Establishing a Human Baseline for the Winograd Schema Challenge

David Bender
Center for Research on Concepts and Cognition
Indiana University
512 North Fess Avenue
Bloomington, Indiana 47408

Abstract

The Winograd Schema Challenge (WSC) is a pronoun resolution task for which deep semantic knowledge is required to achieve high performance. Until now it has been assumed that human performance on the WSC is nearly at ceiling, but evidence for this has been mainly anecdotal. Here we present the results of a large online experiment that both establishes a baseline for human performance on the WSC and demonstrates the importance of human testing, not only as a means of validating a particular corpus, but more fundamentally as a guide in defining desirable characteristics for Winograd Schemas (WS).

The Winograd Schema Challenge

In recent years, several tasks have been proposed which focus on the difficultly of understanding and reasoning with natural language. Typically these tasks involve reading a short fragment of text and then selecting the best answer or alternative from a following list.

For example, the Choice of Plausible Alternatives (COPA) challenge emphasizes causal reasoning, where a given premise is most likely connected to only one of a number of possible choices (Roemmele, Bejan, and Gordon 2011). Similarly, in the domain of Recognizing Textual Entailment (RTE), two fragments of text are presented with the implicit question: does the former logically entail the latter (Dagan, Glickman, and Magnini 2006)?

Some have criticized COPA and RTE as being too restrictive (Levesque 2011). By framing the allowable associations in logical or causal terms, these challenges rule out subtle judgments based on subcognitive pressures and aesthetic preferences. As an alternative, Levesque has proposed the Winograd Schema Challenge (WSC), a pronoun disambiguation task named after Terry Winograd, who first discussed the form (Winograd 1972).

The WSC is made up of individual problems called Winograd Schemas (WS). Each WS contains two nearly identical sentences with clear but very different meanings.

> The older kids were bullying the younger ones, so we rescued them. Whom did we rescue?
>
> The older kids were bullying the younger ones, so we punished them. Whom did we punish?

Each sentence is followed by a question, and to answer it correctly requires that one determine, for an ambiguous pronoun, which antecedent is its likely referent. In the first sentence above, the pronoun *them* plainly refers to the younger students, whereas in the second sentence the same pronoun refers to the older kids. The bullies are punished, not their victims.

Notice that the only difference between the two sentences is that "rescued" is changed to "punished". These are usually called the WS's *special words*, though perhaps a better term would be its *fulcrum*, for they act as the pivot point around which the meaning of the entire sentence shifts.

Many cases of anaphoric resolution are relatively simple. Syntactic cues such as gender agreement or plurality constraints are often sufficient to determine the antecedent of a reference. In contrast, well constructed WS problems are thought to be exceedingly difficult to solve analytically. Instead of focusing on surface-level syntax, people understand examples like the one above, often seemingly without effort, because they have access to a wealth of knowledge about the real world that they are able to bring to bear when they read and think about a situation.

Beyond syntactic constraints, Levesque et al. have identified three potential flaws that can invalidate a WS, or make it too easy. These include **selectional restrictions** where concepts evoked by words in the fulcrum are only applicable to one of the possible antecedents, **statistical correlations**, where words in the fulcrum are more similar to one antecedent than the other (e.g. by their mutual information score in a large corpus, or number of pages returned in a web search), and **one-way ambiguity**, where in one version of the WS the pronoun seems equally likely to refer to either antecedent. (See Levesque, Davis, and Morgenstern (2012) for examples.)

In this paper, WS whose answers can be found using these techniques (or other simple syntactic cues) will be referred to as *Easy WS*. WS resistant to these techniques will be called *Hard WS*.

Determining a Human Baseline

Levesque, Davis, and Morgenstern have created an online corpus of more than 140 WS examples[1]. They predict these

[1] Available at https://www.cs.nyu.edu/davise/papers/WS.html

would be correctly answered by average English-speaking adults extremely easily, with overall accuracy "presumably close to 100%" (Levesque, Davis, and Morgenstern 2012, p.557).

Recently the nonprofit organization Commonsense Reasoning, in cooperation with Nuance Corporation, has announced the beginning of an annual WSC competition,[2] with an inducement prize awarded to the team or individual who produces a program capable of meeting a baseline established for human performance. Because questions for a given year will not be released in advance, the existing online corpus is intended to be used by participants while developing their systems.

But how well do people actually perform on this set of questions? Answering this empirical question is the primary aim of this paper. We hope that determining a baseline for human performance on this corpus will be helpful to those interested in entering the WSC competition, and that it will serve as a standard reference they can use to measure their modeling efforts. We also hope the organizers of the WSC are able to use this work to improve their online corpus.

Methods

Amazon's Mechanical Turk (MT) has been proven as a compelling alternative to laboratory studies across a range of experimental tasks (Chandler, Mueller, and Paolacci 2014). (Also see (Mason and Watts 2010) for an overview.) Numerous psycholinguistic studies have been duplicated using MT, including those that depend on precisely controlling the presentation of stimulus materials and measuring response times at the millisecond level (Crump, McDonnell, and Gureckis 2013).

Participants

A single experiment was performed online using MT. More than four hundred volunteers participated in the study. Subjects were adult residents of the United States who speak English fluently, and were screened by means of a qualification task (see below). Out of 430 subjects who attempted the task, 23 did not finish. Of these, 9 either did not complete or did not pass the qualification task. The remaining 14 qualified, but did not complete the testing task. The 407 participants who completed the testing task were paid between $\$0.50$ and $\$1.50$ each, depending on how many questions they answered correctly, while those who did not finish (or did not qualify) were offered $\$0.25$ as compensation for their time.

Qualification Task

Identifying qualified subjects on MT requires time and effort, but it also yields an important methodological advantage. Workers who demonstrate that they take experiments seriously and possess the skills relevant to a given task produce more reliable data than workers simply trying to make money. For data collectors, this translates into less cause to remove outliers from the results based on performance (or some other *ad hoc* metric), even when such omissions can be arguably justified (Chandler et al.).

To maximize the quality of our subjects, we gave all participants a combined qualification/training task. This preliminary task familiarized subjects with the experiment by asking them to answer a few WS questions. Subjects were required to correctly answer at least 75% of the questions correctly. This cutoff, though arbitrary, was clearly not overly stringent, as less than 1% of all participants failed.

All subjects received the same training examples, and only subjects who passed the training task were allowed to continue on to the testing task. Results are reported for all participants that completed the testing task; none were discarded.

It is problematic, in general, to combine qualification and training tasks. If the training and testing tasks are the same, potential subjects will reach the testing task only if they have are already demonstrated skill at that same task. To avoid begging the experimental question – how well do humans perform at the WSC? – seven out of eight of the training questions were *Easy WS* questions as previously defined. Because the testing task was mostly made up of *Hard WS* questions, we predicted that the qualification task would be easier than the testing task. Our null hypothesis was that the two tasks were equally difficult.

One of the training questions is shown below. Note that the pronoun ("she" or "it" respectively) can be resolved syntactically by distinguishing between gendered and inanimate personal forms.

> The bird flew too close to Tara, so it swerved. Who swerved?
> The bird flew too close to Tara, so she ducked. Who ducked?

Design

The testing task contained 160 WS questions: 143 *Hard WS* from the online corpus (as of February, 2015), along with 17 *Easy WS* questions created specifically for this experiment.

In a volunteer sampling, randomized block design, each of the two versions of the 160 questions was answered at least 50 times, by different subjects, yielding a total of over 16,000 answers. Some questions received a few more than 50 answers (mean 50.9) because the experiment was performed by many subjects simultaneously, and blocks were allocated in a round robin fashion, without waiting for the failure or successful completion of each allocated block.

Each subject answered a block of 40 randomly selected questions with an average of 36 drawn from the set of *Hard WS* and 4 drawn from the set of *Easy WS*. No single subject was given both variants of the same WS.

Easy WS examples were added solely to test (and hopefully reject) the hypothesis that the training/qualification task and the testing task were equally difficult. In other words, we wanted to eliminate the qualification task as a threat to the testing task's internal validity, making sure that by excluding subjects who perform poorly on *Easy WS* examples, we were not preferentially selecting those who perform better at *Hard WS* examples than the general adult English-speaking population. We did not use the training questions themselves for this purpose in order to eliminate

[2] http://commonsensereasoning.org/winograd.html

confounding factors, such as increased effort during training, mistakes made due to learning the experimental controls, and so on.

For each subject we measured two dependent variables: accuracy, and response time for each question. This was primarily an exploratory design; individual differences were considered secondary, and the only factor of interest was the question category (*Easy WS* versus *Hard WS*). Finally, at the conclusion of the study, subjects were also asked to provide their age and offer any comments they might have. The comments were entered as unstructured text.

Materials

We used the psiTurk software system to design and administer the experiment (McDonnell et al. 2012). Subjects performed the task in a browser window. Both mobile phones and tablets were excluded because of their small screen size and the need for keyboard input.

All materials used in the experiment, including WS questions, JavaScript experiment code, instructions, software used for data retrieval and analysis, and more extensive detail regarding the corpus analysis, are available online at https://github.com/benderdave/wsc-exp.git.

Procedure

Subjects were tested individually over the internet and so no environmental controls were in place. Subjects were given instructions explaining the task and directing them to answer each question quickly but without sacrificing accuracy. Subjects read an informed consent form and agreed to participate in the study. All input was done using only four keys on the keyboard: ENTER, SPACE, 1, and 2.

For each WS, both during training and testing, the procedure was as follows. The sentence was displayed in black letters on a white background at the top of the window. After the subject indicated they had finished reading by pressing the SPACE key, a question, whose answer depended on correctly resolving an ambiguous pronoun, was displayed beneath. At the same time, two choices were displayed side-by-side. The left choice was always labeled 1 and the choice on the right was labeled 2. The position of the correct choice (left or right) was selected randomly with equal probability.

The subject then pressed 1 or 2 to indicate their answer. The subject's response time was measured (by the computer) as the time between pressing the SPACE key and making a selection by pressing 1 or 2. Figure 1 shows an example of the window contents while waiting for the subject to respond.

Immediate feedback (correct or incorrect) was given after each trial, along with an updated score. The number of correct answers and the number of trials completed so far were shown as a fraction. The subject then pressed ENTER, whereupon the window was cleared and the next trial started. Instructions, including, when appropriate, a warning not to sacrifice accuracy for speed, were visible at the bottom of the window at all times.

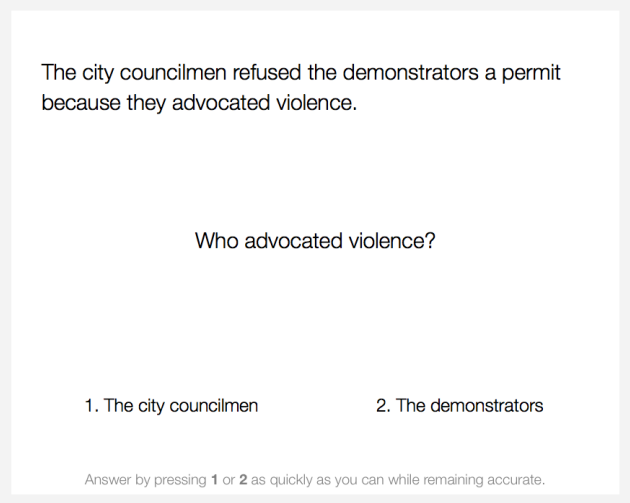

Figure 1: Screenshot of experiment window

After the final question in the test task, subjects were asked to select their age from a pull-down selection widget with allowed values in the range 18-129. Subjects were also given the opportunity to make comments about the experiment. Specifically they were asked if there were any questions that they found confusing or nonintuitive.

Results and Discussion

The 407 subjects who completed testing scored a mean of 92.1% on the normal test questions ($\sigma_s = 0.07$), taking an average of 10.2 minutes ($\sigma_s = 3.9$) to complete the task. Figure 2 shows the distribution of scores. A little more than one out of every seven subjects (58) answered all 40 questions correctly, and only a few had scores close to chance level.

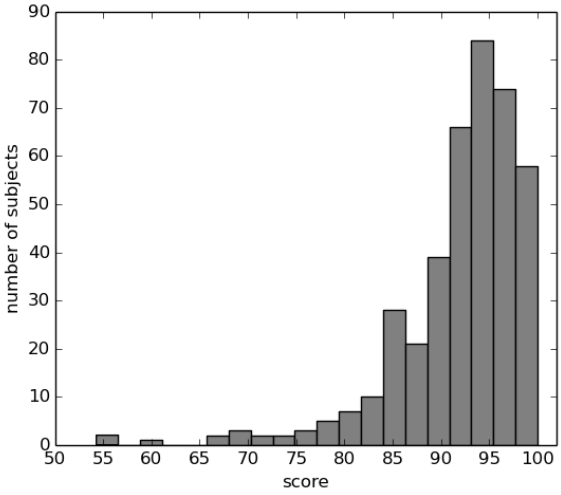

Figure 2: Distribution of scores

Subjects scored a mean of 98.6% on the *Easy WS* questions ($\sigma_s = 0.08$). The difference in scores on *Hard WS* questions and scores on *Easy WS* questions was shown to

be significant in a two sample t-test ($t(808) = -12.33$, $p < 0.001$), so we can reject the hypothesis that subjects perform equally well on both. This provides evidence that our qualification task was, as expected, easier than the testing task. The remaining discussion deals only with the 143 *Hard WS* questions (286 including both versions of each question) taken from the online corpus.

Response times

Mean response time across all subjects was 4.3 seconds ($\sigma_s = 5.2$). Figure 3 shows the response time distribution.

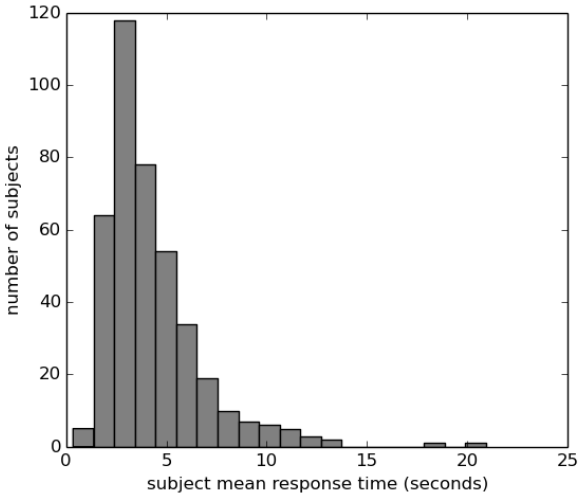

Figure 3: Distribution of response times

One subject had a mean response time of about 350 milliseconds with individual responses well below 100 milliseconds. Two others had mean response times over 17 seconds. In both Figure 3 and Figure 4 these three outliers can be clearly seen.

It is instructive to observe the contribution these subjects make to the overall mean response time (and accuracy). Removing these subjects has very little effect, mainly due to the large number of subjects who participated in the experiment. Mean accuracy rises by only about a tenth of one percent, and mean response time falls by about 70 milliseconds.

Moreover, the subject with mean response time of about 18 seconds scored a perfect 100%. It would be a mistake to throw out this subject's data, assuming, based solely on average response time, that they did not understand the task very well, or didn't pay attention, or don't have a mastery of English. There were outliers in some subject's response times, including delays of up to several minutes. However, there is no way to know with certainty what a subject did during those delays. Data like this may simply represent a relatively rare but real tendency toward very careful deliberation.

The data in Figure 3 prompts an interesting question. Is response time correlated with accuracy? If a subject's response time is taken as a proxy measure for how difficult they find a question, and on average someone is likely to miss more difficult questions than easy ones, will increased average response time indicate a lower score?

Figure 4 shows the relationship between average response time and score. Although there is a significant slightly negative correlation (Pearson correlation, $r = -0.16$), this linear fit explains very little of the score variance ($R^2 = 0.026$). Therefore response time is an imprecise predictor of accuracy.

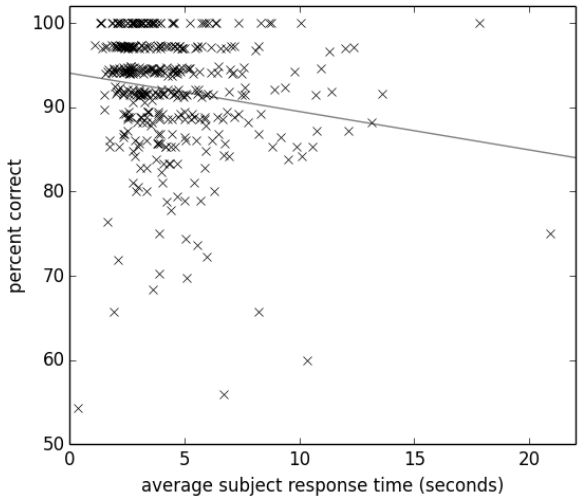

Figure 4: Relationship between mean response time and score

This result suggests that, contrary to intuition, WS questions need not be so simple that they can be solved by an average person extremely quickly, with little or no obvious conscious effort. Complex or difficult WS may be included without too much concern that people won't be able to solve them, and such questions might become desirable if the organizers of the WSC want to increase its difficulty in the future.

Ages

Every subject who completed the experiment submitted their age. Figure 5 shows the distribution of subject ages. The median age was 30, roughly in line with previous reports on MT demographics (Buhrmester, Kwang, and Gosling 2011), suggesting that our pool of subjects was more diverse than it would have been had we performed a laboratory study using only undergraduate students as subjects.

Comments

Given the opportunity, 295 (of 407) participants submitted comments about the experiment. In the course of examining these comments, we identified several qualitative trends worth bringing to light.

Personal preferences and overriding expectations More than anticipated, subjects' answers were guided by what they perceived as the relevant background probabilities for a given question.

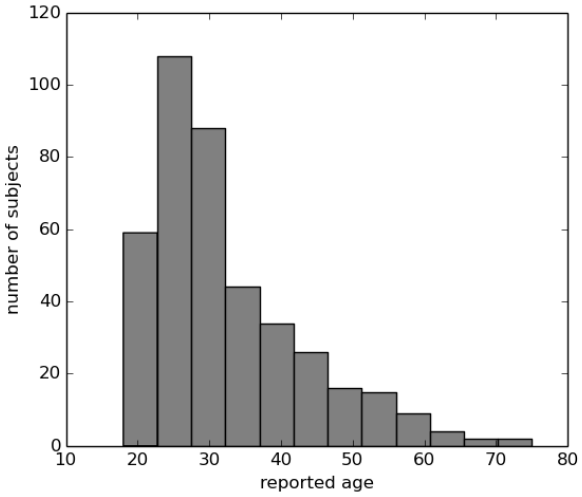

Figure 5: Distribution of reported ages

For example, one question described a situation wherein either a train was delayed or a meeting was short.

> My meeting started at 4:00 and I needed to catch the train at 4:30, so there wasn't much time. Luckily, it was delayed, so it worked out. What was delayed?

Contextual clues, specifically the phrase "so it worked out", direct the reader to the correct referent: *the train*. However, several subjects ignored those clues. Rather, they said, they made their choice based on evidence from their own experience, namely that meetings are very much more likely to be delayed than trains.

Another question described a situation where a group of people showed a preference for one type of cookie over another.

> Everyone really loved the oatmeal cookies; only a few people liked the chocolate chip cookies. Next time, we should make more of them. Which cookie should we make more of, next time?

In this version of the question the correct referent was the oatmeal cookies. However, one subject chose the wrong answer because they couldn't imagine anyone liking oatmeal cookies more than chocolate chip. (According to the subject, only one kind of cookie is more odious than oatmeal: raisin cookies.)

Differences in value judgements WS questions rely on the knowledge people accrete through many years of having experiences in the real world, and such knowledge is never completely free from bias. Even so, several comments revealed a case where differing values can make a question ambiguous. One question involved a very awkward parental situation.

> Pam's parents came home and found her having sex with her boyfriend, Paul. They were embarrassed about it. Who were embarrassed?

Understanding this question depends on one having a sense of who would be most likely to be embarrassed in this situation, the young woman and her boyfriend or her parents. About 10% of the subjects who saw this question, with ages ranging from 19 to 38, thought it was just as likely that the parents would feel embarrassed as the daughter.

Interestingly, there were no concerns expressed about ambiguity in the second variant of this question, which depended on determining who in the situation was the most likely to be *angry*. (96% answered correctly that the parents were most likely to be angry.) Examples like this reveal the subtle effects of value judgements on answers to WS questions.

Unfamiliar concepts Three subjects mentioned that they were unfamiliar with words or concepts used in a particular WS. One subject was unfamiliar with the term *crop duster*. Others didn't recall what a *bassinet* is or were unfamiliar with the name *Xenophanes*. (Note that, in the last case, answering correctly only required recognizing that *Xenophanes* was the name of a person.) Such individual differences in vocabulary are unsurprising, and should be taken into account when assembling a WSC corpus.

Additionally, in future attempts to measure human performance on the WSC, subjects' familiarity with words in the corpus should be measured independently. This way performance on WS containing problematic words, such as those just mentioned, can be correlated with how familiar those words are to subjects, in general.

Rushing and unintentional errors Despite repeated warnings that they read carefully and not rush, many subjects mentioned their tendency to "move too fast" or that they were "rushing" or "reading too quickly" or "trying to go fast". About 6% of the comments contained this sentiment in one form or another. This may be an inevitable consequence of the fact that subjects were working for pay, with no direct supervision.

Additionally, the same number of participants commented that they had pressed one button when they meant to press another. For example, they pressed 1 when they intended to select the choice associated with 2. This type of error is unavoidable without adding a step where subjects confirm their choice.

Fortunately neither of these issues resulted in a large number of subjects losing interest in the study or completing it with intentionally poor results. On the contrary, over 10% of the subjects who left comments mentioned that they enjoyed the task.

Corpus Analysis

Constructing a high quality WSC corpus is a laborious job. To date, most WS questions have been carefully crafted by a small handful of people[3]. Many examples that at first glance seem acceptable, upon closer scrutiny exhibit one of the flaws discussed in the first section (selectional restrictions, statistical correlations, or one-way ambiguity.) Here

[3]One exception is notable. The authors of a recent model of pronoun resolution created a large corpus using a group of undergraduates (Rahman and Ng 2012), but many of the results were *Easy WS* questions.

we discuss only general corpus quality and one-way ambiguity. Evaluating susceptibility to statistical correlations is beyond the scope of this work.

Given that such care has been taken in creating Levesque, Davis, and Morgenstern's online corpus, one would expect that most of them would be answered correctly by almost all subjects. This was shown to be mainly true, though there were some interesting exceptions. Figure 6 shows all 286 questions, ranked according to what percentage of subjects who saw the question answered it correctly.

As is shown in Figure 6, over 75% ($286 - 69$ out of 286) of the questions were answered correctly by at least 90% of subjects who were given them. Almost 17% were never answered incorrectly. On the other hand, 26 questions were answered incorrectly by 20% of those who saw them, and for one question, our subjects' accuracy was less than chance.

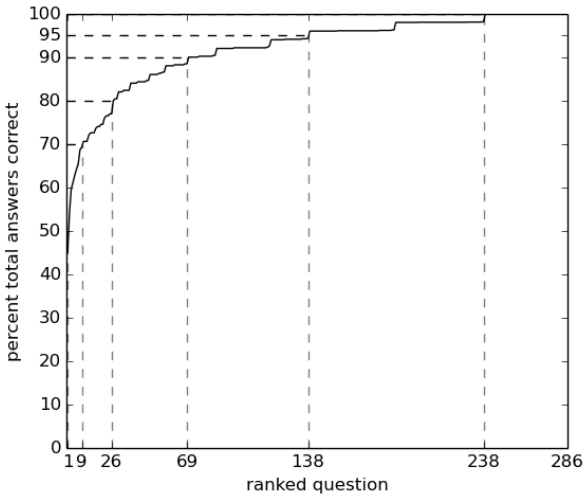

Figure 6: Ranked overall question score

This question, the first version of number 40, is worth looking at more closely. (Hereafter the first variant of a WS question will be called *version A*, and the second variant will be called *version B*.)

> I couldn't put the pot on the shelf because it was too tall. What was too tall?

Most participants selected "the shelf" as their answer for this question. This is incorrect, because although shelves can be high or low, their height above the floor is not normally thought of as an object having extent, as would usually be implied by the adjective "tall". Rather, the shelf itself would have to be stretched upward in order to be considered tall – a very unlikely picture.

And yet, looking only at version A of the question, subjects may not have had reason to focus on that particular word. The full WS is shown below, and "tall" is part of its fulcrum.

> I couldn't put the pot on the shelf because it was too tall. What was too tall?
>
> I couldn't put the pot on the shelf because it was too high. What was too high?

Without specifically asking about this question in a post-task questionnaire or interview, we can only speculate as to why so many subjects missed this question. However, one plausible interpretation is that it is only when we look at both versions together that the correct answer to the version A, *the pot*, is evident. This likely happens because when we see both versions together, the words that make up the fulcrum (*tall/high*) are the only ones that differ between the sentences. This inescapably draws our attention to them, and in full context it becomes easy to see "tall" in contradistinction to "high".

Subjects almost always answered version B of this question correctly (98%), and many of those that missed version A commented on how confusing they found it. Unsurprisingly, this shows that question number 40 suffers from one-way ambiguity, though it is difficult to notice when looking at version A in isolation.

Many other questions exhibit this same imbalance. Figure 7 shows the fifteen WS with the largest difference in accuracy between versions A and B.

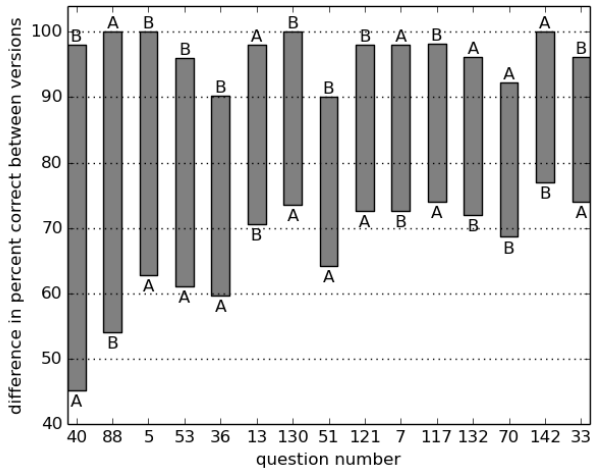

Figure 7: 15 WS questions with largest differences between versions A and B

For each bar the in the Figure, the lower boundary sits at the accuracy score associated with the more ambiguous version, whereas the bar's upper boundary lies at the accuracy level of the less ambiguous version. As an example, for WS question number 36, shown below, subjects only answered version A correctly about 60% of the time, but answered version B correctly 90% of the time.

> In the middle of the outdoor concert, the rain started falling, and it continued until 10. What continued until 10?
>
> In the middle of the outdoor concert, the rain started falling, but it continued until 10. What continued until 10?

In version A, the conjunction *and* indicates that the rain starting to fall should be taken jointly with it continuing to fall. However, subjects may have seen the concert as more salient than the rain, or may have been drawn to an image of both performers and audience soldiering on in the face of bad weather. Whatever the reason, these and similarly one-

way ambiguous questions should be modified or discarded from a high-quality WSC corpus.

Looking back at Figure 6, an interesting question arises. What would our subjects' mean accuracy be if we omit the 26 questions with lower than 80% accuracy? In this case, overall accuracy would rise to 94.4%, and if we omit questions with lower than 90% accuracy, overall accuracy rises still further to 96.2%. Of course, *post hoc* adjustments such as these should be regarded with suspicion, but they they are still intriguing.

In summary, there is a trade-off made when tightly controlling a corpus's contents, as have Levesque, Davis, and Morgenstern. Such effort can help reduce the occurrence of flaws, but it cannot eliminate them. In particular, several dozen of the WS in the currently available online corpus suffer from one-way ambiguity. This may be in part due to the fact that WS questions are almost always viewed as pairs.

It is recommended that any proposed WS corpus be tested against a broad range of human subjects to root out any examples of one-way ambiguity. In addition, other methods should be developed to identify WS that have issues with statistical correlations or selectional restrictions.

Conclusion

There are three main conclusions can be made as a result of this study.

First, participants performed at 92% on the WSC corpus currently available online. There is little to indicate this level of performance is not representative of the general population of English speaking adults who live in the US, so it should serve as a reasonable baseline for human performance on the WSC.

Second, although response times significantly influenced accuracy, there was great variation. The mere fact that a WS requires significant mental effort should not immediately disqualify it.

Finally, even a carefully curated corpus of WS was found to contain questions with one-way ambiguity. It can be difficult to identify points of conflict in value judgements and other factors that influence the understanding of a WS. Because of this, validating a WSC corpus by testing it against human subjects is critical to ensuring high quality.

Acknowledgements

We are grateful to Dr Robert Goldstone of the Percepts and Concepts Laboratory at Indiana University for his guidance and support.

References

Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6:3–5.

Chandler, J.; Mueller, P.; and Paolacci, G. 2014. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behavior research methods* 46:112–30.

Crump, M. J. C.; McDonnell, J. V.; and Gureckis, T. M. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE* 8(3).

Dagan, I.; Glickman, O.; and Magnini, B. 2006. The PASCAL Recognising Textual Entailment Challenge. In *ML Challenges: Evaluating Predictive Uncertainty, Vis. Obj. Clas., and RTE, 1st Pascal ML Challenges Workshop, April, 2005*, volume 3944, 177–190.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2012. The winograd schema challenge. In *AAAI 13th International Conference on the Principles of Knowledge Representation and Reasoning*, 552–561.

Levesque, H. 2011. The Winograd Schema Challenge. *AAAI Spring Symposium Series*.

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

McDonnell, J.; Martin, J.; Markant, D.; Coenen, A.; Rich, A.; and Gureckis, T. 2012. psiTurk (Version 1.02)[Software]. New York, NY: New York University.

Rahman, A., and Ng, V. 2012. Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. *Proc. of the 2012 Joint Conf. on Empirical Methods in NLP and Comp. Natural Language Learning* 14(12):777–789.

Roemmele, M.; Bejan, C.; and Gordon, A. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. *AAAI Spring Symposium*.

Winograd, T. 1972. *Understanding Natural Language*. New York, NY: Academic Press.