Running head: PI GROUP: REFERRING EXPRESSIONS

Principal Investigator Group:

Referring Expressions

Jason Chong
Ward McLean
Anton Venema

Simon Fraser University

LING 481, Nancy Hedberg, 29th of November 2007
Clark & Wilkes-Gibbs, in their paper, "Referring as a Collaborative Process" (1990), set up an experiment in which they sought to investigate the ways by which people negotiate the formation of referring expressions in conversation. They elaborate the theory that conversation is a primary locus of language use, and that it is a necessary mechanism of any linguistic exchange for parties to agree upon references presented within the realm of discourse.

Conversation inherently involves a large degree of collaboration between two or more people. In most cases, there is a back-and-forth turn-based alternation during which each participant in the conversation makes what they believe to be a valid contribution. According to Grice (1975), the contributions must be co-operative according to a set of guidelines, or maxims. At the very least, each offering should be relevant and succinct. Each sentential element in the discourse is provided in such a way as to build on the previous foundation of shared information and to enhance it in some way.

In many cases, it seems that conversational participants will collaborate in order to ensure that a common understanding is maintained. This is, perhaps, due to the characteristic temporal limitations of spoken word which inhibit a significant amount of forethought and planning. A frequently occurring example of this can be found in referring expressions. When referring to an object in discourse, it is typically necessary for both the speaker and hearer to establish the same reference – that of the speaker. If for some reason the referential object is not clear to the hearer, the conversation can rapidly break down, as future contributions are based to a large degree on past understanding.

For this reason, the hearer will often provide the speaker with some form of feedback to let them know that the referring expression has been understood and that a shared reference has been established. Until that point has been reached, the speaker and hearer will typically engage in a collaborative process by which they communicate back and forth. Once both parties are satisfied that the referent is fully known, the dialogue can continue on its previous course. There are many examples of this, and so its existence is difficult to dispute, but understanding the nature of how this collaboration takes place requires study.

Accordingly, several different replications of the 1990 experiment were enacted, transcribed, and analyzed. The goal of this paper is to gather the data together from the various replications, analyze the overall results, and produce a critical review of the experiment – its strengths and shortcomings.
Method

Materials

The twelve figures from the Chinese game of Tangram, as displayed in the report of the original experiment, were copied and distributed to each group. Most groups pasted the figures upon index cards for durability.

Subjects

In the original experiment, subjects consisted of students fulfilling undergraduate requirements at Stanford University. Our seven replications involved seven different pairs, typically selected due to their familiarity with the experimenter. Married couples and friends of a range of age groups were selected, as opposed to the random stranger pairings in the original experiment.

Procedure

The different groups conducting the experiment performed the experiment in varying manners. While the general approach was similar, the finer procedural details held some critical differences. In the original experiment, the subjects were separated by an opaque screen, which we believe was designed to restrict the subjects to verbal communication. It is readily imaginable that visual cues could quite easily confound the experiment. Attempts were made by most groups to replicate this experimental feature, but unfortunately not all. The original paper made no mention of randomizing the Tangram figures after each trial. It was implied in at least one location, but was not explicitly stated as one of the procedural requirements. It was quite likely this oversight in the original paper that caused at least one of our groups to use the same ordering for all their trials. The last major departure from the original experiment that we observed was the alternation of director/matcher roles between trials. Only one group made this a part of their procedure, likely as an attempt to provide counterbalancing. It was, however, not a part of the original experiment, and therefore a potential confounding factor.

The matching procedure was conducted properly (i.e. in line with the original procedure) in all groups. Six sequential trials were conducted, each of which involved the director verbally directing the matcher to arrange his or her set of 12 figures such that the matcher’s arrangement matched the director’s. Subjects were informed that their time and accuracy would be measured. For most replications, the entire experiment ran for about 20 to 30 minutes.

All of the trials were recorded and subsequently transcribed, adhering as closely as possible to the format outlined in Santa Barbara research. From these transcriptions, the experimenters could analyze the discourse to determine the development of referring expressions, measured by counting the number of words and turns per figure used by the director as well as the type of the initial noun phrase used by the director to initiate the referring process.
According to the original experiment, each figure was to be assigned at least one turn. One group assigned less than one turn to certain figures if the director described more than one figure without a verbal response from the matcher. Another group enumerated all noun phrases used by the director, rather than just the initial one as detailed in the original paper.

Results

The first thing that is valuable to mention is the variations in experimental design and procedure. Due to the large number of independent investigations, there were an equal number of interpretations of the original paper. The first problem we considered was that of visual cues. Fortunately for us, the groups that did not make use of a fully opaque screen did not exhibit significantly different results from the other groups. This is potentially due to the fact that they used subjects who were not intimately familiar with each other. It is possible that strangers are less likely to build on visual cues, since they typically involve a higher degree of understanding. Additionally, we have to question whether or not visual cues would even present a confounding factor if present. Unless the director physically acts out his or her interpretation of what a given figure or feature looks like, it is difficult to imagine how anything other than verbal communication could be used to reliably communicate the correct reference. As the results obtained by this group were overall in-line with the others, we will leave these questions for future research.

For the group which did not randomize the figure arrangement after each trial, the potential for confounding the data is prominent. The most likely manner by which this would occur would be if the director was confident enough in the matcher’s recognition during the last trial to refer to the figure by position number rather than by characteristics of the figure itself (“the one that was first last time”), or even skip over the figure altogether (“you know number two, so number three...”). Fortunately for us, the subjects did not exhibit this behavior. They used characteristic referring expressions across all trials, and so the data produced maintained its integrity and usefulness in this composite view.

The group that allowed the alternation of matcher/director roles presented the most significant variation in results. It was immediately clear after the first role reversal that something unique was happening by which the former matcher was now taking a dominant director role by attempting to establish new foundational reference points for each figure. Several of these references were significantly different from the initial ones presented by the former director. Interestingly, after the “rehashing” that took place during the second trial, both parties seemed to agree on the foundation, and the remainder of the data appeared to be remarkably similar to the other groups. Most notable was that including their data did not modify the results significantly, and so they were not excluded from consideration.
There were also notable variations in the transcription analysis to consider. For the group who assigned less than one director turn to figures, their transcripts were re-analyzed. The revised data produced by this second analysis was used in place of their original data when producing the composite chart. The other group who defected from the expected analysis counted all noun phrases instead of just the initial ones. Unfortunately, their transcription was not made available to us, and so it was not possible to go back and re-analyze. Because of this and other complications with their table generation, this group’s report was excluded from consideration for initial noun phrase types. Thus, Table 1 was composed of data from only six of seven groups.

While the description of how to classify the initial noun phrases was well-documented in the original paper, each group categorized them according to their own idiosyncratic assessment of the classifications by Clarke & Wilkes-Gibbs. It was determined by our group members after transcription analysis that certain noun phrases were confused or slotted differently depending upon the consensus of the experimenters. Two groups did not incorporate the “Descriptions” category at all. This is potentially due to either its relatively obscure reference in the original paper or to considering it to not be necessary. While we considered revisiting these groups’ transcriptions and re-evaluating their initial noun phrase classifications, we agreed that in the interests of time, we would produce results based entirely on the original assessments.

As indicated in the two graphs and in the table (with an additional graph of the table), the composite results were very similar results to the original experiment. We were particularly delighted with the presence of the slight upward curve in Figure 2, which charted the average number of turns per figure used by the director. This same feature was present in the original experiment’s results. Apparently, the greatest efficiency in referential behavior is obtained in the fifth trial for the majority of cases. The exact reason for the increase in turns during the sixth trial is unknown. Overall, the charts corresponded closely to the originals. Although the specific values differed slightly, the trend was remarkably similar. In the original, nearly 4 turns per figure were needed for the first trial. We were higher, at 5.5 turns per figure. Thereafter, the drop was identical to only 2 turns per figure in the second trial, in spite of the one group who found an increase due to a director/matcher alternation. By the third trial, only 1.5 turns were needed per figure. Fewer turns still were required for each figure in trials 4 and 5, again matching the original results of Clark & Wilkes-Gibbs.

As to the number of words used by the director per figure, displayed in Figure 1, our data matched the pattern evidenced by the original results. While their directors required an average of 42 words per figure in the first trial, ours required 60 words, perhaps reflecting the participants’ greater familiarity with each other and more relaxed conversational tendencies. Thereafter, there was stronger conformity to the original experiment. We had 23 words in trial 2, while Clark found 19. For trial 3, we found 15 words, while Clark had 12. We were identical in trial 4 with 10 words, and nearly identical for trials 5 and 6. In fact, our directors and matchers became more efficient,
requiring a word or two fewer for the last two trials. Again, this may be an indicator of the comfort level of subjects who know each other.

Clark & Wilkes-Gibbs separated attempts to describe a figure (i.e. “The next figure looks like...”) from direct identification (i.e. “The next one is the skater.”). As expected, more descriptions were found in the earlier trials than in the later ones for both sets of data. Our results showed a drop from trial 2 to 6 of 19% down to only 8%. This was less substantial in the original, where it was only 17% down to 14%.

It follows that we found the same pattern Clark & Wilkes-Gibbs did with regards to the increased use of elementary noun phrases in later trials. Starting with the second trial, elementary phrases were increasingly employed, rising from 51% to 79%. The original advanced from 52% to 72%, peaking at 80% in the fifth trial. Our data peaked at 82%, also in the fifth.

We found a greater use of proxy phrases, whereby one participant finished another’s phrase, than what was observed in the original experiment. This could easily reflect the familiarity our subjects had with one another, since finishing someone else’s thoughts not only indicates a deeper-than-average level of understanding, but could be construed as rude interruptions when occurring between strangers. Our participants used a greater number of installment noun phrases yet fewer provisional phrases than the original. Provisional phrases are indicated by a quick effort by the director to explain or clarify his expression in a new clause, whereas an installment phrase involves a pause or try marker before the director initiates the next clause. Not only are they similar, but the presence of episodic noun phrases (episodic tonal groups instead of separate clauses) makes them all very difficult to distinguish. This could perhaps explain the variation between our results and those of the original experiment.

Discussion

The results of the Clark & Wilkes-Gibbs experiment on referring expressions appear to be very robust. In spite of varying noun phrase assessments, our overall patterns strikingly confirm the original findings. It is clear that when faced with the task of getting another person to establish the same common reference point for an object, the individuals involved in the discourse will work collaboratively with each other to achieve a mutually satisfactory reference.

We noted that one of the groups used indirect referring expressions throughout the entire trial. None of the other groups make a significant use of indirect expressions, and none at all were found in the later trials. It would be interesting to examine the conditions under which this particular replication was performed to see if there was some factor that would explain this unusual behavior.
Several of the groups used non-native speakers of English as subjects. While we did not see any clear indication that this played a role in the results we obtained, it is worth noting that the two groups with the highest number of turns in the first trial were non-native English speakers. This potential uncertainty in proper formation of English sentences could have led the subjects to be more interactive in their verbal communication. We did not have access to relevant first/second-language information for all the subjects’ used in the experiments, however, so this conjecture is just that. In any case, the study could be replicated using native and non-native English subject distinctions to see if this has a significant effect on the results.

It would be interesting to see if there is a linguistics scalar involved here, for which certain means of reference are stronger than others on an organized scale, such as:

> Description < Provisional NP < Episodic NP < Elementary NP

Elementary noun phrases would be the strongest form of reference. It should then be possible to take a reference for which an elementary phrase would suffice and use a description in its place. However, it would not be possible to take a reference for which a description is the strongest possible realization and substitute an elementary noun phrase. Determining the exact nature of the scale and where initial references fall on it could warrant further research.

With regard to the increase in the number of turns used by the director in the sixth trial, perhaps the experiment could be replicated with an extended number of trials. While concern would have to be taken as to subject, it would be interesting to see if the behavior is due to some predictable factor.
Figures and Tables

Figure 1: Average number of words per figure used by director

Figure 2: Average number of turns per figure used by director
<table>
<thead>
<tr>
<th>Initial Noun Phrase Types</th>
<th>Trial 2</th>
<th>Trial 3</th>
<th>Trial 4</th>
<th>Trial 5</th>
<th>Trial 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary</td>
<td>51%</td>
<td>54%</td>
<td>78%</td>
<td>82%</td>
<td>79%</td>
</tr>
<tr>
<td>Episodic</td>
<td>6%</td>
<td>19%</td>
<td>7%</td>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>Installment</td>
<td>8%</td>
<td>0%</td>
<td>1%</td>
<td>3%</td>
<td>1%</td>
</tr>
<tr>
<td>Provisional</td>
<td>10%</td>
<td>7%</td>
<td>3%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Dummy</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Proxy</td>
<td>3%</td>
<td>7%</td>
<td>3%</td>
<td>3%</td>
<td>6%</td>
</tr>
<tr>
<td>Description</td>
<td>19%</td>
<td>11%</td>
<td>7%</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>3%</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
</tr>
</tbody>
</table>

*Table 1: Percentage of six types of initial noun phrases for trial 2 through trial 6*
References
