Human tests of materials for the Winograd Schema Challenge 2016

Ernest Davis
Department of Computer Science
New York University
davise@cs.nyu.edu

Leora Morgenstern
Leidos Corp.
leora.morgenstern@leidos.com

Charles Ortiz
Nuance Corporation
charles.ortiz@nuance.com

July 2, 2016

Abstract

A corpus of 108 pronoun disambiguation problems and 89 Winograd schemas was assembled. These materials were evaluated by 21 human participants. Overall, 91% of the answers given by the participants agreed with the answers intended by the test designers. For the running of the Winograd Schema Challenge at IJCAI-2016, it has been possible to extract collections of 60 PDPs and 60 WS-halves whose natural interpretations are very reliably endorsed by the participants in the experiment.

We report here on an experiment to determine human performance on collections of pronoun disambiguation problems (PDPs) and Winograd Schemas (WSs), in preparation for running the Winograd Schema Challenge at IJCAI-2016 (Morgenstern, Davis, and Ortiz, 2016) (Levesque, Davis, and Morgenstern, 2012).

A similar experiment is reported in David Bender (2015). Bender used a population of 403 Mechanical Turkers, each answering 40 questions; his subjects achieved an overall accuracy of 92.1%.

1 Material preparation

A collection of 66 texts comprising 108 PDPs was collected from a number of literary sources. (A single text with multiple pronouns could provide a separate PDP for each pronoun.) These were edited in a number of respects: Content and wording was changed to make the texts clearer to human readers (perhaps also to automated readers) and to clarify the context of text which were now isolated. Genders of pronouns and/or possible referents were altered in order to increase the number of possible referents.

Of the 108 PDPs,
74 had 2 possible referents,
27 had 3 possible referents,

4 had 4 possible referents,
3 had 5 possible referents.
Therefore, guessing at random, the expected score would be 47.6 correct out of 108 = 44%.

All proper names were changed to avoid the possibility that a human reader might recognize the source of the quotations. There is no reason to suppose that any reader did recognize any of the sources, though one reader made a correct guess as to the general era and milieu, and one made an incorrect guess about the author. In one case, the name of a prominent personality was changed; two readers spotted the change and correctly guessed the original.

A collection of 89 Winograd schemas were manually constructed by the authors. Two sets of 89 disambiguation problems, denoted here "WS.a" and "WS.b" were extracted from this collection, each schema supplying one of its halves to WS.a and the other to WS.b. The assignment of halves to sets was done at random.

2 Participants

Participants were recruited by announcements in the NYU Computer Science email chat group and in the department job-listing email list. The announcement required that participants be fluent speakers of English, though they were not required to be native speakers, and that they not be experts in linguistics or related fields.

Twenty-one subjects in total were recruited. Of these, seven were native English speakers, and fourteen were not. Seventeen, including all the non-native speakers, were currently students in the Computer Science Department; three were alumni of the program (one from many years earlier) and one was a high-school student. No other information about the participants was recorded.

Participants were paid for their participation, though the three alumni all declined payment.

3 Presentation of the material

The participants were given the test materials in printed form, and marked their answers in pen on the tests. The instructions they were given can be found at
http://www.cs.nyu.edu/faculty/davise/papers/WS2016Instructions.pdf

Test questions were given in the following form: A text was presented, with one pronoun in boldface. This was followed by a snippet, with a phrase containing the pronoun, to further emphasize which pronoun was the subject of the question. Next, a list of possible referents for the pronoun was given. For example:

> The trophy did not fit in the brown suitcase because **it** was too small.
> **it** was too small.
> A. trophy
> B. suitcase.

The order of answers was always the order of the first occurrence of the referent in the text.

In the PDPs, as mentioned, there were often multiple PDPs for a single text. These were placed consecutively, in the order of the pronouns being queried, and labelled "**SAME TEXT AS PREVIOUS QUESTION**," on the test sheet to avoid confusion. In the WSs, there were

some cases where the text of consecutive problems were somewhat similar; these were labelled "**TEXT IS DIFFERENT FROM PREVIOUS QUESTION**."

Participants were instructed to write the letter corresponding to the most plausible referent next to the question, or to write "X" if they thought the question was genuinely ambiguous. It may be noted that the designers of the tests did not intend any texts to be genuinely ambiguous, and that the automated contestants in the challenge will not be encouraged to mark a question as ambiguous, and will not gain any advantage from doing so. However, part of the point of this test was to gather information about which of the examples are poorly constructed; and from that point of view, it is more useful to us to get a clear statement from the participants that the example is ambiguous than to have them guess randomly, as the automated contestants will presumably do. Of course, since the test designers wanted their examples to pass the test, they preferred not to see the answer of X, and it is certainly possible that participants guessed at that preference and, consciously or unconsciously, tried to accomodate it.

Participants were permitted to ask the test-taker about the meaning of unfamiliar words. In the event, three subjects took advantage of this; two asked about one word and one asked about two (four different words). Participants did not raise any other questions. No other feedback from the participants was obtained, beyond their answers.

There were three tests a participant might take: the PDPs, WS.a, and WS.b. Giving a single participant both WS.a and WS.b could be problematic, in that seeing one half of a schema might influence his/her interpretation of the alternate half. (The automated contestants in round 2 the Winograd Schema challenge will see a collection consisting of only one half of each schema.) Therefore a participant might either take one test, or might take the PDPs plus one version of the WS.

Participants were informed that they were being hired for about an hour's work. At the beginning of the hour, they were handed either the PDPs or one of the WSs. If they completed that test in half an hour, or a little more, then they were invited to work through a second test. In the event, 16 subjects took both the PDP and a WS; 3 took the PDPs but did not take a WS; and 2 took a WS but not the PDP. Somewhat curiously, of the 5 who completed only one of the two tests, 3 were native speakers. In total 19 subjects took the PDPs, 9 took WS.a and 9 took WS.b.

Subjects were not individually timed. All but one of the subjects took between 45 and 75 minutes to complete their work; one subject, however, took only 25 minutes to complete two tests. Note that subjects who completed two tests answered a total of 197 questions. Thus, a subject who completed in an hour took an average of 18.3 seconds per question; the one speedy subject took an average of 7.6 seconds. In Bender's (2015) experiments, his subjects took an average of 15 seconds per question.

4 Results

In all that follows, we will, for convenience, use the phrase "the correct answer" and similar phrases to mean "the answer intended by the test designer."

Subjects numbers 2, 4, 5, 6, 11, 17 and 19 were the native English speakers; the remaining are the non-native speakers.

4.1 Results for the PDPs

The PDP test consisted of 108 problems with 19 subjects: 7 native speakers and 12 non-native speakers. The overall average score was 90.89% with a standard deviation of 7.6%.

Individual scores:

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
0.80	0.99	0.76	0.96	0.93	0.96	0.97	0.93	0.95	1.00

S11	S12	S13	S14	S16	S17	S18	S19	S20
0.97	0.77	0.88	0.87	0.83	0.95	0.97	0.91	0.85

The mean over the native speakers was 0.95; the mean on the non-native speakers was 0.88. Given the small sample size, this is probably not statistically significant, but it is suggestive. In particular subjects 1, 3, and 12 had substantially lower scores, though not anomalously so. If these subjects are excluded, then the average on the remaining subjects overall rises to 0.93 and the average on the non-native speakers rises to 0.92; i.e the difference between the native and non-native speaker vanishes.

We now consider the individual problems in terms of the number of subjects who got them right.

Subjects who answered correctly:	19	18	17	16	15	14	13	12
Number of problems:	42	23	7	13	11	6	5	1

Thus all the problems were answered correctly by at least $12/19 = 63\%$ of the subjects.

If the three weakest subjects are excluded, then there are 54 questions that were answered correctly by all 16 of the remaining subjects, and 22 questions that were answered correctly by 15 out of 16.

4.2 Results for WS.a

The WS.a test consisted of 89 problems with 9 subjects: 1 native speaker and 8 non-native speakers.

The overall mean was 92% with a standard deviation of 6%.

Scores of individual subjects:

S1	S6	S9	S10	S13	S16	S18	S20	S21
0.87	1.00	0.98	0.98	0.92	0.86	0.97	0.84	0.88

Subject 1, who had a notably weak score on the PDPs, had a somewhat low score here, but not strikingly so.

Individual problems:

Subjects who answered correctly:	9	8	7	6	5	4
Number of problems:	50	22	12	4	0	1

Unlike the PDPs, there is one very badly designed problem here, for which fewer than half the test-takers came up with the answer intended by the test-designer. (It is probably not reasonable to say that they got the "wrong" answer.)

4.3 Results for WS.b

The WS.b test consisted of 89 problems with 9 subjects: 3 native speakers and 6 non-native speakers.

The overall mean was 93% with a standard deviation of 3%.

Scores of individual subjects:

S2	S3	S7	S8	S12	S14	S15	S17	S19
0.97	0.92	0.94	0.93	0.91	0.87	0.94	0.93	0.94

Subject 3 and 12, who had notably weak scores on the PDPs, had scores at the mean for this test.

Individual problems:

Subjects who answered correctly:	9	8	7	6	5	4	3
Number of problems:	63	11	6	5	2	1	1

Here there were two very badly designed problems, on which fewer than half the test-takers agreed with the test-designer, and two more in which only five of the nine subjects agreed with the test-designer.

4.4 Two-sided Winograd Schemas

In a valid Winograd Schema, both halves of the schema must be validated with user tests, since the choice of half used should be made at random. The following tables shows the number of schemas that were highly successful on both sides.

Responses on two halves	9 & 9	9 & 8	8 & 8
Number of problems:	38	21	3

5 Extraction of problem sets for 2016 Winograd Schema

For the running of the Winograd Schema Challenge at IJCAI 2016, our plan is to use a set of 60 PDPs for the first round; if the challenge proceeds to a second round, to use a set of 60 WS

halves for the second round.

We did not simply take the PDPs that scored the highest on the subject tests because there were other considerations involved. For instance, the collection for the subject tests included some texts with as many as 5 pronouns queried; in the challenge collection there are no more than 3 queries about any single text. The collection that we will use comprises:

33	problems for which	19	subjects gave the same answer
19	problems for which	18	subjects gave the same answer
6	problems for which	17	subjects gave the same answer
2	problems for which	16	subjects gave the same answer

Note that 16 out of 19 is still 84%, a high level of agreement. Overall, 96.75% of subject answers were "right" over this corpus; this is not, of course, a statistically sound measure.

The collection of 60 Winograd schemas to be used in round 2 of the competition comprise:

38 halves of schemas where both halves were solved correctly by 9 people.

21 halves of schemas where one half was solved correctly by 9 and the other half by 8

1 half of 1 schema where both halves were solved correctly by 8.

6 Publication of test materials

Once a round of a Winograd Schema challenge has been run, the problems used in that challenge (and their alternates, in the case of a Winograd Schema) will be published.

Regrettably, we cannot publish any of the other problems, because we may want to use them in subsequent runnings of the challenge. Even problems that scored very badly in subject tests cannot be published, because we may want to try to modify them.

7 Conclusions

This kind of data obviously cannot support any general conclusions about human abilities to do pronoun reference disambiguation. The subject pool is small. The subjects are neither representative of any particular population nor are they homogeneous in their English language abilities. They are presumably bright, well-educated, and accomplished test-takers; on the other hand, two-thirds are not native speakers. The problems are not representative of pronoun disambiguation generally; they were deliberately selected or contrived to emphasize certain cognitive abilities. The test is not ecologically sound as a test of reading comprehension; in ordinary reading, there is almost always more context, there are more distractions, and it may not be possible to understand the text without finding the referent of every pronoun.

However, as far as it goes, the data collected in these tests supports the statement that the bar for "near human" abilities on the Winograd Schema Challenge to be given at IJCAI 2016 can conservatively be set at 90% correct. This result corresponds with the result found by Bender (2015) of 92.

Acknowledgements

Thanks to Hector Levesque and Gary Marcus for discussions. Thanks to the Nuance Corporation for providing funding for the tests.

References

D. Bender, "Establishing a Human Baseline for the Winograd Schema Challenge," MAICS-2015, 39-45.

H. Levesque, E. Davis, and L. Morgenstern, "The Winograd Schema Challenge," KR-2012.

L. Morgenstern, E. Davis, and C. Ortize, "Planning, Executing, and Evaluating the Winograd Schema Challenge," *AI Magazine,* Spring 2016.