The Design of an Experiment in Anaphora Resolution for Referring Expressions Generation

Diego Jesus de Lucena
University of São Paulo (USP/EACH)
Av. Arlindo Bettio, 1000
São Paulo, Brazil
diego.si@usp.br

Ivandré Paraboni
University of São Paulo (USP/EACH)
Av. Arlindo Bettio, 1000
São Paulo, Brazil
ivandre@usp.br

Abstract

We present a pilot experiment to measure the effects of redundancy in the resolution of definite descriptions as performed by a small number of human readers. Although originally intended to provide evidence of how much redundancy should ideally be included in generated anaphoric descriptions, preliminary findings reveal a number of little explored issues that are relevant to both referring expressions generation and interpretation.

Keywords

Referring expressions generation, Anaphora resolution.

1. Introduction

Human speakers routinely make use of redundant information when referring to world or discourse objects via definite descriptions, and they often do so even when the sole purpose of referring is the identification of the target object. By contrast, Natural Language Generation (NLG) systems usually implement referring expressions generation (REG) algorithms that are far less prepared to include redundant information in their output.

One possible reason for this difference between real language use and NLG output is the fact that generating redundancy without a proper reason comes with a price, namely, false logic implicatures in the sense defined by H. P. Grice [1]. For instance, in a context containing only one object of type 'door', a redundant (from the point of view of identification) reference to the colour attribute of the referent as in "please open the *red* door" may cause the hearer to consider whether there is any special reason for mentioning such 'unnecessary' information at all.

To avoid this sort of mishap, most REG algorithms to date (including one of the most influential works in the field, the Incremental algorithm in [2]) attempt to avoid the inclusion of any information not strictly necessary for the identification of the intended referent. Referring expressions produced in this way are suitably brief, but they may look unnaturally so. In extreme cases, certain instances of short descriptions may actually make the identification of the intended referent a daunting task. One such example is the case of *deictic* references in structurally-complex (e.g., spatial) domains. Deictic referents may be unidentifiable unless a certain amount of redundant information is added [3]. For example, a distinguishing description such as "the girl in white shoes" is not of much help if, say, the referred person is part of a large crowd. Redundancy in this case (e.g., "the girl in white shoes, *next to the elevator*") may facilitate the resolution of these expression (here understood as the task of interpreting the referring expression and identifying the intend referent.)

The implication of this for REG algorithms is that redundancy should be somehow taken into account at least when generating instances of *space deixis*, and this is precisely the kind of insight needed to design NLG systems that describe objects in physical contexts. For other kinds of application, however, this may be only a minor issue. This is the case, for example, of systems that generate textual reports or documents making intensive use of *anaphoric* referring expressions. In these cases, it is far less clear whether the same principle of 'redundancy as a means to help resolution' is applicable, and if not, what role redundancy should play at all.

In this work we investigate the effects of redundant information in anaphora resolution to gather evidence on how to generate more human-like anaphoric descriptions. More specifically, we designed a small pilot experiment to measure reader's search behaviour under a number of controlled situations of anaphora resolution. Preliminary findings suggest that some of the existing evidence on deixis may not hold for anaphora, and reveal a number of little explored issues that are relevant not only to REG, but to research on language interpretation as well.

2. Background

Probably the best-known REG algorithm to date is the Incremental algorithm in [2]. The input to the algorithm is a context set *C* containing a number of objects – a target object and its distractors – with their corresponding semantic properties (represented as attribute-value pairs as in 'colour-blue'), and the intended referent *r* to be described by means of a definite description. The goal of the algorithm is to compute a list of properties *L* such that *L* denotes the intended referent *r* and no other distractor in *C*. Redundancy in this case is merely a by-product of a more general strategy to cope with the computational complexity of the task: once an attribute is selected for inclusion in *L*, it can never be removed (and, crucially, not even if a subsequent addition makes this attribute redundant), which gives the name 'incremental to the approach.

In previous work [3] we describe an experiment to measure the effort involved in the resolution of deictic descriptions in spatial domains, whose results suggest that under certain circumstances the inclusion of logical redundancy may be necessary if the hearer is to identify

International Conference RANLP 2009 - Borovets, Bulgaria, pages 225–229

the intending referents at all. The findings in [3] however do not cover anaphora, and it is unclear whether they are still applicable to these cases. For a start, unlike space deictic expressions, anaphors do not normally convey location information to help find the antecedent term[1], e.g., in a context with only one object of the type 'cup', the redundancy in "the cup *on the table*" may facilitate search for the intended referent in a deictic situation, but less clearly so in an anaphoric context.

Secondly, anaphora resolution involves not only searching for the antecedent term in the text (as when searching for domain objects in space deixis) but also interpreting multiple candidate descriptions (including those of the competing discourse objects, or distractors, and the real antecedent term.) Descriptions of distractor objects may vary greatly in the number of attributes that they share with the referring expression, which may somehow have an impact on the overall resolution effort. For instance, given the antecedent term "the large white cat" and the anaphor "the white cat", the reader may come across distractors such as "the small cat", "the large black cat" and so on, each of them representing a particular obstacle to resolution.

Finally, the work in [3] does not distinguish between *discriminatory* and *non-discriminatory* redundancy, that is, it is not clear how redundant information impacts resolution when it may help ruling out distractors (e.g., the use of a redundant attribute 'white' in a context in which all distractors are black) or not (e.g., the same, in a context in which some of the distractors are also white.)

3. Experiment Design

We are interested in collecting evidence of how redundancy may affect anaphora resolution (i.e., the task of interpreting the referring expression and then identifying the antecedent term in the previous text), so that in the future this information could be taken into account in the development of more human-like REG algorithms. To this end, we designed an experiment in which subjects were instructed to identify anaphoric antecedents of descriptions conveying various degrees of redundancy in a number of documents in electronic format, while their navigation steps and resolution times were recorded with millisecond precision.

Subjects. 38 native speakers of Brazilian Portuguese.

Procedure. All subjects were shown 13 documents in electronic format in random order. Each document conveyed a short text in a randomly selected domain (e.g., cars, pets, books etc.) Each text was shown one paragraph at a time. Subjects were told to read each paragraph and scroll down to reveal the next one. Upon reaching the end of the text, an instruction of the kind 'Click on the expression that refers to a X in the text' was displayed, in which X was an unambiguous anaphoric expression. Although subjects most likely did not read the entire text, but simply skimmed through it to find each instruction at the end, the experiment setting forced them to browse the text in linear order from the beginning, and did not allow them to skip to the instructions. This was done to provide a general idea of the text topic and size.

After interpreting the given instruction and hitting a 'start' button at the end of each document, the subject was free to backtrack and locate the referred antecedent term in the previous text. Because the text was shown one paragraph at a time, the subject was forced to use the navigation (up / down) arrows to find the required information, which ensured incremental interpretation.

Each text conveyed a number of clickable elements representing the actual antecedent *a* and alternative candidates. To prevent subjects from trying to find the answer simply by looking for special formatting (e.g., hyperlinks), clickable elements were visible only when the mouse pointer was passed over them. Clicking on a wrong answer or going beyond the antecedent position would produce an error message and a new text to be randomly selected, that is, the experiment could only be finalized once the 13 correct answers were found[2].

After each correct answer the subject was directed to the next text, until the end of the experiment. The instruction conveying the referring expression was permanently shown at the bottom of the screen to remind the subjects of their task. All navigation steps and times were recorded during the entire resolution procedure, that is, from the moment that the 'start' button was hit until the correct answer was selected.

Redundancy. We would like to test whether adding logically redundant attributes (either discriminatory or not) to a referring expression may affect resolution times. Redundancy in this case is viewed as a combination of two factors: the number of redundant attributes conveyed by the referring expression and their discriminatory power. Starting from a basic expression conveying the referent type and a single discriminatory attribute (e.g., 'the black cat'), we will consider the addition of four degrees of redundancy: minimal descriptions or zero redundancy (0), one discriminatory attribute (+1), one non-discriminatory attribute (-1) and two attributes (2), in this case being one attribute discriminatory and the other not. Other attribute combinations were not considered for practical reasons[3].

Referential Context. Besides looking into situations of reference with various degrees of redundancy, we will also vary the degree of complexity of the context by making use of *distractors*, that is, discourse objects of the

[1] Unless we were to consider the special case of *textual anaphora* as "the first word in the above paragraph" [4].

[2] This allowed us also to filter out subjects that produced an overly large number of mistakes, and who may not have taken the experiment seriously.

[3] The degree of redundancy expressed by two simultaneous attributes seemed less relevant to our present investigation, and perhaps less common in language use as well.

same type as the antecedent *a*, and which are placed between the anaphoric expression and *a*, so that the reader is forced to take them into account during resolution. Contextual complexity will be modelled as the number of distractors found before the antecedent *a*, which may vary from zero to two (recall that the reader searches backwards from the referring expression.)

Of course another relevant factor in contextual complexity would be the degree of contrast of the distractor compared to *a*, that is, the number of attributes that the distractor shares with *a*. In our experiment setting this means that we should take into account two kinds of distractor: a less contrastive kind that we call *c1*, which differs from *a* by one attribute, and a more contrastive variety called *c2*, which differs from *a* by two attributes. For instance, a possible *c1* distractor for the referent of "large black cat" would be "large white cat", and a *c2* distractor would be "small white cat".

The issue of how many attributes are shared between distractor and antecedent terms may be an interesting one, but in order to avoid testing every referring expression in 6 different contexts (3 context sizes * 2 distractor types) we presently do not take distractor types into account, that is, we eliminate this possible effect by using a uniform distribution for individual *c1* and *c2* distractors, and also for the presentational order of (*c1*,*c2*) pairs. In practice this means that (depending on the random text selection used in each experiment) different subjects may come across more objects of type *c1* or *c2*. As we discuss later, this has no impact on our hypothesis testing regarding *c1* and *c2* objects since we will limit this investigation to the cases in which both appear simultaneously.

The four degrees of redundancy (0, +1, -1 and 2) and the three degrees of contextual complexity (0, 1 or 2 obstacles) give rise to 4 * 3 = 12 situations of reference to be examined, each of them corresponding to a statement in the experiment.

Table 1 – Research statements

#	Redundancy	Distractors
01	0	0
02	0	1
03	0	2
04	+1	0
05	+1	1
06	+1	2
07	-1	0
08	-1	1
09	-1	2
10	2	0
11	2	1
12	2	2

Research questions. Results in [3] suggest that adding redundant attributes to *deictic* descriptions in spatial domains (e.g., buildings divided into rooms etc.) may facilitate resolution. However, given that anaphora resolution involves interpreting multiple candidate descriptions and matching them to the anaphor term to decide which one is co-referential, our central hypothesis is that the effect of redundancy may in this case be *precisely the opposite*, that is: longer descriptions demand more time for the identification of discourse objects (i.e., the antecedent term and distractors.) But this is not to say that one should expect an increase in the overall resolution time when redundancy is included: unlike space deictic descriptions that use redundant information to *locate* the intended referent (e.g., "the cup *on the table*" in a context in which there is only one such cup), anaphora resolution is not generally facilitated in this way, and it is unlikely that finding an antecedent in the text will take any longer if we write e.g., "the cup on the table" instead of simply "the cup". For that reason, instead of looking into overall resolution times we will examine the identification times of individual discourse components, namely, the antecedent term *a* and *c1* and *c2* distractor objects.

We will use the notation *time(r, x)* to represent the average identification time spent in contexts conveying 0..2 distractors while examining the object *x* (which can be the antecedent *a* or a distractor *c1* or *c2*) given a description of redundancy degree *r*. All times are measured from the moment in which *x* is displayed on screen until the moment that a navigation action is performed (e.g., moving up or down in the text, or selecting *x*) in which presumably the antecedent or distractor has been identified as such. The relationship between redundancy and identification will be tested by comparing identification times of a given short description (0 degree of redundancy) and a long one (+1, -1 or 2 degrees of redundancy.) Additionally, we will also compare short (+1 and -1 descriptions) with long (2) descriptions when relevant.

In our pilot experiment we consider the following four research questions (*h1*-*h4*):

h1: The use of longer descriptions increases the identification time of anaphoric antecedents.

$$time(0, a) < time(r, a), r \neq 0.$$

This hypothesis will be tested by comparing the time spent examining the antecedent of short descriptions (statements 01..03) and those conveying +1, -1 or 2 degrees of redundancy (statements 04..12) Additionally, we will compare descriptions conveying one degree of redundancy (r=+1 and -1) with those conveying two degrees (r=2). In all cases we expect longer descriptions to demand longer identification time.

h2: The use of longer descriptions increases the identification time of less contrastive distractors.

$$time(0, c1) < time(r, c1), r \neq 0.$$

This hypothesis will be tested by comparing the time spent examining *c1* distractors (and hence deciding that *c1* was not the correct antecedent term) in contexts involving both *c1* and *c2*. More specifically, we will compare the time spent on *c1* given a short description (statement 03) with the time spent given descriptions of +1, -1 or 2 degrees of redundancy (statements 06, 09 and

12.) In all cases we expect longer descriptions to demand longer identification time[4].

h3: The use of longer descriptions increases the identification time of more contrastive distractors.

$$time(0, c2) < time(r, c2),\ r \neq 0.$$

This hypothesis is analogous to *h2*. We will compare the time spent on *c2* given a short description (statement 03) with the *c2* time given descriptions of +1, -1 or 2 degrees of redundancy (statements 06, 09 and 12) In all cases we expect longer descriptions to require longer identification time.

h4: Identifying more contrastive distractors (c2) is faster than identifying less contrastive ones (c1).

$$time(r, c2) < time(r, c1),\ r \neq 0.$$

This hypothesis states that *c1* distractors always require longer identification time than *c2* regardless of the degree of redundancy of the referring expression. This will be tested by comparing the time spent on *c2* and *c1* in all situations in which both occur (statements 03,06,09 and 12.) In all cases we expect more contrastive distractors to require shorter identification time.

We had originally made additional predictions about the possible relationship between degrees of redundancy and misidentification (e.g., selecting *c1* or *c2* instead of the antecedent term.) However, misidentification turned out to be almost inexistent in our data due to our strict implementation that forces the subjects to carefully select each antecedent to obtain the required 13 correct answers to reach the end of the experiment. Thus, this analysis was not possible in our experiment setting.

Materials. 146 purpose-made documents in electronic format, conveying one statement each. Two documents were only intended to familiarize the subject with the experiment setting, and had no other research purpose. The reminder 144 research documents represent our 12 possible document configurations in 12 different domains (pets, vehicles etc.) This level of variation was deemed necessary to avoid domain and other linguistic effects[5], and also to prevent the subjects from relying on memory. It should however be made clear that each subject had only to find the correct answer in 12 different (and randomly selected) research documents, being each one in a different domain and presented in random order (besides the practice documents at the beginning.)

All documents kept the same sentence structure and number of words between the referring expression and the antecedent term. The language used was kept as simple as possible, and making use of highly visual, concrete discourse objects. Besides the basic entity type, referring expressions conveyed three kinds of attribute: colour, location and size. In all statements, colour and location were discriminatory attributes that could be made redundant or not depending on the contents of the description in which they appeared, whereas size was always non-discriminatory and thus always redundant.

4. Preliminary Results

38 Information Systems students completed the experiment. Table 2 shows the average identification time for the antecedent *a* and distractors of type *c1* and *c2* in situations involving 0, 1, -1 and 2 degrees of redundancy (denoted as *r0*, *r1*, *r-1* and *r2*.)

Table 2 – Average identification times (seconds)

object	r0	r1	r--1	r2
a	2.72	3.58	2.93	3.67
c1	1.81	2.00	2.16	3.03
c2	2.01	1.84	2.04	1.79

Informally speaking, it is immediate to observe that for the three kinds of objects (antecedent *a*, *c1* and *c2*) the data show a tendency (either of increase or decrease) from *r0* to *r2* that is interrupted only by the *r-1* cases. In fact, if we disregard the *r-1* column in Table 1 above we notice that the identification times of both antecedent and *c1* always *increase* according to the degree of redundancy, and, analogously the identification times of *c2* always *decrease*. In addition to that, the above identification times show that the difference from *r0* to *r2* is fairly large, but less so between *r0* and *r1* or *r-1*. This seems to suggest that our experiment setting was not entirely adequate to measure the subtle effect of adding one single attribute to a non-redundant description, or to distinguish between discriminatory (*r1*) and non-discriminatory (*r-1*) redundancy. Accordingly, our results below were mainly significant when comparing the difference between *r0* and *r2*, and for that reason we will refer to them simply as short/long descriptions.

The results for hypotheses *h1-h3* (those comparing identification times for *a*, *c1* and *c2* in short and long descriptions) using Wilcoxon signed-rank test are significant as stated in the following Table 3-5. In *h3* the observed effect was in the opposite direction.

Table 3 – h1: antecedent identification

Test	N	T	%	p
r0 < r1	24	47.50	71.05	0.0034
r0 < r2	21	42.00	60.53	0.0106
r−1 < r2	22	59.00	71.05	0.0284

[4] Contexts involving one distractor only (statements 02,05,08 and 11) convey *either c1* or *c2*, which does not allow us to draw a balanced comparison between them.

[5] For example, a reader more familiar with (or more interested in) a particular subject may pay more attention to that text, and this may impact resolution.

Table 4 – h2: c1 identification

Test	N	T	%	p
r0 < r−1	19	49.00	55.26	0.0642
r0 < r2	24	87.50	63.16	0.0742
r1 < r2	23	77.50	65.79	0.0658

Table 5 – h3: c2 identification

Test	N	T	%	p
r0 > r2	20	60.50	63.16	0.0966

Results for hypothesis *h4* (the comparison between *c1* and *c2* identification times) are significant as in Table 6 below, that is, only for the longest descriptions.

Table 6 – h4: c2 < c1 test

Redundancy	N	T	%	p
r2	20	35.00	76.32	0.0090

5. Discussion

When the shortest (*r0*) and the longest (*r2*) descriptions are compared, all tests showed significant change in resolution times of the antecedent, *c1* and *c2* distractors. On the other hand, as suggested in the previous section, our experiment setting was unable to detect significant differences between *r0* and *r1* / *r-1* descriptions in most tests. The exceptions were two significant effects in antecedent identification (those between *r0* and *r1*, and between *r-1* and *r2*) and one effect in *c1* identification (between *r0* and *r-1*.)

Despite these limitations, these results seem to suggest that redundancy does increases identification times of antecedent terms and *c1* (i.e., less contrastive) objects, which can be explained by the fact that reading longer descriptions simply takes longer. The effect is particularly significant for antecedent identification (*h1*), but also observable for *c1* (hypothesis *h2*.)

On the other hand, results for *c2* (i.e., more contrastive) objects were remarkably opposite to the predictions in *h3*: redundancy in this case actually *decreases* identification times, that is, making resolution easier. This is further confirmed by the findings for *h4*, in which the identification of *c1* objects took much longer that the identification of *c2* (recall that these were measured under exactly the same situations of reference.)

A possible explanation for this difference is the cognitive load involved in the identification of various competing descriptions (of *a*, *c1* or *c2*). Given a referring expression *i* and a candidate referent *j*, the reader is supposed to interpret both *i* and *j* and decide whether they match. Even though redundancy does increase reading times, matching *i* and *j* may still require relatively little cognitive effort when both *i* and *j* share a large number of properties (or words) if compared to matching a more dissimilar description pair. In other words, we hypothesize that for a closely-related pair of descriptions (i.e., an anaphor and a *c1* object, or an anaphor and the actual antecedent term) the readers may benefit from some form of shallow processing to quickly decide, for example, that the reference "the small black cat" is not the same as (*c1*) "the small white cat", but that it does co-refer with an antecedent term "the black cat".

By contrast, when facing a more complex match between "the small black cat" and a candidate conveying two unexpected attributes (*c2*) as in "the large white cat", it may be necessary to resort to a somewhat deeper analysis, which would explain why *c2* reading times are longer. Although we presently do not seek to validate this claim, this intuition seems to be consistent with the behaviour reported in informal interviews with some of the experiment subjects. We believe that more research will be required to clarify this issue.

Finally, with respect to the comparison between deixis and anaphora resolution, our preliminary results for anaphora are quite dissimilar from those reported in [4] for space deixis. Although this was to a large extent to be expected (as hypothesised in *h1-h3*), the present outcome seems to suggest a far more complex picture that once again will require further investigation.

6. Final Remarks

Despite the small scale of our pilot experiment, preliminary results suggest a number of interesting issues to be taken into account in the design of REG algorithms. Chief among them, we found that redundancy may in fact *increase* anaphora resolution times. While this is not to say that redundancy should be simply avoided (e.g., redundancy may reduce misidentification or improve comprehension), this insight is quite contrary to existing knowledge on the generation of deictic descriptions.

In addition to that, redundancy seems to have different effects depending on the kinds of redundant attribute used (i.e., discriminatory or not) and affects more or less contrastive distractors in different ways. All in all, there is no straight answer as to whether to generate redundant descriptions or not, but rather that a number of context details need to be taken into account.

The experiment provided us with a large and complex data set that we have only started to analyse. As future work we intend to make additional inferences from these results and redesign a number of aspects of the experiment setting including a larger number of subjects.

7. Acknowledgements

This work has been supported by CNPq and FAPESP.

8. References

[1] Grice, H. P. "Logic and Conversation". In P. Cole and J. L. Morgan (eds.) *Syntax and Semantics*, Vol. iii: Speech Acts. New York, Academic Press, pages 4-158, 1975.

[2] Dale, R. and E. Reiter. "Computational interpretations of the Gricean maxims in the generation of referring expressions". Cognitive Science 19(2), 1995.

[3] Paraboni, I., K. van Deemter and J. Masthoff "Generating Referring Expressions: Making Referents Easy to Identify". Computational Linguistics 33(2), 229-254.

[4] Lyons, John. *Semantics*. Cambridge Univ. Press, 1977.