WinoFlexi: A Crowdsourcing Platform for the Development of Winograd Schemas

Nicos Isaak[1 [0000−0003−2353−2192]] and Loizos Michael[1,2]

[1] Open University of Cyprus
nicos.isaak@st.ouc.ac.cy loizos@ouc.ac.cy
[2] Research Center on Interactive Media, Smart Systems, and Emerging Technologies

Abstract. The Winograd Schema Challenge, the task of resolving pronouns in certain carefully-structured sentences, has received considerable interest in the past few years as an alternative to the Turing Test. Systems developed to tackle this challenge have typically been evaluated on a small set of hand-crafted collections of sentences, since the development of new sentences by individuals is itself a rather challenging task, requiring care and creativity. In this paper we approach the problem of developing Winograd schemas via the introduction of *WinoFlexi*, a flexible online crowdsourcing system. Our empirical evaluation of the system's performance suggests that *WinoFlexi* allows crowdworkers to develop Winograd schemas of quality similar to that of most typical existing collections.

Keywords: Winograd Schema Challenge, Crowdsourcing

1 Introduction

The Winograd Schema Challenge (WSC) has been proposed as a novel litmus test for machine intelligence. Unlike the Turing Test, which is based on short free-form conversations during which a machine attempts to imitate a human, machines passing the WSC are expected to demonstrate the ability to think without having to pretend to be somebody else [1]. Passing the challenge requires resolving pronouns in certain sentences where shallow parsing techniques seem not to be directly applicable, and where the use of world knowledge and the ability to reason seem necessary [2, 3]. Although the challenge is, by design, easy for humans, the development of new Winograd schemas is, itself, too troublesome for humans lacking inspiration and creativity [4].

In this paper, we present *WinoFlexi*, a flexible online collaboration system that allows members of crowdsourcing platforms to collaborate *explicitly* for the development of Winograd schemas. To the best of our knowledge, this is the first work that attempts to use crowdsourcing for this task. We envision the use of this platform as a source of Winograd schemas for use in WSC-based CAPTCHAs [5] and in WSC competitions for the evaluation of systems that attempt to pass the challenge [4].

WinoFlexi uses a combination of tools that enhance the schema-development process: *i)* it is more cheat-proof than existing crowdsourcing platforms, and *ii)* it uses test questions that are closer to the schema-development process that benefit non-dubious workers and ban dubious ones. Our empirical study with workers from an existing crowdsourcing platform, showed that *WinoFlexi* can be used for the development of Winograd schemas that are comparable to the most typical existing schema collections.

2 The Winograd Schema Challenge

Winograd schemas comprise of two Winograd halves, with each half consisting of a sentence, a definite pronoun or a question, two possible pronoun targets (answers), and the correct pronoun target [1]. The following schema (a pair of halves) illustrates the key characteristics of Winograd schemas: *1.) Sentence: Erica called Jennifer on the phone because she was not responding to email. Question: Who was not responding to email? Answers: Jennifer, Erica. Correct Answer: Jennifer. 2.) Sentence: Erica called Jennifer on the phone because she was not able to email. Question: Who was not able to email? Answers: Jennifer, Erica. Correct Answer: Erica.*

Given just one of the halves in a schema, the aim is to resolve the definite pronoun in the question to one of its two co-referents. The avoid trivializing the task, the co-referents are of the same gender, and are either both singular or both plural. The two halves differ in a special word or phrase that critically determines the correct answer. Schemas that do not *strictly* follow these rules are called "schemas in the broad sense".

It is believed that the WSC can provide a more meaningful measure of machine intelligence when compared to the Turing Test, exactly because of the presumed necessity of reasoning with commonsense knowledge to identify how the special word or phrase affects the resolution of the pronoun. By extension, it is believed that a system that contains the commonsense knowledge to correctly resolve Winograd schemas should be capable of supporting a wide range of AI applications. Although, as expected from its reliance on commonsense knowledge, English-speaking adults have no difficulty with the challenge, the development of the schemas themselves is a very challenging task [4]. According to Levesque [1] in order to build quality Winograd schemas one needs to avoid two pitfalls: having questions whose answers are in a certain sense too obvious, and (more importantly) having questions whose answers are not obvious enough.

To the best of our knowledge, the availability of Winograd schemas is limited. Currently, only two widely-used WSC collections exist: *i)* Rahman and Ng's collection [6], which consists of 942 schemas and was developed by students (built under the "broad sense"); *ii)* Levesque and Davis's [1] collection, which consists of 150 schemas and was developed under the strict rules of the WSC (referred to later as the Winograd-library).

The availability of Winograd schemas seems disproportional to their demand and their potential impact. A recent study [5] showed that the WSC can form the basis of a new type of CAPTCHA, which might encourage more AI researchers to work on the problem of actually trying to tackle the WSC, and perhaps, to help towards the building of machines able to reason with commonsense knowledge. On the other hand, the development of carefully-crafted pronoun resolution tasks towards the development of Winograd schemas is a hard process [4]: it requires creativity and inspiration, and it is too troublesome to be done on a regular basis to support, for instance, competitions on the WSC or the testing of systems that might have been trained on existing collections of Winograd schemas. Perhaps not unrelated to the limited availability of Winograd schemas is the fact that the first and only WSC competition was organized in 2016 [4].

Towards addressing this disparity, we turn to crowdsourcing. Currently, many skilled labor activities are carried out online via crowdsourcing platforms. These platforms can eliminate geographic constraints and help workers to pursue work that they find valuable [7]. This work utilizes such platforms to develop *WinoFlexi*, in an effort to bring

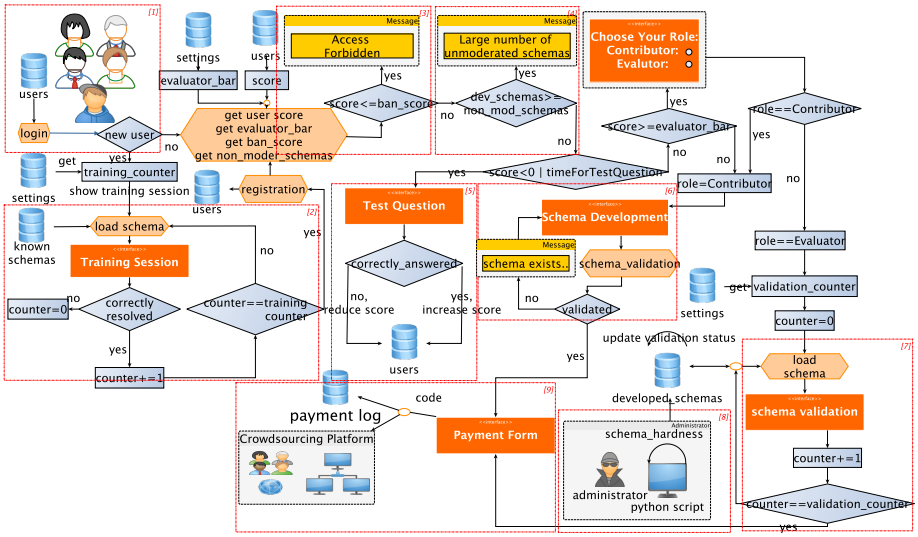

Fig. 1: *WinoFlexi*'s Architecture for the Development of Winograd Schemas. The various parts of the architecture are marked in red rectangles, and are discussed in Section 3.

together researchers and people from across disciplines, concerned with the acquisition and use of language data in the context of knowledge-based applications like the WSC. The design of appropriate crowdsourcing mechanisms for our particular task and the evaluation of the developed Winograd schemas is the focus of the rest of this paper.

3 Crowdsourcing Platform Architecture

We continue to present our platform and its constituent modules (see Fig. 1), and discuss how the crowd collaborates to built schemas under *WinoFlexi*'s evaluation mechanisms. Recognizing that the schema development process is tedious and troublesome, *WinoFlexi* is built to act as an assistant with effective incentive mechanisms for the crowd.

3.1 Registration and Training Session

The first step for each worker is to apply as a Contributor to our platform, where they register their credentials (`http://cognition.ouc.ac.cy/mcSchemaBuilder`; see *part-1* in Fig. 1). Workers need not be domain experts but need to have a strong command of English to ensure that schemas have no spelling, syntactic, or grammatical errors, and comply with the schema development process. To maximize the quality of the developed schemas, every Contributor has to complete a training task (see *part-2* in Fig. 1). During the training phase workers are familiarized with the development process by being asked to correctly resolve randomly selected schemas from the

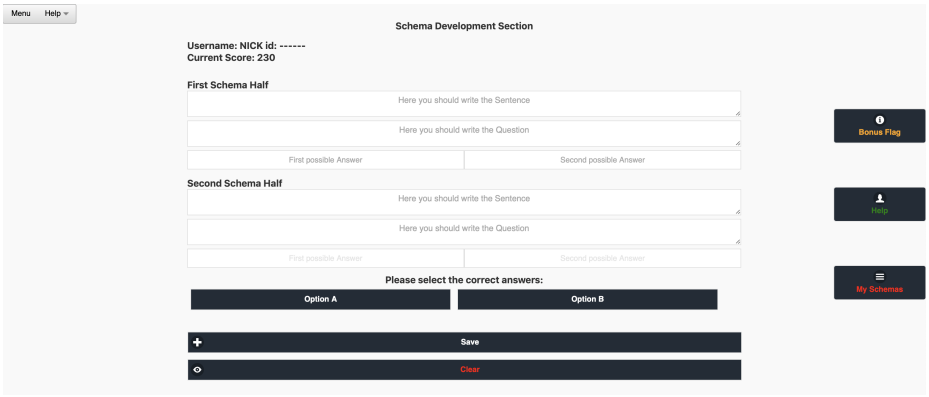

Fig. 2: The Contributor Dashboard.

Winograd-library. The length of the training phase can be increased either by the system administrator or automatically by *WinoFlexi* to ensure that the quality of the produced schemas meets expectations. In particular, if the *auto-training* flag is enabled, then the length of the training phase for every new registered Contributor is determined by how much the number of invalid schemas produced so far exceeds the number of valid ones.

3.2 Contributing and Evaluating

Workers both contribute in the development of schemas, and evaluate their quality.

Contributors: Contributors are workers who develop schemas (see *part-6* in Fig. 1), using the dashboard shown in Fig. 2. When a Contributor adds a schema, *WinoFlexi* does some basic checks: *i)* It checks if each schema half comprises a sentence, a question, and two pronoun targets. *ii)* It checks if the correct pronoun target of each schema half has been selected. *iii)* It checks if the sentence, the question, and the two pronoun targets of each schema half are related. *iv)* It checks if the two halves are related. Relatedness is checked using the heuristic approach shown in Fig. 3 applied to each of the pairs sentence-question, sentence-first_pronoun_target, sentence-second_pronoun_target.

Evaluators: Workers who validate schemas are called Evaluators (see *part-7* in Fig. 1). Contributors are allowed to take on this second role if they meet two requirements: first, the percentage of their valid and approved (by other Evaluators) schemas among those that they have contributed that far exceeds a certain threshold (which we have set to be 90%, corresponding to the bar for *near adult human* abilities on the WSC [3]); second, their score (which we discuss later) is above a certain other threshold. Contributors who are also Evaluators choose the role in which they interact with *WinoFlexi* at login time. At the beginning of the development process, the only Evaluator is the system administrator. The evaluation process comprises of answering a number of yes/no questions using the dashboard shown in Fig. 4. Affirmative responses to all but the first question are necessary to characterize a schema as valid. Additionally, the Evaluators have access to a similarity tool to detect if the Contributors are following a pattern to develop

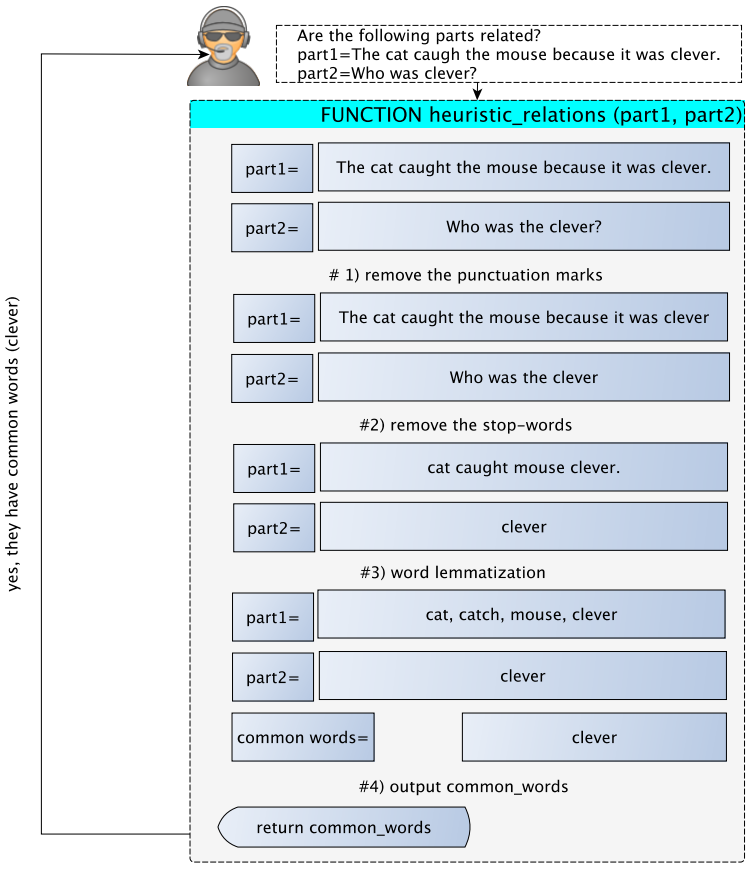

Fig. 3: Heuristic Relations to Eliminated Problems with Schema Cohesion.

similarly-looking schemas. The tool acts like a *leakage-detector* [7] that queries the WinoFlexi-library and Winograd-library to determine if a newly-contributed schema is "leaked", in that it is significantly similar to an existing schema. Each approved schema increases the Contributor's score and each "leaked" schema decreases it, affecting whether the Contributor will meet the requirements to become an Evaluator.

3.3 Quality-Assurance Measures

Additional mechanisms are used to ensure the quality of the developed schemas.

Test Questions: Many crowdsourcing platforms use tests as a method of assessment, offering their certification mechanisms to verify that a given worker indeed holds a particular skill [7, 8]. Previous works indicate that more interactive studies may motivate participants to read instructions more carefully leading to better compliance [9]. Our approach is based on the adaptive interjection of test questions and on rewarding

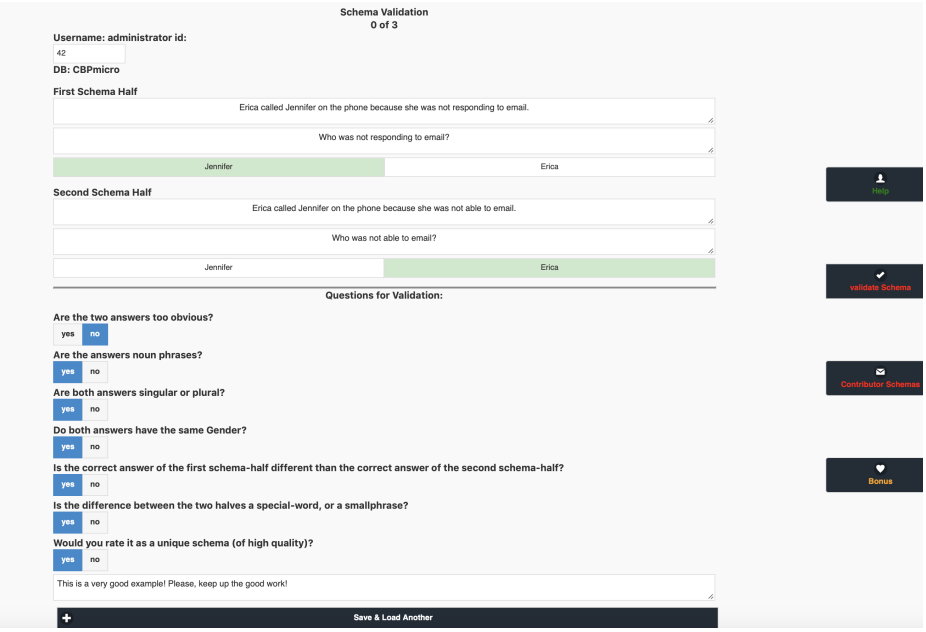

Fig. 4: The Evaluator Dashboard.

the worker with a positive score for successfully resolving them (see *part-5* in Fig. 1). *WinoFlexi* can be enabled to display test questions as often as necessary, to both Contributors and Evaluators; this can be manually handled by the system administrator, or automatically controlled by the system. By default, a test question has a 10% probability of being displayed after every login. If the *auto-testing* flag is enabled, this probability is adjusted in a manner analogous to how the length of the training phase is adjusted. Test questions are selected from the WinoFlexi-library (validated contributed schemas) and the Winograd-library; both collections include schemas that strictly follow the WSC rules. Correct / wrong answers to test questions increase / decrease a worker's score.

Ban Score: Online certification of skills is still problematic, since dealing with cheating is a major challenge. The *ban-score* mechanism automatically bans workers who have a sufficiently low score (see *part-3* in Fig. 1), with the threshold identified empirically.

Un-Validated Schemas: To prevent workers from entering a large number of potentially invalid schemas, there is a mechanism that limits the number of schemas each worker can develop before they undergo the validation process (see *part-4* in Fig. 1).

Winograd Schema Hardness: *WinoFlexi* leverages existing tools for the WSC to generate feedback to the Contributors (see *part-8* in Fig. 1). Towards this goal, we follow a single-step approach for labeling schemas with a hardness score which indirectly shows if a schema is considered hard to answer by a machine; Winograd schemas are accordingly labeled as such by the computed hardness index. For this purpose we use a recent

tool [3] that can take any Winograd schema and output a score that shows its hardness index. The hardness index is presented to the Contributors and the Evaluators. If the majority of a Contributor's schemas are easy (respectively, hard) then our system prompts them to develop schemas that are harder (respectively, easier) to solve.

3.4 Payment and Rewards

Payment Procedure: Most of the microtasks on the crowdsourcing platforms are priced individually, and workers are paid a base rate multiplied by the number of correctly completed tasks. Whatever their motives are, workers want to earn money and seek out tasks to maximize their expected earnings. To make sure that only the workers who developed schemas are going to get paid, we enhanced *WinoFlexi* with a payment verification plug-in (see *part-9* in Fig. 1). Upon each schema development (or validation), Contributors and Evaluators are prompted with a notification message and a code which is automatically generated and inserted into our database. Each worker has to provide the same code on their crowdsourcing platform to receive the actual payment.

Rewards: Workers, recruited through crowdsourcing platforms, must receive a small fixed payment for participating in the experiment, and/or a bonus for high quality results [8]. Past work has shown that the quality of work produced in a crowdsourcing working session can be influenced by the presence of financial incentives, such as bonuses. *WinoFlexi* adopts this philosophy and rewards Contributors based on "relative performance", namely only the worker that performs best receives rewards.

4 Experimental Design and Results

In recent years, a growing number of researchers have been using well-known crowdsourcing platforms [9]. A large body of work has shown **MicroWorkers (MW)** to be a reliable and cost-effective source for various fields and research purposes [8, 9]. Platforms like MW offer a framework that enables the employers to submit individually designed tasks to the crowd. MW has almost 1.5 million subscribed workers, and offers more than 40 million tasks. The MW platform offers many features which can influence the completion time and the results. Moreover, it provides campaign creators with predefined groups of workers from different regions that are organized according to their skills (e.g., best rated countries, writers, workers with certain language qualification tasks). To attract the worker's attention we used a simplified title (*title: Develop Groups of Sentences, Questions & Answers that Meet Certain Criteria*) and promoted it on the MW platform. Workers were given instructions explaining the task directing them to develop schemas without sacrificing accuracy. It was made clear that the development of invalid schemas might ban them from the system. Furthermore, we promoted *WinoFlexi* only under the Hired-Section of *English Speaking Countries + En*, meaning that only members of that group were able to participate. Our selected workers have both English proficiency, and admission tests passed. For our task, we offered a compensation of $1.00 for each developed schema or for the validation of three schemas in a row. We also advertised a bonus for quality schemas without stating the amount.

Table 1: Snapshot of the Contributors' Developed Schemas on *WinoFlexi*.

1	Erica called Jennifer on the phone because she was not responding to email.	Who was not responding to email?	Jennifer, Erica
	Erica called Jennifer on the phone because she was not able to email.	Who was not able to email?	
2	If Rachel listened to Mrs. Sheila, she would have given her full marks.	Who would give full marks?	Mrs. Sheila, Rachel
	Had not Rachel ignored Mrs. Sheila, she would have got full marks.	Who would have got full marks?	
3	The martial artist defended himself from the drug dealer because he was violent.	Who was violent?	The drug dealer, The martial artist
	The martial artist defended himself from the drug dealer because he was under attack.	Who was under attack?	

The experiments ran for one week, and yielded more than 165 schemas (see Table 1), from 50 workers, aged 18 to 65. From the developed schemas, 135 (81%) were valid, and 30 invalid. The highest score of a worker was 250 points and the lowest was -70; the Contributor with the lowest score was automatically banned by *WinoFlexi*. The majority of the workers had a non-negative score, and the top three workers had a score of at least 170, which well-exceeded the second condition for qualifying as an Evaluator. The total cost of our campaign was $258.00. The Contributors were paid $165.00 for the schema development process, with an additional $63.00 given as bonuses. On the other hand, $30.00 were paid to Evaluators for the schema evaluation process.

Our experimental evaluation shows that *WinoFlexi* supports the development of *valid* schemas, with a cost of approximately $1.91 per schema. Considering the challenge difficulties, we believe that this is a fair cost. Mean response time across all workers was 1.48 minutes, and the average time for the best worker was 1.66 minutes. 60% of the bonuses were offered to the top five workers. We believe that our adopted approach leads to more bonus opportunities for workers who submit schemas of good quality.

Evaluators were not observed to show a preference for the evaluation process over the schema design process. Although the evaluation process seems more straightforward, workers might have preferred the schema design process for the following reasons: *i)* they were more familiar with the schema design process than the evaluation process; *ii)* through the schema design process they were eligible for rewards, such as cash bonuses; *iii)* they did not want to leave other Contributors unpaid, or lower their score.

The general picture emerging form the analysis above is that *WinoFlexi* is a platform where workers can collaborate for the schema development process. However, there is a key question when considering this approach that we have not addressed yet: How does the quality of the developed schemas compare to that of schemas in existing collections?

4.1 Quantitative Analysis

Co-reference Resolution: Our baselines are three co-reference resolution systems that were used on the Winograd-library [4], namely the *Stanford-Core-NLP* system, *Wikisense* [10], and *Knowledge-Parser* [2]. Showing a positive correlation of the performance of the three systems on the Winograd-library and the WinoFlexi-library would offer evidence that *WinoFlexi* can be used to develop schemas of good quality. For our experiment, we randomly selected 50 schemas (100 schema-halves) from each library. On the

Winograd-library, **Stanford-Core-NLP** correctly resolves 37% schema-halves, incorrectly resolves 39% of them, and does not make any decision on the remaining 23%. On the WinoFlexi-library, it correctly resolves 44% schema-halves, incorrectly resolves 44% of them, and does not make any decision on the remaining 12%. **Wikisense** correctly resolves 59% schema-halves of the Winograd-library, incorrectly resolves 31% of them, and does not make any decision on the remaining 9%. On the WinoFlexi-library, it correctly resolves 56% schema-halves, and incorrectly resolves 44%. **K-Parser** correctly resolves 38% schema-halves of the Winograd-library, incorrectly resolves 36%, and does not make any decision on the remaining 26%. On the other hand, on the WinoFlexi-library, it correctly resolves 37% schema-halves, incorrectly resolves 37% of them, and does not make any decision on the remaining 26%. Comparison of the results shows that the performance of the three systems on the WinoFlexi-library is analogous to their performance on the Winograd-library. According to our results, the two libraries have correlation coefficients of 0.925 (Stanford-Core-NLP), 0.987 (Wikisense), and 0.995 (K-Parser), respectively. The results provide evidence that our developed schemas are of the same or similar quality with the Winograd-library schemas.

Hardness Metric Tool: For the purpose of this experiment, we randomly selected 57 schema-halves of the WinoFlexi-library, and compared their hardness index to that of 57 schema-halves of the Winograd-library taken from a previous work [3]. Fig. 5 shows in more detail how the computed hardness index varies across schema-halves, suggesting that indeed, the two sets have comparable average hardness indices and analogous variability in their hardness indices. The general picture emerging from the analysis shows that despite the fact that our workers were not initially familiar with the schema development process, through *WinoFlexi*'s mechanisms they were trained to design schemas of good quality. Furthermore, the data presented here provides evidence that the *WinoFlexi* schemas avoid Levesque's pitfalls, meaning that the questions of the schemas are neither too obvious, nor are their answers not obvious enough.

Schema Structure: Next, we compare the structure of all the crowd-generated schemas (WinoFlexi-library) to that of all the expert-generated schemas (Winograd-library), as a way to determine if using crowdworkers sacrifices quality in exchange for scalability.

For this experiment, we developed a tool that identifies the sentence pattern of each designed schema. Given as input an English sentence, it outputs its pattern/type which can be either a simple, a compound, a complex, or a compound-complex sentence. Simple sentences have only one independent clause (SV; where S=Subject and V=Verb), while compound sentences can have two or more independent clauses (e.g., "SV and SV"). On the other hand, complex sentences can have one independent clause plus one or more dependent clauses (e.g., "SV because SV"), and compound-complex sentences can have two or more independent clauses plus one or more dependent clauses (e.g., "SV and SV because SV."). The connector in each complex sentence shows how the dependent clause relates to the independent clause. Based on the typical connectors found in Winograd schemas, we consider the following groupings of connectors for our analysis: *i)* Cause/Effect: because, since, so that; *ii)* Comparison/Contrast: although, even though, though; *iii)* Place/Manner: where, how, however; *iv)* Possibility/Conditions: if, whether, unless; *v)* Relation: that, which, who; *vi)* Time: after, as, before.

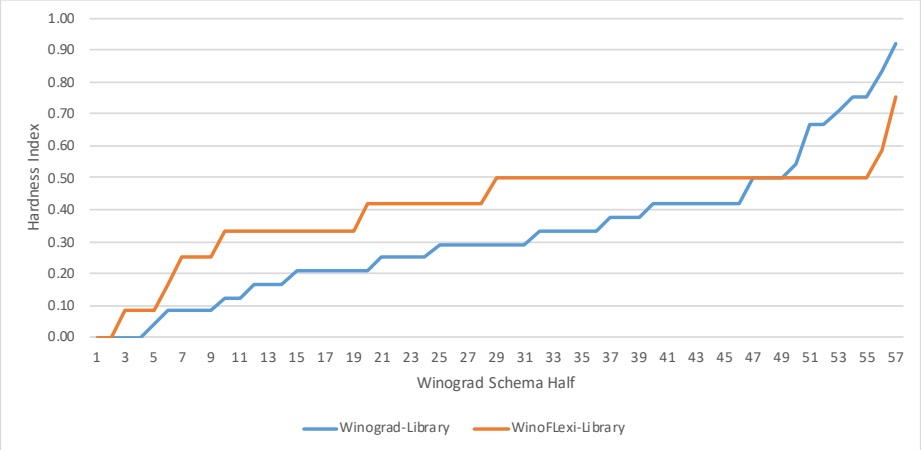

Fig. 5: Hardness Index Variability across 57 Schema Halves of the Winograd-library and 57 Schema Halves of the WinoFlexi-library. Each group is sorted based on the hardness index.

The results showed that 9% of the crowd-schemas are based on simple sentences, 8% on compound sentences, and 83% on complex sentences. On the other hand, 41% of the expert-schemas are based on simple sentences, 14% on compound sentences, and 45% on complex sentences. Most of the developed schemas (both expert and crowd) are based on complex sentences. The expert-schemas that were designed with complex sentences had 30% "Cause/Effect", 8% "Comparison/Contrast", 1% "Place/Manner", 4% "Possibility/Condition", 18% "Relation", and 39% "Time" relationships. On the other hand, the crowd-schemas had 52% "Cause/Effect", 1% "Comparison/Contrast", 2% "Possibility/Condition", 1% "Relation", and 44% "Time" relationships. The results provide evidence that with *WinoFlexi*'s help the crowd was able to develop quality schemas that are based on a variety of sentence patterns, similar to the expert developed schemas. Additionally, the fact that crowd-schemas are not based on simple sentences, like the expert-schemas are (41%), might show that the crowd did not sacrifice quality in exchange for scalability. Considering the challenge difficulties, it seems that *WinoFlexi* can motivate and inspire researchers for the faster development of new schemas.

4.2 Qualitative Analysis

Based on the valid developed schemas, and taking into account comments received from Contributors, we present below a qualitative analysis of *WinoFlexi*'s outputs.

Evaluation Procedure: Certain outputs suggest that *WinoFlexi*'s evaluation might need to be optimized, and schemas might need to be evaluated by more than one Evaluator. For instance, the following was mistakenly considered as a valid schema: *1.) Sentence: Karen loved going to salons to get her nails done. They always looked so nicely decorated. Question: What looked nicely decorated? Answers: The Salons, The Nails. 2.) Sentence: Karen loved going to salons to get her nails done. They always looked so*

nicely manicured. Question: What looked nicely manicured? Answers: The Salons, The Nails. This schema cannot be considered as a valid one because the second half is resolvable with selectional restrictions; salons cannot be manicured.

Inspiration and Creativity: One of the problems during schema development is the lack of inspiration and creativity. It seems that the collective intelligence of the crowd is able to mitigate this issue. For instance, the workers developed schemas which are based on a variety of subjects, like cartoon heroes (spiderman, hulk), animals (hyenas, zebras), hospitals (psychiatrists, medications), people in general (fights, burglars), things (cards, drains). The following is an example schema: *1.) Sentence: Spiderman spun his web around the Hulk because he was falling. Question: Who was falling? Answers: Hulk, Spiderman. 2.) Sentence: Spiderman spun his web around the Hulk because he was annoyed. Question: Who was annoyed? Answers: Hulk, Spiderman.*

Enjoyment and Curiosity: Based on comments that we received, certain workers were motivated by an intrinsic incentive such as enjoyment and curiosity for new knowledge, and not only from potential rewards. Worker *Member0xx*, for example, offered the following comment: *"I am terribly sorry, on my most recent schema I accidentally selected the wrong option. The schema is about putting a shirt in the dryer. I hope it is something you can fix. Thank you for your time and allowing a platform to develop these schemas, I very much enjoy trying to figure out new ways to create a valid schema."*.

5 Conclusion and Future Work

We have presented *WinoFlexi*, an online crowdsourcing system built explicitly for the development of Winograd schemas. Despite the acknowledged difficulty of the task when assigned to individuals, our empirical evaluation offers evidence that online crowd platforms and systems like *WinoFlexi* might offer a viable alternative.

Among possible directions for future research, of interest would be the automation of parts of the process of schema development and validation, without taking humans out of the loop. Sentences upon which schemas could be built, for example, could be automatically detected by crawling the Web, and offered to the *WinoFlexi* crowdworkers for further processing and validation. This human-machine teaming might prove to lead to a more efficient utilization of human time, and might yield a more diverse set of schemas, perhaps expanding the creativity and inspiration of the crowdworkers. In terms of validation, one could attempt to identify heuristics employed by humans when evaluating schemas, and might seek to help Evaluators focus their attention to those aspects of a schema that might be more salient when determining its validity.

Acknowledgments

This work was supported by funding from the EU's Horizon 2020 Research and Innovation Programme under grant agreements no. 739578 and no. 823783, and from the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination, and Development. The authors would like to thank Ernest Davis for sharing his thoughts and suggestions on this line of research.

References

1. Hector Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema Challenge. In *Proceedings of the 13th International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.
2. Arpit Sharma, Nguyen H. Vo, Somak Aditya, and Chitta Baral. Towards Addressing the Winograd Schema Challenge - Building and Using a Semantic Parser and a Knowledge Hunting Module. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 25–31, 2015.
3. Nicos Isaak and Loizos Michael. A Data-Driven Metric of Hardness for WSC Sentences. In Daniel Lee, Alexander Steen, and Toby Walsh, editors, *Proceedings of the 4th Global Conference on Artificial Intelligence*, volume 55 of *EPiC Series in Computing*, pages 107–120. EasyChair, 2018.
4. Leora Morgenstern, Ernest Davis, and Charles L. Ortiz. Planning, Executing, and Evaluating the Winograd Schema Challenge. *AI Magazine*, 37(1):50–54, 2016.
5. Nicos Isaak and Loizos Michael. Using the Winograd Schema Challenge as a CAPTCHA. In Daniel Lee, Alexander Steen, and Toby Walsh, editors, *Proceedings of the 4th Global Conference on Artificial Intelligence*, volume 55 of *EPiC Series in Computing*, pages 93–106. EasyChair, 2018.
6. Altaf Rahman and Vincent Ng. Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
7. Maria Christoforaki and Panagiotis Ipeirotis. Step: A Scalable Testing and Evaluation Platform. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing*, 2014.
8. Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. Anatomy of a Crowdsourcing Platform — Using the Example of microworkers.com. In *Proceedings of the 5th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pages 322–329. IEEE, 2011.
9. Eyal Peer, Sonam Samat, Laura Brandimarte, and Alessandro Acquisti. In Kristin Diehl and Duluth Carolyn Yoon, editors, *Beyond the Turk: An Empirical Comparison of Alternative Platforms for Crowdsourcing Online Research*, volume 43 of *NA - Advances in Consumer Research*, pages 18–22. MN : Association for Consumer Research, 2015.
10. Nicos Isaak and Loizos Michael. Tackling the Winograd Schema Challenge Through Machine Logical Inferences. In David Pearce and Helena Sofia Pinto, editors, *STAIRS*, volume 284 of *Frontiers in Artificial Intelligence and Applications*, pages 75–86. IOS Press, 2016.