context-aware ubiquitous Arabic vocabularies learning system. In *Control System, Computing and Engineering (ICCSCE), 2016 6th IEEE International Conference on* (pp. 1-6). IEEE.

de Freitas, S., & Liarokapis, F. (2011). Serious Games: A New Paradigm for Education? In M. Ma, A. Oikonomou, & L. C. Jain (Eds.), *Serious Games and Edutainment Applications* (pp. 9–23). London: Springer London. https://doi.org/10.1007/978-1-4471-2161-9_2

Diab, M. (2004, September). The feasibility of bootstrapping an Arabic wordnet leveraging parallel corpora and an English wordnet. In *Proceedings of the Arabic Language Technologies and Resources, NEMLAR, Cairo.*

Erradi, A., Almerekhi, H., and Nahia, S. (2013, July). Game-based micro-learning approach for language vocabulary acquisition using LingoSnacks. In *Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on* (pp. 235-237). IEEE.

Giibel, S., Gutjahr, M., & Hardy, S. (2013). Evaluation of Serious Games. Serious Games and Virtual Worlds in

Education, Professional Development, and Healthcare, 105.

Knewton (2012). *The Gamification of Education*. Retrieved 31 December 2017, from https://www.knewton.com/infographics/gamification-education/.

Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & Gaming*, *45*(6), 752-768.

Toma, I., Alexandru, C.-E., Dascalu, M., Dessus, P., and Trausan-Matu, S. (2017). Semantic Taboo -- A Serious Game for Vocabulary Acquisition. *Romanian Journal of Human - Computer Interaction*, *10*(2), 241–256. Retrieved from https://hal.archives-ouvertes.fr/hal-01618507

Xu, J., Fraser, A., and Weischedel, R. (2002, August). Empirical studies in strategies for Arabic retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 269-274). ACM.

Testing TileAttack with Three Key Audiences

Chris Madge, Massimo Poesio, Udo Kruschwitz, Jon Chamberlain
Queen Mary University London, University Of Essex
{c.j.madge, m.poesio}@qmul.ac.uk {udo, jchamb}@essex.ac.uk

Abstract

The Game-With-A-Purpose (GWAP) approach has shown some success and promise in language resource collection. However, player recruitment and accuracy can be challenging. In this work, *TileAttack*, a GWAP designed to gather annotations for text segmentation, is presented to the online linguistic community, an indie gaming community and the crowdsourcing community. We evaluate the results of this experiment both through traditional accuracy measures and adapted metrics from Free-to-Play games. With the addition of a tutorial, we find a high level of recall is achieved from crowdsourced non-expert workers.

Keywords: Gamification; Crowdsourcing; HCI

1. Introduction

Many Natural Language Processing (NLP) tasks require large amounts of annotated text to train statistical models, or as a gold standard to test the effectiveness of NLP systems. These are often hand-annotated contributions (Palmer et al., 2005) using annotation tools. These annotation tasks may be carried out using pre-built annotation tools such as MMAX2 (Müller and Strube, 2006), web-based crowdsourcing focused WebAnno (Yimam et al., 2013), or the wiki style web-based *GMB Explorer* (Basile et al., 2012). However, those tools are aimed at expert annotators and require some understanding on the part of the user. Willing and inexpensive experts can be difficult to recruit. This process can be time consuming, expensive and tedious. Consequently, this requirement for annotated data remains an obstacle to progression for some NLP tasks. One proven method of reducing the time to gather the annotations is crowdsourcing (Snow et al., 2008). However, this doesn't scale very well. When attempting to build large corpora gamification can be cheaper (Poesio et al., 2013), provide more accurate results and better contributor engagement (Lee et al., 2013).

In this work, we look at gathering **mentions**. These are candidate entities for co-refererence that are usually detected in a co-reference pipeline in a step often referred to as Mention Detection. They are typically noun phrases, pronouns and named entities. Historically, the task of mention detection was rarely considered in isolation, but rather as a step in part of a pipeline for co-reference resolution (Peng et al., 2015). A rule-based approach, (e.g. pick all noun phrases (Haghighi and Klein, 2010)) was generally preferred with such systems usually aiming for high recall and compromise on precision, placing more confidence/importance on the co-reference resolution step (Kummerfeld et al., 2011) and being satisfied that incorrectly identified mentions will simply remain singletons which can be removed in post processing (Lee et al., 2011). However, this approach can result in a propagation of errors with singletons then being incorrectly identified as co-referent, particularly in the case of pleonastic entities (Lee et al., 2017). It has been pointed out by multiple researchers that this is a very important step for overall co-reference quality (Stoyanov et al., 2009; Hacioglu et al., 2005; Zhekova and Kübler, 2010). Recently, systems are now once again looking at machine learning approaches with the mention detection step being considered in isolation (Lee et al., 2017; Nguyen et al., 2016). This area is still identified as an area of challenge, particularly in under resourced languages (Soraluze et al., 2012) or domains, like biomedicine (Kim et al., 2011).

Games-with-a-Purpose (GWAPs) harness human effort as a side effect of playing a game (Von Ahn and Dabbish, 2008). GWAPs have been successful in many applications attracting large numbers of users to label datasets and solve real world problems (Lafourcade et al., 2015). Examples include *The ESP Game*, in which by playing, players contribute image labels (Von Ahn and Dabbish, 2004), and *FoldIt*, in which players solve protein-structure prediction problems (Cooper et al., 2010). In contrast, gamification has been described as "the use of game design elements in non-game contexts" (Deterding et al., 2011). Gamification has been very effective in motivating text labelling. For example, *Phrase Detectives* has been particularly effective in motivating participation in gathering anaphoric annotations (Poesio et al., 2013). However, there are limited examples of GWAPs for NLP. Creating a GWAP that produces annotations as a side effect, rather than applying gamification to motivate annotation, presents a greater challenge. The former requires mapping the task completely into a game, whilst the latter typically adds a layer of game-like themes and carefully selected motivational game mechanics. In exchange for this additional challenge, GWAPs have the potential for much higher player engagement.

One of the goals of gamified solutions is to provide a positive and engaging user experience. Designing an interface for an application can present multiple challenges. This is particularly evident in application for text annotation. Text often has complex properties which can be difficult to visualise and present in an easy to use interface. The aforementioned tools take different approaches, for example, to embedded and overlapping annotations. There is no standardised and accepted interface for text annotation tools. Borrowing ideas from game interfaces can reduce the barriers to reach a wider audience of non-expert users. Designing for motivation carries additional complexity.

Games such as *Puzzle Racer* have demonstrated the feasibility of inexpensively creating an engaging GWAP that

produces annotations. Furthermore, they report the annotations that are gathered are of a high quality and at a reduced cost compared with other methods (Jurgens and Navigli, 2014). However, such games have yet to achieve the player uptake or number of judgements comparable to GWAPs in other domains. GWAPs for annotation tasks often present additional unique challenges compared to those for image labelling and other similar tasks. For example, users can differentiate between image features easily, but not so easily with text features (Mason and Watts, 2010). The linguistic complexity of some text annotation tasks may not be immediately obvious or difficult to map into a game domain. Additionally, it may be challenging to find a representation that both entertains users and is easy to understand. *TileAttack* supports any text segmentation task with or without embeddings (e.g. noun-phrase embedding), that may be aligned, non-aligned or overlapping, making it broadly applicable to a variety of text annotation tasks including Named Entity Recognition, Information Extraction and Mention Detection.

In this work, we experiment with the GWAP *TileAttack*. [1] *TileAttack* is designed to gather mentions, a crucial step of the co-reference resolution pipeline which discovers potential referring expressions including noun-phrases and possessive pronouns (Lee et al., 2011). The following example shows the nested mentions enclosed in braces, (taken from the Phrase Detectives corpus (Chamberlain et al., 2016)) :

{*A Wolf*} *had been gorging on* {*an animal* {*he*} *had killed*}

In our previous work on testing game mechanics, we identified two additional important challenges with *TileAttack*: increasing player recruitment; and low annotation accuracy (Madge et al., 2017). This appears to be a challenge of effectively communicating the task to the players whilst retaining their interest. This is also a challenge in games. From studies in game design, the best approach is believed to be one that allows the player to play immediately, learning through a tutorial, without needing to read a manual (Sweetser and Wyeth, 2005). Naturally, traditional annotation tools, take a more utilitarian tool-like approach offering a manual and expecting a prior understanding of the task for which the tool will be used. *TileAttack* includes a game-like tutorial that plays similarly to an ordinary round but with more player feedback.

For Gamification and GWAPs to really achieve scale, they require communication of an arbitrarily complex task to a group of non-experts in a game setting. GWAPs are often tested against students from a department that have some interest or understanding in the task. In this experiment we ask if the current *TileAttack* is effective in the recruitment of non-experts and gathering accurate annotations with three distinct audiences: a linguistic community; a gaming community; and via crowdsourcing.

2. Related Work

The first Game With A Purpose was Von Ahn's *The ESP game*. This game was created to crowdsource image labels for web images, which may be used to train a supervised machine learning system. Human annotators play a game against a timer in which they were anonymously paired and rewarded scores for agreeing common labels to describe an image. In the interest of acquiring a comprehensive set of labels for each image, the game used a feature called *taboo words*. This resisted players contributing obvious image labels by displaying labels in a game as unavailable, once they had been contributed so many times. (Von Ahn and Dabbish, 2004)

The ESP game's design of rewarding based on agreement addresses the problem that an annotation task's latent correct labels are unknown by the system at the time the player is rewarded. Instead, given some input, it uses the agreement of multiple players output labels as a basis to determine whether points should be rewarded. This strategy has been described by Von Ahn as *output-agreement* (Von Ahn and Dabbish, 2008).

The GWAP concept was later applied in multiple fields to motivate player contribution including annotating text data for training NLP supervised learning systems. One notable example of a GWAP for text annotation is *Phrase Detectives*, in which players annotate and validate anaphora (Poesio et al., 2013). *Phrase Detectives* has gamification-like mechanics to motivate play such as points, leaderboards and levels, but also makes use of a game-like detective theme and tutorial section.

More recently, there have been increasingly game-like approaches taken (Vannella et al., 2014; Jurgens and Navigli, 2014). *Puzzle Racer* is a GWAP for image-sense annotation. Players tie images to word senses by racing through a series of gates, attempting to pass through gates that match a certain word sense (Jurgens and Navigli, 2014). Whilst a great example of a GWAP for NLP annotation, the game describes itself as "purely visual" and has a task itself that maps to images leaving the task not too far from being image labelling, rather than a typical NLP annotation task. *Puzzle Racer* recruited students incentivised by monetary prizes for top scoring players, and demonstrated a reduced cost over traditional crowdsourcing methods.

The *Wordrobe* suite of games (Bos et al., 2017) supports multiple games that perform similar annotations to that of *TileAttack* including tasks such as Named Entity Recognition and finding the referents of pronouns. However, unlike *TileAttack*, the *Wordrobe* games perform preprocessing to identify potential text segments, and then ask the player to identify which of those potential segments are correct. Whilst this fits nicely into a common game design that runs throughout the suite of games, it does constrain the players choices to potentially incorrect items. In comparison, *TileAttack* is only constrained to token boundaries.

3. TileAttack

TileAttack is a web-based two player blind game in which players are awarded points based on player agreement of the tokens they mark. The visual design of the game is inspired by *Scrabble*, with a tile like visualisation (shown in Figure 1).

In the game, players perform a text segmentation task which involves marking spans of tokens represented by tiles.

[1] `https://tileattack.com`

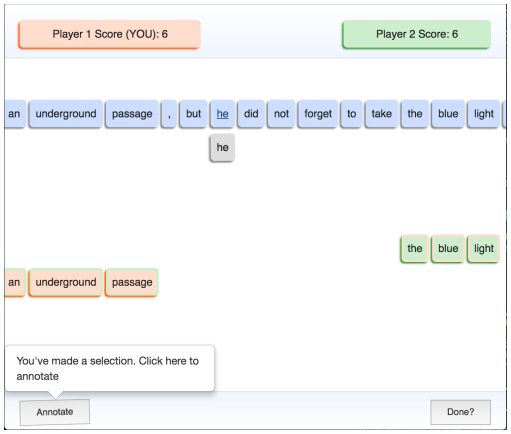


Figure 1: In game screenshot from *TileAttack*

Figure 2: Tutorial screenshot from *TileAttack*

Our approach was to start with a game design that begins from as close as possible to an existing working recipe. We chose a design that is in many respects analogous to *The ESP Game*, but for text annotation. This provides the opportunity to test what lessons learned from games similar to *The ESP Game* still apply with text annotation games, and how, in the domain of text annotation, these lessons can be expanded upon. Like *The ESP Game*, we use the "output-agreement" format for the game, in which two players or agents are anonymously paired, and must produce the same output, for a given input (Von Ahn and Dabbish, 2008).

3.1. Gameplay

Following the documentation, but before the game, players are shown a two round tutorial (shown in Figure 2). For crowdsourced players, completion of this tutorial is mandatory. In the tutorial the player marks two sentences. They are informed of what entities are present in the sentence and how many mentions there are. They can incorrectly mark multiple items, which will be highlighted with a flashing red border, but will only be allowed to proceed once they have discovered all the correct items (shown by the glinting effect). They receive immediate and direct feedback to inform them of their progress.

In each game round, the player is shown a single sentence to annotate. The players can choose to select a span from the sentence by simply selecting the start and end token of the item they wish to mark using the blue selection tokens. A preview of their selection is then shown immediately below. To confirm this annotation, they may either click the preview selection or click the *Annotate* button. The annotation is then shown in the player's colour. When the two players match on a selection, the tiles for the selection in agreement are shown with a glinting effect, in the colour of the player that first annotated the tiles and a border colour of the player that agreed. The players' scores are shown at the top of the screen.

Players receive a single point for marking any item. If a marked item is agreed between the two players, the second player to have marked the item receives the number of points that there are tokens in the selection, and the first player receives double that amount. The player with the greatest number of points at the end of the round wins.

When a player has finished, they click the *Done* button, upon which they will not be able to make any more moves, but will see their opponent's moves. Their opponent is also notified they have finished and invited to click *Done* once they have finished. Once both players have clicked *Done*, the round is finished and both players are shown a round summary screen. This screen shows the moves that both players agreed on, and whether they won or lost the round. Clicking *Continue* then takes the player to a leaderboard showing wins, losses and the current top fifteen players. From this page they may click the *Next Game* button, to start another round.

4. Experiment

4.1. Task

In this experiment we will test *TileAttack* with three separate audiences discussed below. The results of the experiment will be compared on both accuracy, and as an evaluation of player recruitment, using a set of metrics adapted from Free-to-Play games (Xicota, 2014).

In this game, players mark "mentions". These entities would normally be collected by a mention detection system and are typically used as part of larger NLP pipelines, such as relation extraction systems or co-reference resolution systems (Lee et al., 2011). To determine how successfully players are annotating the corpus, they are given sentences from the gold standard Phrase Detectives corpus (Chamberlain et al., 2016) to annotate.

4.2. Recruitment

To test *TileAttack's* ability to attract players in a gaming audience, it has been integrated with the *Kongregate* platform. [2] *Kongregate* is a popular indie game platform with an audience exposure of approximately 40,000 players.

To test *TileAttack* with a group interested in the field of linguistics, *TileAttack* has been added to a new NLP games portal. The Linguistic Data Consortium - University of

[2] https://www.kongregate.com/

Pennsylvania (LDC) project, *LingoBoingo* [3]. The LDC advertised their new portal during that month via social media channels and a newsletter. This audience is most comparable with the previous experiment, that also focused on online communities interested in linguistics (Madge et al., 2017).

To test *TileAttack's* ability to gather annotations and the benefit of the new tutorial irrespective of the game qualities, *TileAttack* has been integrated with Amazon Mechanical Turk, a crowdsourcing platform that remunerates workers on behalf of requesters to carry out small tasks. These tasks are known as *Human Intelligence Tasks* (HITs). A requester can choose from one of several Amazon Mechanical Turk templates to upload data into, or creating a custom integration. They may also specify the number of unique workers to carry out each HIT, and requirements for those workers that include qualifications. These qualifications can be awarded by the requester and serve as a flag to positively or negatively filter workers.

In our implementation, we make use of the *ExternalQuestion API*. This results in *TileAttack* being displayed in a HTML IFrame in the MTurk requester interface as a custom question. Having successfully taken part we award workers with a qualification. This satisfies the requirement of each worker participating only once, by serving as a flag on their account that is checked to prevent future tasks being displayed to them.

4.3. Experimental Design

For both *Kongregate* and LDC players, their experience is exactly as described in TileAttack's usual gameplay.

TileAttack is integrated into Amazon Mechanical Turk. Workers are shown the game documentation, with game references removed. They are then taken to the tutorial. They must complete the tutorial before they are allowed to perform the annotation task itself. Having completed the tutorial they are then asked to annotate six sentences. The core game mechanics, including scores or any evidence of a second player, are removed. The game like interface remains. Having completed the tutorial and five sentences, the participants are then remunerated for their participation (0.50 USD). Each participant is only allowed to take part once.

5. Results

Of the participants that attempted the crowdsourcing task, approximately 15% continued to completion. We take all completed games in these results, including contributions from crowdsourcing participants that did not fully complete the crowdsourcing task.

5.1. Annotation Quality

The player's annotations are compared with that from the expert annotated Phrase Detectives corpus (Chamberlain et al., 2016). This corpus provides expert annotated data as corrections to an automated pipeline. The game does not attempt to apply the corrections from the corpus. This analysis of annotation quality uses a subset of the sentences that were expert approved without requiring corrections.

[3] https://lingoboingo.org/

	LDC	Kongregate	MTurk
Precision	60.3	16.3	72.7
Recall	55.2	17.5	66.7
F-Measure	57.6	16.9	69.5

Table 1: User-based annotation accuracy from *TileAttack* used by 3 groups

	LDC	Kongregate	MTurk
Precision	60.1	29.6	38.9
Recall	61.7	60.7	89.3
F-Measure	60.9	39.8	54.2

Table 2: Item-based annotation accuracy from *TileAttack* used by 3 groups

	LDC	Kongregate	MTurk
Games	109	20	352
Items	56	5	9
Avg. Annotations	1.8	3.6	26.4
Participiants	19	7	73

Table 3: User play data from *TileAttack* used by 3 groups

As we are interested in both the design of the system and its ability to gather accurate annotations, we take two measurements of accuracy. Table 1 is the average accuracy for each user, in each game. We use this to judge how successful the system was in communicating the task to a specific audience and enabling contribution. This is comparable to the previous experiment, albeit without a tutorial, in which *TileAttack* players achieved 56.6% precision and 59.4% recall (Madge et al., 2017).

Table 2 is the average accuracy over all items (taking a union of all annotations provided by all users in that group, for that item). This allows us to judge on the whole, how successful the system is at gathering annotations. It is also important to measure both due to the way tasks are distributed to players.

Table 3 shows the number of participants for each group, the games they played, how many items were annotated and the average annotations per item. A higher number of annotations per item is very likely to raise recall. This occurs when there is a wide spread in the number of games played by the users. If a few users play many games, the system will present those users with games they have not seen before, so many individual annotations per item will be received for that group. This does impact the results shown in Table 2, but not those in Table 1. The average annotations per item are far higher for the MTurk players, as the system ensured everyone played six games, so items were more evenly annotated.

The crowdsourced players (MTurk), on average achieved a high average precision and recall. Their contribution over the items had a much higher recall, but also a much lower precision. These players were forced to take the tutorial and motivated financially. This demonstrates the system does appear to be effective in gathering annotations.

TileAttack did not appear to be successful in terms of accuracy on the *Kongregate* platform. Over a period of one

month on the Kongregate platform, only 7 players chose to play *TileAttack*. They rated the game at 1.3/5 stars.
LDC players achieved precision and recall comparable to that of online linguistic groups in the previous experiment (Madge et al., 2017).

5.2. Analysis using Free-To-Play Metrics

	LDC	Kongregate	MTurk
LTJ (mention)	8	2	40
LTJ (sentence)	1	2	8
AJpP (mention)	8	2.5	16
AJpP (sentence)	1	2	2
ALP (secs)	115	180	193
MAU	19	7	73
Retention (1 day)	0	0	0

Table 4: Free-to-Play metrics for *TileAttack* used by 3 groups

Table 4 shows adapted free to play metrics for TileAttack. These metrics are defined as follows: *Lifetime Judgements (LTJ)* is the average number of items annotated per player over their lifetime of play. *Average Judgements per Player (AJpP)* is the average number of items marked per player, per gaming session. *Average Lifetime Play* is the average session length in time. *Monthly Active Users (MAU)* is the number of users in a month, the active part refers specifically to those that finished a game. *Retention and churn* is the players that were kept and lost respectively, over some time period.

6. Conclusion

TileAttack presents a fast and usable interface for sequence labelling with embedding. The system, including the design of features such as the tutorial, appear to be effective in communicating the nature of the desired annotation to non-experts. When players are financially incentivised, *TileAttack* does now achieve a high level of recall. Obviously, the strengths of a crowdsourcing approach is based on robust aggregation methods that extract the wisdom of the crowd and filter out outliers. However, here we aim to obtain high-quality annotations in the first place independent of various aggregation methods that may be added later.
In our continued progress with the *TileAttack* game, we have demonstrated, with the recent addition of a tutorial, we can reach a fair level of accuracy using non-expert annotators. If the crowdsourced participants were permitted to continue contributing, we may reasonably expect that the accuracy of their contribution may increase further with their experience.
Whilst *TileAttack* did not perform very well on *Kongregate*, this was by far the most challenging setting so far. Set alongside indie games, *TileAttack* still fails to attract the volumes of players necessary to annotate a large corpora. Now the interface and instructions appear to be satisfactory, more work must be done for *TileAttack* to work in a game setting. This will involve further testing of game design concepts and mechanics to improve both *TileAttack's* ability to attract and retain players.

7. Bibliographical References

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). A platform for collaborative semantic annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 92–96. Association for Computational Linguistics.

Bos, J., Basile, V., Evang, K., Venhuizen, N., and Bjerva, J. (2017). The groningen meaning bank. In Nancy Ide et al., editors, *Handbook of Linguistic Annotation*, volume 2, pages 463–496. Springer.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2016). Phrase detectives corpus 1.0 crowdsourced anaphoric coreference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z., et al. (2010). Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760.

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pages 9–15. ACM.

Hacioglu, K., Douglas, B., and Chen, Y. (2005). Detection of entity mentions occurring in english and chinese text. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 379–386. Association for Computational Linguistics.

Haghighi, A. and Klein, D. (2010). Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393. Association for Computational Linguistics.

Jurgens, D. and Navigli, R. (2014). It's all fun and games until someone annotates: Video games with a purpose for linguistic annotation. *TACL*, 2:449–464.

Kim, Y., Riloff, E., and Gilbert, N. (2011). The taming of reconcile as a biomedical coreference resolver. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 89–93. Association for Computational Linguistics.

Kummerfeld, J. K., Bansal, M., Burkett, D., and Klein, D. (2011). Mention detection: heuristics for the ontonotes annotations. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 102–106. Association for Computational Linguistics.

Lafourcade, M., Joubert, A., and Le Brun, N. (2015). *Games with a Purpose (GWAPS)*. John Wiley & Sons.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning:*

Shared Task, pages 28–34. Association for Computational Linguistics.

Lee, T. Y., Dugan, C., Geyer, W., Ratchford, T., Rasmussen, J. C., Shami, N. S., and Lupushor, S. (2013). Experiments on motivational feedback for crowdsourced workers. In *ICWSM*.

Lee, H., Surdeanu, M., and Jurafsky, D. (2017). A scaffolding approach to coreference resolution integrating statistical and rule-based models. *Natural Language Engineering*, pages 1–30.

Madge, C., Chamberlain, J., Kruschwitz, U., and Poesio, M. (2017). Experiment-driven development of a gwap for marking segments in text. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '17 Extended Abstracts, pages 397–404, New York, NY, USA. ACM.

Mason, W. and Watts, D. J. (2010). Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with mmax2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.

Nguyen, T. H., Sil, A., Dinu, G., and Florian, R. (2016). Toward mention detection robustness with recurrent neural networks. *arXiv preprint arXiv:1602.07749*.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Peng, H., Chang, K.-W., and Roth, D. (2015). A joint framework for coreference resolution and mention head detection. In *CoNLL*, volume 51, page 12.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM TiiS*, 3(1):3.

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

Soraluze, A., Arregi, O., Arregi, X., Ceberio, K., and De Ilarraza, A. D. (2012). Mention detection: First steps in the development of a basque coreference resolution system. In *KONVENS*, pages 128–136.

Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. (2009). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 656–664. Association for Computational Linguistics.

Sweetser, P. and Wyeth, P. (2005). Gameflow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3):3–3.

Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., and Navigli, R. (2014). Validating and extending semantic knowledge bases using video games with a purpose. In *ACL (1)*, pages 1294–1304.

Von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *SIGCHI*, pages 319–326. ACM.

Von Ahn, L. and Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8):58–67.

Xicota, D. (2014). Free to play and its Key Performance Indicators.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *ACL (Conference System Demonstrations)*, pages 1–6.

Zhekova, D. and Kübler, S. (2010). Ubiu: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99. Association for Computational Linguistics.

LieCatcher: Game Framework for Collecting Human Judgments of Deceptive Speech

Sarah Ita Levitan, James Shin, Ivy Chen, Julia Hirschberg
Dept. of Computer Science, Columbia University
New York, NY USA
{sarahita@cs.columbia.edu, js4785@columbia.edu, ic2389@columbia.edu, julia@cs.columbia.edu}

Abstract

We introduce "LieCatcher", a single-player web-based Game With A Purpose (GWAP) that allows players to assess their lie detection skills, while simultaneously providing human judgments of deceptive speech. Players listen to audio recordings from the Columbia X-Cultural Deception (CXD) Corpus, a collection of deceptive and non-deceptive interview dialogues, and guess if the speaker is lying or telling the truth. They are awarded points for correct guesses, and lose lives for incorrect guesses, and at the end of the game, receive a score report summarizing their performance at lie detection. We present the game design and implementation, and discuss plans for using the human annotations for research into the acoustic-prosodic properties of believable, trustworthy speech. This game framework is flexible and can be applied to other useful speech annotation tasks, and we plan to make the game available to the public to extend for other tasks.

Keywords: games with a purpose, speech annotation, deception, trust

1. Introduction

In recent years, much progress has been made in developing and improving human language technologies. Some of these advances have been made using supervised learning methods, which rely on an abundance of annotated data. For example, a state-of-the-art commercial automatic speech recognition (ASR) system can rely on as much as 5000 hours of annotated speech (Hannun et al., 2014) Speech corpus annotation is a critical component of any speech related research. Traditionally, this annotation has been done by a small group of highly skilled annotators. This is a time consuming process, with extensive training required, and it is also expensive. In recent years, crowdsourcing has revolutionized the annotation process. Instead of relying on a few skilled annotators, crowdsourcing allows us to collect annotations from a large group of unskilled crowd workers, quickly and cheaply. Because this work is unskilled, it is important to take steps to control the quality of the annotations. An alternative approach to collecting annotations involves the use of Games With A Purpose (GWAP). The idea behind GWAP is to motivate people to solve computational problems by presenting the problem as a series of simple steps in an enjoyable game format.

In this work we have designed and implemented a GWAP with the goal of collecting human judgments for a corpus of deceptive speech. In our ongoing research, we are examining human ability at deception detection. The corpus contains dialogues between interviewer/interviewee pairs, where the interviewer asks 24 biographical questions, and the interviewee aims to deceive her partner for a random half of the questions. The interviewer records his judgment of each question, i.e. whether he thinks his partner is telling a lie or the truth. With this paradigm, we have record of a single human judge for every interviewee response. However, we are interested in exploring human perception of deception at a larger scale, exploring individual differences in how people perceive deception, as well as exploring trust. To do this, we need many instances of human judgments for each utterance. A previous perception study of human performance at deception detection recruited 32 participants to listen to audio recordings ranging from 25-50 minutes long, and annotate them with their judgments of deception (Enos et al., 2006). This process typically requires an experimenter to schedule, train, and supervise the participants, and it can be a time consuming and expensive ordeal. In addition, although the human judges are paid for their time, there is no explicit motivation for the judges to perform well at the specific task that they are working on, and it is conceivable that they will become disinterested in the task and even answer randomly.

Here we introduce a GWAP to collect large-scale human annotations of deception. This framework has several advantages. It enables the rapid, large-scale collection of human annotations - multiple users can play in parallel, and they can play the game from any location, at any time. It is inexpensive - players are unpaid, motivated by the enjoyment of the game, and there is no need for a human to train the players. There is explicit incentive for players to perform well at the task, in the form of points and loss of game lives. In addition, the game implementation is flexible and makes it easy to manipulate conditions, so that we can design experiments to test theories of human perception of deception.

The rest of this paper is organized as follows. Section 2. reviews related work, and Section 3. details the speech corpus that we use for the game. In Section 4., we describe the design and implementation of LieCatcher. Section 5. describes an initial pilot study that we conducted to get early feedback about the game design. We conclude in Section 6. with a discussion of ongoing and future work.

2. Related Work

Games with a purpose (GWAP) have been previously used for annotation of language corpora, including text and speech modalities. "tashkeelWAP" (Kassem et al., 2016) is a web application with a single-player and a two-player game where Arabic speaking players digitize Arabic words with their diacritics that were not correctly recognized by OCR systems. "Phrase Detectives" [1] is another annotation game, where players label relationships between words and phrases, to create a rich language resource of anaphoric co-references (Chamberlain et al., 2008).
"Voice Race" (McGraw et al., 2009) and "Voice Scatter" (Gruenstein et al., 2009) are GWAP that are educational for their players, and also useful for obtaining speech annotations. In "Voice Race", A player is presented with a set of word definitions on flashcards, and they must quickly say the corresponding words. In "Voice Scatter", the player chooses flashcards to study, and when presented with a term, speaks the definition into a microphone, earning points for correct responses. This game elicits spontaneous speech in longer sentences. By using speech recognition as well as contextual information from the games, the spoken utterances can be labeled orthographically with near perfect accuracy. These games are enjoyable as well as educational, and provide labeled speech data as a by-product of the games.

3. Corpus

For this work, we examined the Columbia X-Cultural Deception (CXD) Corpus (Levitan et al., 2015) a collection of within-subject deceptive and non-deceptive speech from native speakers of Standard American English (SAE) and Mandarin Chinese (MC), all speaking in English. The corpus contains dialogues between 340 subjects. A variation of a fake resume paradigm was used to collect the data. Previously unacquainted pairs of subjects played a "lying game" with each other. Each subject filled out a 24-item biographical questionnaire and were instructed to create false answers for a random half of the questions. They also reported demographic information including gender and native language, and completed the NEO-FFI personality inventory (Costa and McCrae, 1989).

The lying game was recorded in a sound booth. For the first half of the game, one subject assumed the role of the interviewer, while the other answered the biographical questions, lying for half and telling the truth for the other; questions chosen in each category were balanced across the corpus. For the second half of the game, the subjects' roles were reversed, and the interviewer became the interviewee. During the game, the interviewer was allowed to ask the 24 questions in any order s/he chose; the interviewer was also encouraged to ask follow-up questions to aid them in determining the truth of the interviewee's answers. Interviewers recorded their judgments for each of the 24 questions, providing information about human perception of deception. The entire corpus was orthographically transcribed using the Amazon Mechanical Turk (AMT)[2] crowd-sourcing platform, and the speech was segmented into *inter-pausal units* (IPUs), defined as pause-free segments of speech separated by a minimum pause length of 50 ms. The speech was also segmented into turn units, where a turn is defined as a maximal sequence of IPUs from a single speaker without any interlocutor speech that is not a *backchannel*. Finally, the speech was segmented into question/answer pairs, using a question detection and identification system (Maredia et al., 2017) that employs word embeddings to match semantically similar variations of questions to a target question list. This was necessary because interviewers asked the 24 questions using different wording from the original list of questions.
In total, there are 7,141 question/answer pairs, each associated with a question number (1-24), start and end times in the full session recording, transcribed text, and truth value (T or F).

[1] http://anawiki.essex.ac.uk/phrasedetectives/

4. Game Design and Implementation

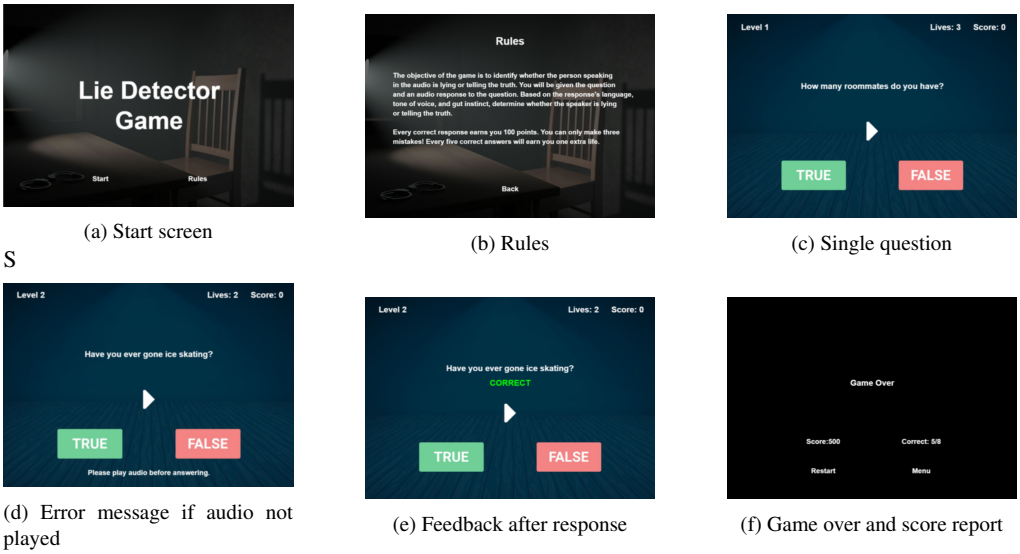
4.1. Game Design

The game design is simple and flexible. The player is presented with a series of audio recordings from the CXD corpus, each one paired with the text of the interviewer question that prompted the interviewee's response. The player listens to each interviewee audio clip, and selects whether they think the speaker is lying or telling the truth. The player can listen to the audio an unlimited number of times, but is required to listen to the full audio before selecting a "True" or "False" button. Each player is given 3 lives; a correct guess earns the player 100 points, while an incorrect judgment causes the player to lose one life. The game ends when the player has lost 3 lives, and the final screen of the game is a display summarizing the player's performance. The points and lives, as well as the final score summary, serve to motivate the player to try their best to succeed at the game.

Figure 1 displays screenshots from the main 6 stages of the game: (a) Start screen, where users select "play" or "rules", (b) Rules, which lists the rules of the game, (c) Single question, which shows the text of a question along with a play button to listen to the audio, along with "True" and "False" buttons to select the deception judgment, (d) Error message displayed if the audio was not played before selecting a button, (e) Feedback after the player selects a button, showing the correct answer, and (f) Game over and score report displaying information about player performance when the player loses all his lives.

There are many decisions to make in creating this framework. How many lives should players start with? How many times can the player listen to the audio? Should the players receive instant feedback about their judgments, or only at the end of the game? Should the audio clips be randomly chose, or perhaps ranked in some manner (e.g. by difficulty)? Some of these decisions may significantly impact player performance. For example, it is possible that

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



(a) Start screen

(b) Rules

(c) Single question

S

(d) Error message if audio not played

(e) Feedback after response

(f) Game over and score report

Figure 1: Main states of the game.

players would benefit from receiving instant feedback about their judgments as they play the game. We are interested in exploring the effect of these parameters, and therefore have implemented these options in a flexible manner, so that we can experiment with different settings and observe their effects. In addition, we plan to extend the game to accommodate for different levels and designs (where certain levels would have differing conditions – limited time, or limited number of listens of the audio, or no feedback, or more difficult audio recordings), which should be interesting to study as the data grows and the game is played more frequently over time.

4.2. Game Implementation

We used PhaserJS [3] for LieCatcher's framework. PhaserJS is a 2D game framework for creating HTML5 games for web browsers. We chose this framework because of its lightweight features and intuitive javascript syntax. PhaserJS is a state-based game framework, meant to support small games. In a state-based game framework, every scene in the game is its own state that the user is in (i.e., "Menu", "Rules", "Stage1", "Stage2", etc.). Because of this, assets must be loaded quickly, so as not to slow the gameplay. Other larger game development engines as Unity allow support for multithreaded applications, but this comes with additional overhead, and is not necessary for our lightweight game.

For the backend of the game, we stored the audio files in a MongoDB database [4] hosted on our own server. Since PhaserJS does not natively support database queries, we set up endpoints on our site server using ExpressJS URLs that returned queries from our database. When loading assets, phaserJS queries the appropriate site server endpoints and receives request responses corresponding to the data of interest. Specifically, in each state, loading assets is typically done synchronously in a "pre-load" method before the assets are placed into the scene. Because it takes a significant amount of time to load over 7,000 audio files, we instead loaded the audio files asynchronously in a queue in the background during gameplay, as to not interfere with the user experience. One audio file is loaded in the background while the player plays each stage (i.e. each audio clip). The weakness to this approach is that a player may spend less time playing a certain stage than the time it takes to load one audio file. However, the longest audio files load in under 5 seconds, so loading times are not a major issue to the user experience.

When a player loses all three lives, they are sent to the game over screen, and during this state, we send user session data to the user database. The data include the IDs of questions that were correct and incorrect, the time it took to answer each question, the player score, the date, and the number correct and total answered. We used the JS fetch API as a request handler to pass JSON data into the request bodies. This was done to collect data from the user session and store it into a separate user database.

5. Pilot Study

In order to get early feedback about the game, we conducted a pilot study where 40 students played the game and answered a pre-game and post-game survey. For the purpose of the study, we structured the game as 2 levels, with 10 audio samples in each level. In level 1, players were not provided with any feedback about the correctness of deception judgments. That is, they received no points for correct judgments, did not lose lives for incorrect judgments, and there was no message on the screen to indicate whether their judgment was correct or incorrect. In level 2, players received immediate feedback about their judgments with a displayed "correct" or "incorrect" message, as well as earning 100 points for each correct judgment. At the end of the 10 audio clips in level 2, players were given a score

[3] https://phaser.io
[4] https://www.mongodb.com

report for the level 2 questions.

Before playing the game, players filled out a pre-game survey. They were asked to report their gender and first language spoken, and answered three questions: (1) How often can you spot a lie in daily life? (on a scale of 1 to 5, with 1 being almost never and 5 being almost always) (2) How often do you think people lie in daily life in order to achieve some gain, either material or social? (also on a scale of 1 to 5) (3) Do you have experience in law enforcement or in another job where spotting lies is important? If yes, please describe.

After answering the pre-game survey, participants played the pilot game. We introduced two quality control questions to ensure that players were paying attention and listening to the audio, and not selecting buttons randomly (e.g. with their audio turned off). In each level, one of the audio segments was a recording that said "Please wait 5 seconds and select TRUE" or "select FALSE".

After playing the game, participants answered a post-game survey and provided feedback about their experience. Questions included: Did you find the game to be easy to use? Which level did you prefer (level 1 or level 2)? How would you rate your ability to detect deception after playing this game? How well do you think your score on the game reflects your ability to detect lies in the real world? Did you like the premise of the game? Would you recommend the game to a friend? Did you like the game graphics? Players also provided feedback about the quality control questions, and general ideas about the game. They also reported strategies that they used in making their judgments. Some of the survey questions were adapted from a study of human judgments of deception by Enos et al (Enos et al., 2006) and others game evaluation questions were adapted from (Sturm et al., 2017).

5.1. Pilot Study Survey Responses

40 students participated in the pilot study, 26 female and 14 male. 77% of the participants were native speakers of English, and the rest were native speakers of other languages (e.g. Chinese), but were proficient in English. Only one player reported job experience with lie detection.

35 of the players reported using a laptop or desktop computer to play, while 5 players used a mobile device. They played using various browsers, including Chrome, Firefox and Safari, without compatibility issues. Overall, the feedback about the game was positive. 85% found the game easy to use, and 75% reported that they would or might recommend the game to a friend.

Player responses were mixed about whether they thought the game is a good way to assess ability to detect lies. 57% responded yes or maybe, while 43% responded no. 73% of players preferred level 2, where feedback was given, to level 1. This information is useful for future game design choices. The feedback about the quality control questions was informative - some players thought it was a great idea to check attention, while others found it slightly confusing. In the future, we might inform players to expect such questions distributed throughout the game, to avoid confusion. 70% of respondents liked the premise of the game, 18% were neutral, and 12% did not like the premise. 50% liked the game graphics, while 35% were neutral and 15% did not like them. Going forward, we plan to incorporate ideas from this initial player feedback in order to improve the player experience.

5.2. Pilot Study Player Behavior

Players were overall 49.86% accurate in their predictions, not including check questions. The minimum correct number of questions by a player was 5 correct, while the maximum was 13. The median and mean was 9 correct, with a standard deviation of 1.94. 100% of players answered the check questions correctly and made sure to listen to directions and wait five seconds, indicating that players were attentive in making their decisions. Overall, however, players were still on average approximately as accurate as random guessing.

There was a noticeable difference in player performance in between the levels. For level 1, the average number of correct questions was 4.1 out of 9, with a median of 4 and standard deviation of 1.18. The overall accuracy of all players was 45%. In contrast, level 2 players averaged 4.9 correct answers of 9, with a median of 5 and standard deviation of 1.27. The overall accuracy for level 2 was 55%.

Some questions had collective responses strongly in favor of an answer choice. In particular, question 5 had 33 responses to "T" and 7 to "F" with an accuracy of 17.5%, question 9 had 34 responses to "T" and 6 to "F" with an accuracy of 85%, and question 14 had 33 responses to "T" and 7 to "F" with an accuracy of 82.5%. There were no questions with responses strongly in favor of "F", indicating that for the given audio sample pool, players were more inclined to trust confidently than to accuse.

There was a negligible difference in performance between female and male players. Female players were 50% accurate with a trust rate of 60%, while male players were 49% accurate with a trust rate of 62%.

6. Conclusions and Future Work

We presented LieCatcher, a GWAP where players can learn how well they perform at deception detection, while providing human annotations of deception. This game framework allows for the rapid and large-scale collection of human annotations of deceptive speech, and can easily be extended to other speech annotation tasks. We plan to make the game implementation publicly available for further development. We conducted a pilot study to get early player feedback about the game. The initial feedback is promising, and we plan to incorporate some of the feedback to further improve the game.

We are now in the process of testing the game on student volunteers. So far we have received feedback that the game is entertaining; people enjoy assessing their abilities at lie detection. Once this is completed and preliminary feedback is addressed, we plan to distribute the game on crowdsourcing platforms such as Amazon Mechanical Turk to collect large-scale annotations. After this data collection phase, we will conduct an analysis of acoustic-prosodic properties of trustworthy speech. We also plan to explore the role of gender and culture (of the speaker as well as the listener) on trust.

7. Acknowledgements

This work was partially funded by AFOSR FA9550-11-1-0120 and by NSF DGE-11-44155.

8. Bibliographical References

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

Costa, P. and McCrae, R. (1989). Neo five-factor inventory (neo-ffi). *Odessa, FL: Psychological Assessment Resources*.

Enos, F., Benus, S., Cautin, R. L., Graciarena, M., Hirschberg, J., and Shriberg, E. (2006). Personality factors in human deception detection: comparing human to machine performance. In *INTERSPEECH*.

Gruenstein, A., McGraw, I., and Sutherland, A. (2009). A self-transcribing speech corpus: collecting continuous speech with an online educational game. In *International Workshop on Speech and Language Technology in Education*.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Kassem, L., Sabty, C., Sharaf, N., Bakry, M., and Abdennadher, S. (2016). tashkeelwap: A game with a purpose for digitizing arabic diacritics.

Levitan, S. I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M., and Rosenberg, A. (2015). Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 1–8. ACM.

Maredia, A. S., Schechtman, K., Levitan, S. I., and Hirschberg, J. (2017). Comparing approaches for automatic question identification. SEM.

McGraw, I., Gruenstein, A., and Sutherland, A. (2009). A self-labeling speech corpus: Collecting spoken words with an online educational game. In *Tenth Annual Conference of the International Speech Communication Association*.

Sturm, D., Zomick, J., Loch, I., and McCloskey, D. (2017). "free will": A serious game to study the organization of the human brain. In *International Conference on Human-Computer Interaction*, pages 178–183. Springer.

Cheap, Fast and Good!
Voting Games with a Purpose

Karën Fort, Mathieu Lafourcade, Nathalie Le Brun
STIH/Sorbonne Université, Paris, France - LIRMM, Montpellier, France - Imaginat, Lunel, France
karen.fort@sorbonne-universite.fr, mathieu.lafourcade@lirmm.fr, imaginat@imaginat.name

Abstract

We present in this paper the voting games with a purpose that were developed around JeuxDeMots, a central game aiming at creating a lexical network for French. We show that such lightweight applications can help collect quality language resources very efficiently and we advocate for a common platform for such voting games for language resources.

Keywords: crowdsourcing, GWAP, voting games

1. Introduction

JeuxDeMots[1] is a game with a purpose (GWAP) aiming at creating a lexical semantic network for French (Lafourcade, 2007). The game was created in 2007, in the wake of the ESP Game (von Ahn and Dabbish, 2004), and is therefore one of the first GWAPs for natural language processing with Phrase Detectives (Chamberlain et al., 2008), long before wordrobe (Bos and Nissim, 2015) and ZombiLingo (Guillaume et al., 2016).

Since September 2007, more than 4,000 players have registered on JeuxDeMots and 1,523,321 games have been played. As of today (January 2018), the network contains more than 2 million terms linked by more than 180 million relations.

The idea to develop complementary games came naturally, as the main game interface and features did not seem adequate to gather some specific information. More specifically, very simple click-only games, which can be played casually without registering and on smartphones, looked promising. Also, multiplying game designs would compensate for the multiple biases of the JeuxDeMots original design, hence producing wider coverage and more accurate lexical and semantic data.

The first voting game which was added aside of JeuxDeMots is PtiClic (Lafourcade and Zampa, 2009). Its aim is to distribute terms according to their relation to a given target term. The player has to click and drag terms toward the appropriate box associated to a specific semantic relation (see Figure 1). Once finished, the proposals are compared to those of others players.

Now 12 complementary games are available through a portal[2] (see Figure 2), 10 of which are simple voting games.

To our knowledge, very few voting GWAPs exist for language resources creation. Apart from the ones we just mentioned, the only GWAP that relates to this type of game is wordrobe (Bos and Nissim, 2015)[3]. On this platform, players are invited to participate to a variety of tasks, related to semantic disambiguation (noun *vs* verb, co-reference identification, named entity annotation, etc) and to choose an answer from a limited list of solutions. Although the concept resembles that of the JeuxDeMots portal, wordrobe tasks are much more complex and require more concentration and some more advanced (at least school-level) knowledge.

Figure 1: PtiClic: the term *repos* (*rest*) is the target term. Each term of the cloud should be dragged and dropped in one of the three boxes on the right-hand side.

Figure 2: The JeuxDeMots portal. Note that Totaki and top10 are not voting games.

[1] See: http://www.jeuxdemots.org/.

[2] See: http://imaginat.name/JDM/Page_Liens_JDMv2.html.

[3] See: http://wordrobe.housing.rug.nl/Wordrobe/public/HomePage.aspx.

2. A Galaxy of Voting Games

2.1. Common Features

We define voting games as very simple games in which the players have to choose between a predefined, limited number of answers, without any training. The selection (or vote) of the player is compared to those of the other players, and more specifically to the state of the resource, in order to perform two tasks: a) including the answer in the resource and b) computing some reward points, which are part of the game functionalities.

Contrary to a quiz game, in which the correct answers are known, we obviously cannot compare the votes which are cast to a reference. Therefore, the majority of answers is used to generate rewards corresponding to what is considered as the right one. It has to be noted that two games (distant in time) with the same instructions might not yield the same results, as the underlying resource might have been modified in the meantime. The created resource (in our case, the `RezoJDM` lexical-semantic network) is dynamic and evolves over time.

We decided to exclude from the definition more complex games, like the ones allowing for free-text answers, like `Totaki` (a guessing game where clues are given to the system which tries to infer the target word) or `top10` (another guessing game, where the players can identify words selected by the system from a simple definition). We also exclude games requiring training, like `Argotario` (Habernal et al., 2017).

The interface of the simplest voting games is quite easy to develop, as it generally consists of a question, a term, and a couple of buttons to choose from. Beside being simple to master, such games are also well-adapted to mobile devices (smartphones and tablets) and they can be played quickly, anytime, anywhere.

An example of such an interface is presented in Figure 3, for `LikeIt` (Lafourcade et al., 2015): the balloons represent the possible answers (in this case, "Yes, I like the idea" / "I don't mind" / "No, I don't like the idea"), the term to decide on is centered and highlighted (here, *obscurcir*, i.e. *to darken*) and the votes on the previous term (*pendre*, i.e. *to hang*) are shown in a horizontal colored bar at the top of the page.

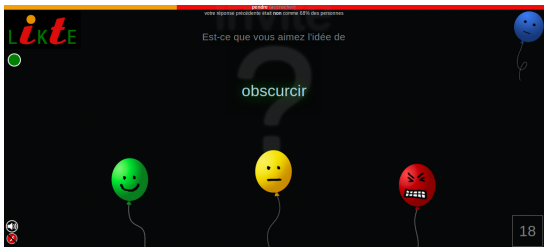


Figure 3: Interface of `LikeIt`, a polarity game with the term *obscurcir* (*to darken*).

Obviously, given the simplicity of the games, a nice design helps attracting players, so funny images are used (balloons in `LikeIt`) instead of simple buttons.

The main challenge is for the system to select adequate terms to be proposed to the player. The approach, based on the idea of potential information propagation consists of the following steps:

- identify a set of symbols/values that we want to tag the terms with. Adding a neutral value if needed. For example, in the case of `LikeIt` the values are: $\{pos\_positive, pos\_negative, pos\_neutral\}$;
- select of a term to tag. Randomly choose a target T (in a set of terms), which is already tagged (not with the neutral value). Then, there is p chance that you propose this term and $p-1$ that you propose one of its neighbors in the network. We set p to 0.5 in our experiments.
- bootstrap by tagging manually with a non neutral value at least one word. In the case of `LikeIt`, we tagged *bon* (*good*) with one positive vote and *mal* (*evil*) with one negative vote.

This simple selection algorithm allows to crawl the network, tagging terms through a propagation approach by maximizing the chance of proposing a target term that is relevant to the task. Increasing the value of p tends to slow down the propagation but increases the number of votes for each term.

2.2. Obtained Results

As reported in Lafourcade et al. (2015), during the first 3 months of `LikeIt` more than $25,000$ terms have been polarized (i.e. tagged with a combination of positive, negative and neutral votes), with a total of over $150,000$ votes. After 7 years, more than $360,000$ terms have been polarized for 75 million votes representing 70% of the terms contained in the network at that time. The Polarimot project (Gala and Brun, 2012) aimed at building a similar resource of polarized terms, but with classical means, i.e. manually. For this project, in the course of 3 months, 3 experts tagged (with 3 votes) a set of $2,400$ terms. The comparison between the resources (Polarimot and `LikeIt`) showed that for the common terms (corresponding to the $2,400$ terms of Polarimot) the obtained polarities are almost identical. The only difference (for less than 20 terms) is a more subtle polarization for terms that are polysemous with some contrastive polarity (like *affection* that refers both to love or to disease).

`SexIt` (Lafourcade and Fort, 2014) is based on the same principle as `LikeIt` (and the same internal engine). The purpose of the game is to assess if a given term is related to sex (in its broadest meaning). As reported in Lafourcade and Fort (2014), the propagation algorithm is especially efficient in crawling the underlying network to propose relevant terms.

`Selemo` (see Figure 4) is a voting game in which the number of choices depends on the target term and relation. The point of the game is to select the most (or least) relevant associations amongst those displayed. For example, what is the most relevant: "a bird can fly" or "a bird can sing"? In 4 months, more than $300,000$ relations have been tagged

Figure 4: Interface of `Selemo`. In this example, are the listed characteristics (*edible, hot, delicious, ...*) relevant or not to *casado* (a Costa Rican dish)?

(as relevant or not relevant) with this game. The accuracy of the results when 3 votes or more were cast is 100%, for 2 votes it is of 95% and 70% for just 1 vote.

Figure 5: `Askit` aims at assessing uncertain semantic relations, especially concerning polysemous words. In this example, can *archives* have the characteristic *pleine* (*full*)?

`AskIt` (see Figure 5) allows to validate/invalidate proposed relations inferred automatically from the `JeuxDeMots` lexical network. The `AskIt` engine selects a relation concerning a word meaning and ask if it holds for another meaning. For example: Does a *bank (river)* contain money? This strategy allows to build contrastive knowledge, which is instrumental in word sense disambiguation, especially when taking advantage of negative (i.e. inhibitory) relations. Since its launch in 2010, this game has allowed to validate/invalidate 1.5 million relations (corresponding to around 23 million votes) with an accuracy of 99.83%.
Similarly, `Emot` (see Figure 5) proposes a target term and a set of emotion/sentiment from which the player has to select the most appropriate (Lafourcade et al., 2016). Since its launch, more than 660,000 emotion/sentiment relations

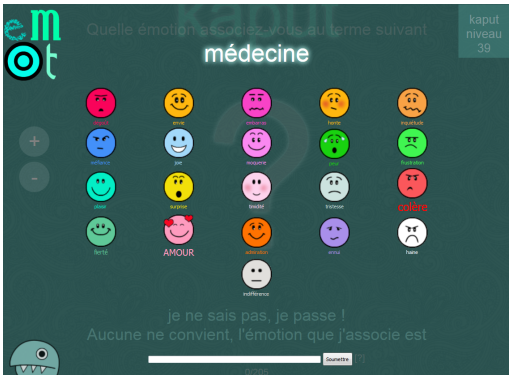


Figure 6: `Emot` aims at collecting sentiment associations with words. In this example, what are the sentiments that best correspond to *médecine* (*medicine*)?

have been created for 120,000 terms by 24 million votes.
`ColorIt` (Lafourcade et al., 2014) is based on the principle of `Emot` but adapted to color/appearance information. Since its start, more than 20,000 terms have been colorized with more than 3.7 million votes.
`PolitIt` (Tisserant and Lafourcade, 2015) is based on the same principle, but is adapted to political associations (for example, *market economy* with *liberalism*). Since its start, more than 8,000 terms have been *politized* with more than 500.000 million votes.
`Yakadirou` (see Figure 7) allows to associated a place preposition to a place relation. For example, in the relation *cat r_place sofa* what is the most relevant preposition: on, over, under? More than 380,000 bets have been placed in 2 years.

Figure 7: `Yakadirou` aims at associating prepositions to relations of place. In this example, what is the preposition of place to associate to *marchandise* and *postal parcel*?

`Tierxical` (see Figure 8) is a bit different from the previous games as it allows to bet on the first mostly associated terms for a given target term. The choice of the player slightly impacts the distribution of the relation weights. More than 750,000 bets have been placed in 5 years.

3. Limitations of the Approach

3.1. The Perils of Majority Voting

Although the influence of the other players' vote is limited, as the previous answers are only shown **after** the vote is

Figure 8: `Tierxical` aims at reordering word associations from the strongest to the weakest. In this example, what are the 3 best synonyms for *débiteur* (*debtor*)?

cast, majority voting still presents some important drawbacks.
First, the players are all considered equally, so a person who just plays around clicking randomly is considered the same as a highly skilled player.
Second, players can easily cheat if they agree on casting the same vote ("always click on Yes", for example).
These two limitations should be compensated by the number of players, provided enough of them play honestly.
Therefore, in such games, attracting a lot of players is especially important.

3.2. The Perils of Simplification

Another danger of voting GWAPs is that they can lead to over-simplification. One example of such a drift in a (microworking) crowdsourcing task is presented in (Bowman et al., 2015), in which in order to identify entailment relations, workers were asked if most people would say that if the first sentence is true, then the second must be too.
In our case, the voting tasks are complementary to a central game, `JeuxDeMots`, which allows to compensate, at least partly, for this effect.

4. Conclusion

Voting games are easy to develop and they provide a very efficient way of collecting large amounts of speakers' decisions in a very limited time. A common platform for such games would allow to easily gather language data, with very little development work.
In our case, the created resources are copyleft and can be downloaded directly from the games' Web sites, with a click on the upper left hand-side image.[4]

[4]For example, for `LikeIt`: `http://www.jeuxdemots.org/JDM-POLA-FR/?C=M;O=A`.

References

Bos, J. and Nissim, M. (2015). Uncovering noun-noun compound relations by gamification. In *Proc. of the Nordic Conference of Computational Linguistics (NODALIDA)*, pages 251–255, Vilnius, Lithuania, May.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase Detectives: a web-based collaborative annotation game. In *Proc. of the International Conference on Semantic Systems (I-Semantics)*, Graz, Austria.

Gala, N. and Brun, C. (2012). Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction dun lexique pour lanalyse dopinions. In *Proc. of Traitement Automatique des Langues Naturelles*, Grenoble, France, June.

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

Habernal, I., Hannemann, R., Pollak, C., Klamm, C., Pauli, P., and Gurevych, I. (2017). Argotario: Computational argumentation meets serious games. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Lafourcade, M. and Fort, K. (2014). Propa-l: a semantic filtering service from a lexical network created using games with a purpose. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, May.

Lafourcade, M. and Zampa, V. (2009). Pticlic : a game for vocabulary assessment combining jeuxdemots and lsa. In *Proc. of Conference on Intelligent text processing and Comptational Linguistics (CICLing)*, Mexico City, Mexico, March.

Lafourcade, M., Le Brun, N., and Zampa, V. (2014). Crowdsourcing word-color associations. In *Proc. of the International Conference on Application of Natural Language to Information Systems (NLDB)*, Montpellier, France, June.

Lafourcade, M., Le Brun, N., and Joubert, A. (2015). Collecting and evaluating lexical polarity with a game with a purpose. In *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Hissar, Bulgaria, September.

Lafourcade, M., Le Brun, N., and Joubert, A. (2016). Construire un lexique de sentiments par crowdsourcing et propagation. In *Proc. of Traitement Automatique des Langues Naturelles*, Paris, France, July.

Lafourcade, M. (2007). Making people play for lexical acquisition. In *Proc. of the 7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thailand, December.

Tisserant, G. and Lafourcade, M. (2015). Politit, du crowd-sourcing pour politiser le lexique. In *Proc. of Etudier le Web politique : Regards croiss, Institut des Sciences de l'Homme*, Lyon, France, June.

Name of the game	Created complementary resource	Information
`LikeIt`	polarized lexicon	150,000,000 votes - 740,000 terms - 1,700,000 polarities
`AskIt`	negative relations	25,000,000 votes - 860,000 negative relations
`SexIt`	sex/no sex relations	410,000 votes - 19,000 terms
`PolitIt`	political relations	540,000 votes - 8,900 politically tagged terms
`ColorIt`	color relations	3,700,000 votes - 20,000 colorized terms - 37,000 color relations
`Selemot`	annotations	23,000,000 votes - 1,500,000 annotations
`Yakadirou`	prepositions of place	380,000 votes - 27,000 place preposition annotated relations

Figure 9: Obtained results for the voting games.

von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proc. of SIGCHI conference on Human factors in computing systems*, CHI '04, pages 319–326, New York, NY, USA. ACM.

The JeuxDeMots Project is 10 Years Old: What We have Learned

Alain Joubert, Mathieu Lafourcade, Nathalie Le Brun
LIRMM, Montpellier, France - Imaginat, Lunel, France
alain.joubert@lirmm.fr, mathieu.lafourcade@lirmm.fr, imaginat@imaginat.name

Abstract

Here we present the assessment of 10 years of experience concerning the JDM project, a set of GWAPs for NLP, among which a main game combined with many satellite games aims to build a large lexical-semantic network for the French language. We highlight the lessons learned from this experience for creating lexical resources through a never ending process. We emphasize that combining automatic inference processes with player activity is particularly relevant to build such a resource. **Keywords:** crowdsourcing, game with a purpose, inferences, lexical semantic network

Introduction

The JeuxDeMots (JDM) project, whose very first GWAP was launched 10 years ago, in July 2007 (Lafourcade, 2007), aims to build a very large lexical-semantic network for the French language. Such a resource is usable in any application needing some semantic analysis of textual information and some reasoning capabilities about world fact and common sense. As a graph, the lexical network contains terms (words, groups of words, expressions, inflected forms, and symbolic informations) connected by typed semantic relations. It was an ambitious project, in the same spirit as Wordnet (Miller, 1995) for the aimed goal, and experience showed us it was feasible: the resource is freely available (C0 license) with a monthly updated export. Building such a resource may be made through different ways:

manual acquisition is a costly, long and fastidious work, where information would not likely be updated without further funding (the typical example being Wordnet or Framenet (Baker et al., 1998));

automatic construction from corpora: result can be biased by the corpus itself or the extraction method. Moreover, to correctly extract semantic relations, it is necessary to carry out a semantic analysis, which is precisely the object of the resource that one wishes to build;

myriadization of paid parcel work, with the risk that the data obtained are not of the expected quality (Fort et al., 2011); this type of method is based on the fact that many Internet contributors, often referred to as *turkers*, are willing to collaborate and are generally (lowly) remunerated for this collaboration.

Hence, we developed a collaborative game on the web in a crowd-sourcing way, where players would not-knowingly construct the resource by playing. As far as we know, prior 2007, such a method had never been used in the NLP domain.

In this paper, we first recall on which principles the main game relies, before addressing the adjustments we have had to perform. Then we present the generated resource, and the automatic methods that densify the network by consolidating (correction and completion) the data obtained from the games. We also point out several useful aspects of such a network in the field of NLP. Then, we discuss the lessons learned after ten years of using the JDM model. Finally, we emphasize that combining automatic inference processes with player activity is particularly relevant to build and densify such a resource.

1. JeuxDeMots

JDM is a GWAP (Game With A Purpose, see Von Ahn (VonAhn, 2004) and (Lafourcade et al., 2015)), that is to say a collaborative game which has a definite purpose beside entertaining (for example, collecting data or solving problems).

In a JDM match, two players collaborate anonymously and in an asynchronous way. A match of JDM is to propose a term and an instruction, asynchronously, to two players who do not know each other, and then to confront their answers. For example : *Give generics of goldfish*, or *which are the parts of motor-vehicle ?* Each player has a limited time to provide the answers he / she deems relevant, and when both sets of answers are confronted, the system only retains the common answers, to limit the risk of error: it is believed that answer is likely to be relevant when given by two players who have had no opportunity to consult each other. When both players give the same answer, but it does not exist in our database, the term is added. At the end of a game both players are rewarded with points and virtual gifts.

The number of terms and relations increases through player activity: we started with a $150,000$ terms data base and no relation; ten years later (2017), the network has more than 2.6 million terms and 180 million relations.

As we said, JDM is a game, but it is a useful game. The play aspect is essential to attract and retain players, and make them going on participating. But it's also a useful game, and as designers of JDM, we must never lose sight that the goal of the game is to build a lexical network. We will return in detail on this dual aspect in section 3, when we will develop what 10 years of JDM experience has taught us.

1.1. Evolutions of the Game

For a game to be attractive and attractive, to avoid monotony is essential. That's why we have tried to develop different game modes, to stimulate the emulation between the players by all sorts of rankings; we created the possibility for the players to give themselves gift-parties, to challenge themselves to duels, to choose from about 30 "skills" (ie the type of relations on which to answer, as synonym,

cause, consequence, family, agent, patient, instrument, location, feature, part, etc.) and to test many other parameters of play. The idea was to offer the possibility to play the main game in all sorts of ways, with all kinds of configuration. Moreover, we have gradually created, in addition to the main game, 12 "satellite" games, so that a player can temporarily abend the main game to try another game, and thus participate in the consolidation and verification of the data obtained through the main game. Indeed the analysis of the first data made some adjustments of the main game necessary, but also gave us the idea to create new games to verify, reinforce, or correct some data.
For the main game of JDM, the turn-over is relatively high: most players are active for about 3 weeks, sometimes even for several months, even years... Some have been playing JDM for 10 years! The initial game has therefore benefited from many improvements and additions over time, as it is detailed in (Lafourcade et al., 2015). We will highlight in particular:

The opportunity to **retry your chance** after a disappointing game, and even to sue the other player. The trials are held in public, the other players play the role of jurors, which is yet another way to create animation and conviviality.

The ability to **play on the theme of his choice**: a player will give more relevant answers in a field in which he is expert or passionate.

The ability to **choose the level of difficulty** of the proposed terms. It is strategic to offer some easy vocabulary (e.g. *tiger* or *land*) to a novice player, so that he is not discouraged and earns points quickly. But after a few games, most players prefer harder terms (e.g. *Higgs boson*), and it's adjustable in the game's options.

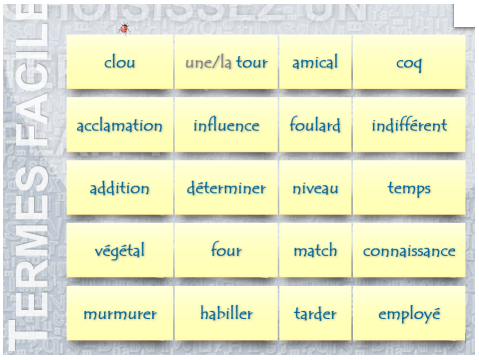


Figure 1: Selecting easy terms in JeuxDeMots amongst a randomized set of terms.

The ability to **offer games**, with the terms and relations of their choice, to other players (and even to attach to this gift a personal message). It is a way of entrusting the sampling of terms to the players, and thus increase their productivity: the players spontaneously choose interesting term / relation pairs, that is to say for which there are many and interesting answers.

A **chat** to communicate in real time with other connected players or the JDM administrator. This reinforces the sense of belonging to a community and allows them to help each other, to help newcomers and to guide them in discovering the many features of the games, to explain how to answer for difficult relations, exchange "tricks" to play better and earn more points, etc.

Figure 2: Offered Gifts in JeuxDeMots, allowing a relevant sampling of terms to be annotated.

However, the most important development was the creation of these "satellite" games, in addition to the main game. For players, these games offer another type of interaction: many are click or vote games, fast, easily playable on a smarphone in common situations such as in a waiting room or public transport. For the lexical network under construction, these "satellite" games compensate for the bias of the main game. Some of them, like Totaki, validate the data collected (Joubert et al., 2011), others, like Askit, correct errors related to polysemy, others focus on specific types of relations: polarity of terms for likeIt, feelings and emotions for Emot, colors and appearance for ColorIt... Tierxical helps refine the relations weighting, Askyou allows to validate or invalidate pending proposals, etc.

1.2. Evolution of Players

It soon turned out that a significant percentage of players, very interested in the "purpose" dimension of the GWAPs, expressed the desire to take a more active and concrete part in the construction of the lexical network. It is for these players that was set up the Diko, a contributive interface: the volunteer players can go and make contributions directly in the entry and for the relation(s) that inspire them. (Lafourcade et al., 2015). This role of active contributor is well suited to people sensitive to the challenge of participatory or citizen science.
To minimize the risk of error related to these contributors, who remain amateurs, a system of validation of their contributions by majority vote has been set.

2. Obtained Resource: a Very Large Knowledge Base

The lexical-semantic network (dubbed RezoJDM), under permanent construction, has been produced by the players, contributors and automated inference mechanisms (aka bots) and can be considered as a knowledge base encompassing both common sense, specialized and lexical informations.

Figure 3: A given play of JeuxDeMots and its outcome.

In addition to being typed (for example: *r_isa*, *r_agent*, *r_patient*, *r_domain*, etc.) a relation is also weighted. Its weight depends on the number of players who have proposed it. A weight can be negative (< 0), it indicates a negative relation (for example, *an ostrich* ***can not*** *fly*). Similarly, a false relation is made negative rather than deleted, to keep in mind that it was proposed and then invalidated. Thus, since relations can be proposed by automated processes, negatively weighting a false relation avoids the system to propose the same erroneous relation in a recurring way. Thus, inference mechanisms can also rely on negative relations.

Needless to say that this resource evolves over time with the addition of new terms and relations (at the very least new named entities). Its construction is not supposed to ended one day (at least theoretically).

Since the startup, the network gained on average around 20000 terms and 1.4 million semantic relations each month. Although the progression is not strictly regular, it is globally linear in time and we do not observe (yet) a beginning of flattening of the progression curve.

2.1. Common Sense & Domain Knowledges

The RezoJDM is a knowledge base containing mostly common sens facts. In order to process texts from specific domains, some efforts were done to integrate specialty domains, for example in health domain (anatomy, medicine, radiology, oncology) (Lafourcade and Ramadier, 2016) or in culinary domain (cooking, ingredients, nutritional facts) (Clairet and Lafourcade, 2017).

2.2. Densification with Automatic Inference

New relations can be inferred from existing ones through automatic endogenous inference, or from other (external resources) by extracting exogenous semantic relations.

Endogenous inferences rely on mechanisms of deduction, induction and abduction (Zarrouk and Lafourcade, 2014). For example : *a cat is a feline* and *a feline has part claws*, so we can deduce that probably *a cat has part claws*.

Exogenous extraction of semantic relations is undertaken from other resources, such as Wikipedia (Lafourcade and Joubert, 2013), or from fictions (French literature) corpora or non fiction and journalistic (Le Monde) corpora.

The contributions are tagged with the name of their author, whether human or automatic mechanism and are pending validation, either through satellite games, or by a game administrator. As shown in (Zarrouk and Lafourcade, 2014), inferred relationships may be wrong, especially when the inference is made from polysemous terms. Manual intervention by an expert is then required.

2.3. Error Detection

Even though the error rate is relatively low in the JDM network, well below 0.1%, we have developed an automatic error detection mechanism, (Lafourcade et al., 2017). which, from a so-called "primary" error, reported by a player or a contributor, will detect and report the errors secondarily induced by the automatic mechanisms of inference.

3. Lessons from the JDM Experience

Our 10 years of experience and exploitation of the JDM model have allowed us to identify a number of characteristics that a GWAP must have in order to be sustainable. (Lafourcade and Joubert, 2013).

3.1. About the Gameplay

Ideally, a GWAP should:

- be attractive, fun and interesting, which is essential to attract a large number of players: such a game must present a ludic interest at the interface level to attact gamers, but even more at the content level in order to keep them;
- be easy to understand, both in terms of the game modes and instructions to respect; a too complex game, or requiring a long learning, will discourage a large number of players;
- arouse addiction : this is possible thanks to the features of the game, as for example the instant replay by simple click, but also the modalities of play and the possibilities of interaction with the other players (lawsuits, gifts, theft of words, duels,...) that encourage people to come back;
- allow the filtering of players : flatter and make them feel useful (which is true) but also make them feel guilty if they do not play well (eventually make them give up the game if they do not improve). It's a good way to keep only the good players and guarantee the quality of the produced resource.

3.2. Benefits for NLP

The durability of a GWAP certainly depends on its attractiveness to the players, but also on how it meets the expectations of its designer. He must be able, by comparing the data he gets to what he wanted to obtain, to make the adjustments and modifications necessary to obtain usable data. The advantages for the NLP community are multiple:

- The data obtained is the result of non-negotiated contributions since the two players whose answers will be confronted have no way of communicating.
- The resource obtained is low cost compared to that which would be built manually, and it is acquired quickly (more than 40000 relations per day);
- The data acquisition procedure is ethical, unlike other approaches, such as Amazon Mechanical Turk (Fort et al., 2011). The principle of GWAP does not raise any ethical problem as long as it remains free and does not offer prizes that look like disguised salaries.

3.3. Issues in Cheating and Vandalism

As shown in (Lafourcade et al., 2015), we also noticed some cases of cheating and vandalism:

- Cheating : some players have managed to bypass some restrictive game rules, such as time limitation. This kind of cheating does not question the quality of the resource obtained, but it may disgust and discourage the players who do not cheat, and this can result in a disaffection for the game. In the context of JDM, we noticed that it was the first hours which constituted the critical phase for this type of risk.
- Vandalism is intended to corrupt the database by knowingly inserting erroneous data. Designers must minimize this risk at all costs, as detecting errors introduced is quite difficult, and must be done manually by experts. In fact, we think it is almost impossible to detect this type of error in an automatic manner. The fact that we only validate the common answers of a pair of players who do not know each other limits the risk of vandalism. As a result, assuming that the system could be able to detect an incongruous information (which is already far from being obvious and which poses the insolvable question of criteria), to systematically classify it as wrong and eliminate it would be counterproductive: incongruous does not necessarily mean wrong.

4. Impact of Automated Inferences

As mentioned above, automatic extraction or inference of semantic relation is at the core of the development of the lexical network.

4.1. Bots Behaving as Players

We recall the principle of the game: the game of a player is compared to another game on the same term and the same instruction (type of relation), and the common answers supply the network. The other part is randomly selected by the system. How to be sure that a game with the same term and the same instruction is available?
To deal with this issue, we devised fake player (bot) which produce pending games when needed. Of course, for a given term and, if they are enough true player games available, no bot is invocated for generating games. The state of the network directly dictates the nature and quality of the bot's answers. In such a way, along with players, the network feeds itself.
Player bots make use of various strategies, but the principle is to select proposals (randomly between 10 and 40) from the network according to three criteria: a) the most activate relations, b) the least activated relations, and c) the relations waiting to be (in)validated. Thus, player bot may induce the validation of waiting and original contributions.
One should notice that a bot never plays against itself but only against true players. A player bot never contributes directly to the network, it does only indirectly through games done with human players.
The average number of common responses between a bot player and a human player is about 12, while that number is about 5 between two human players.

4.2. Bots Behaving as Contributors

Thanks to automatic mechanisms of inference, robots act as contributors and add relations to the network. These relations are proposals, which must be validated by the human players-contributors, who vote for or against. As mentioned above, the inference is done according to different approaches: deduction, induction, and various types of abduction.
Moreover, some bots are able to deduce certain rules from the structure of the network. A rule is a) a set of conditions that must be verified for a given term and b) a conclusion wich is a relation to be added to the term. For example : *$x r_isa 'animal aquatique'* $\rightarrow$ *$x r_lieu 'eau'* (Eng: *if $x is a kind of aquatic animal then $x could be located in water*).
A bot proposes a rule as soon as it finds at least 3 examples and no counter-example (negative relations). If validated (by human administrator), the rule is applied to the network and the found conclusion is directly inserted (no validation required). So far, 4469 rules have been validated and led to the automatic creation of over 50 million relations (out of 180 million in January 2018).
So far, the error rate of automatic contributions is less than 1 for 10000 and 97% of such errors have been automatically detected.

4.3. Snowball Effect

The automatic inference mechanisms work from what is already validated in the network. To give a simple example of inference based on deduction, if we know that *pigeon* is_a *oiseau* (*bird*), then *pigeon* will inherit the general properties of *bird* (that is to say, semantic relations of *bird* with other words).
As a mean, each relation introduced by a player in the lexical network leads to 57 new correct relations inferred (from various bots and strategies), and the number of incorrect proposed relations tends to decrease as the network grows, from 20 in 2012 to 13 in 2014 and finally 5 in 2016.

We estimated that without the action of bot-players nor the mechanisms of inference the number of relations in RezoJDM would be of around 3 million, instead of more than 180. In addition, as the snowball effect results in an increase in the number of relationships validated automatically via the game (because the number of common responses between a bot and a human player is statistically higher than between two human players), both quality and quantity of the data collected is much better than it would have been with only human players.

5. Conclusion

The JDM project has largely demonstrated the interest of combining GWAPs and inference mechanisms to build a reliable and large-scale lexico-semantic resource. More precisely, this resource has been built largely by the activity of players and direct contributors, but also critically supplemented by mechanisms of automatic inferences. Those mechanisms have been instrumental concerning the significant volume and quality of the resource.

Our approach is monolingual and language independent. As a research perspective, we are currently developing a multilingual game similar to JDM, with which we expect to obtain a very large lexical database in a large amount of various languages. Such an approach could be especially instrumental in collecting cross-lingual lexical information for languages with a reduced number of speakers.

References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bos, J. and Nissim, M. (2015). Uncovering noun-noun compound relations by gamification. In Beáta Megyesi, editor, *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 251–255.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008). Phrase Detectives: a web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*.

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013). Using games to create language resources: Successes and limitations of the approach. *Gurevych I., Kim J. (eds) The People's Web Meets NLP. Theory and Applications of Natural Language Processing. Springer*, oct.

Clairet, N. and Lafourcade, M. (2017). Towards the automatic detection of nutritional incompatibilities based on recipe titles. In *CD-MAKE 2017*, pages 346–366. Reggio di Calabria, Italy, Aug 29-Sep 1, 2017.

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

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

Joubert, A., Lafourcade, M., Schwab, D., and Zock, M. (2011). Evaluation et consolidation d'un réseau lexical via un outil pour retrouver le mot sur le bout de la langue. In *Proc. of the 18th conference on Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier.

Lafourcade, M. and Joubert, A. (2013). Bénéfices et limites de l'acquisition lexicale dans l'expérience jeuxdemots. In *Ressources Lexicales: Contenu, construction, utilisation, évaluation*, pages 187–216. Linguisticae Investigationes, Supplementa 30, John Benjamins.

Lafourcade, M. and Ramadier, L. (2016). Semantic relation extraction with semantic patterns experiment on radiology reports. In *10th edition of the Language Resources and Evaluation Conference (LREC 2016)7*. Portoroz, Slovenia, Aug 23-28 May 2016, 6 p.

Lafourcade, M., Joubert, A., and Le Brun, N. (2015). Games with a purpose (gwaps). page 158. Wiley-ISTE, July.

Lafourcade, M., Joubert, A., and Le Brun, N. (2017). If mice were reptiles, then reptiles could be mammals or how to detect errors in the jeuxdemots lexical network? In *Proc.of International Conference on Recent Advances on Natural Language Processing (RANLP 2017)*, Varna, Bulgaria, September.

Lafourcade, M. (2007). Making people play for lexical acquisition. In *Proc. of the 7th Symposium on Natural Language Processing (SNLP 2007)*, Pattaya, Thailand, December.

Liu, H. and Singh, P. (2004). Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, oct.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, December.

VonAhn, L. (2004). Labelling images with a computer game. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 319–326.

Zarrouk, M. and Lafourcade, M. (2014). Inferring knowledge with word refinements in a crowdsourced lexical-semantic network. In *In proc. of the the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 346–366. Dublin, Irlande, 2014, 9 p.