Running head: SURFACE ISSUES IN GRE

Managing ambiguity in reference generation: the role of surface structure

Imtiaz Hussain Khan

Department of Computer Science

King Abdulaziz University

Jeddah K.S.A.

ihkhan@kau.edu.sa

Kees van Deemter and Graeme Ritchie

Department of Computing Science

University of Aberdeen

Aberdeen AB24 3UE, U.K.

{k.vdeemter,g.ritchie}@abdn.ac.uk

Abstract

This paper explores the role of surface ambiguities in referring expressions, and how the risk of such ambiguities should be taken into account by an algorithm that generates referring expressions, if these expressions are to be optimally effective for a hearer. We focus on the ambiguities that arise when adjectives occur in coordinated structures. The central idea is to use statistical information about lexical co-occurrence to estimate which interpretation of a phrase is most likely for human readers, and to avoid generating phrases where misunderstandings are likely. Various aspects of the problem were explored in three experiments in which responses by human participants provided evidence about which reading was most likely for certain phrases, which phrases were deemed most suitable for particular referents, and the speed at which various phrases were read. We found a preference for "clear" expressions to "unclear" ones, but if several of the expressions are "clear" then brief expressions are preferred over non-brief ones even though the brief ones are syntactically ambiguous and the non-brief ones are not; the notion of clarity was made precise using Kilgarriff's Word Sketches. We outline an implemented algorithm that generates noun phrases conforming to our hypotheses.

Keywords: natural language generation, generation of referring expressions, ambiguity management, surface ambiguity in GRE, surface realization

Managing ambiguity in reference generation: the role of surface structure

Introduction

When designing a computer system that can produce written text in natural language (e.g. English) from some underlying non-linguistic representation of information (Reiter & Dale, 2000), an important component is the *Generation of Referring Expressions* (GRE). GRE algorithms generate linguistic expressions that allow a reader/hearer to identify an intended set of referent(s). Such algorithms typically start from information stored in a non-linguistic, computer-oriented form, such as a conventional database or tables of numerical data. The GRE task involves both *Content Determination* (i.e. 'what to say') and *Linguistic Realization* (i.e. 'how to say it'); most existing research focusses on the first of these issues. However, even when Content Determination yields a single unambiguous result, ambiguity can be introduced subsequently as a result of Linguistic Realization, thus causing a risk of confusion for the reader. This paper asks what would be the best way to manage this risk. We are considering relatively simple uses of language, and are not considering literary, rhetorical or humorous usage, where the role of ambiguity may be different.

Following van Deemter (2004), we hypothesize that GRE algorithms can generate referring expressions that are "better for readers" if these algorithms take linguistic ambiguity into account, by assessing how likely an ambiguity is to cause misunderstanding. Earlier work suggests that every sentence is potentially ambiguous between many parses, even though we may not notice this ambiguity (Abney, 1996; Wasow, Perfors, & Beaver, 2005). This suggests that it may not be feasible to avoid all referential ambiguities all the time, and that the choice of referring expression should

sometimes involve a balancing act in which degree of ambiguity is balanced against other properties of the generated expression, such as its length or fluency (van Deemter, 2004).

This paper examines how GRE should deal with structural ambiguity, focussing on ambiguities of the form *the Adj Noun*$_1$ *and Noun*$_2$, henceforth *coordination* ambiguities. We call referring expressions of this form **scopally ambiguous**, because the scope of *Adj* is ambiguous between wide-scope (*Adj* applies to both nouns) and narrow-scope (*Adj* applies only to *Noun*$_1$). As we are interested in these phenomena from a generation perspective, we make the assumption that the speaker/generator always has an intended referent set to be described, and hence there is a well-defined **intended interpretation** to convey.

Our approach to the problem is to assess the likelihood of each interpretation of an NP, and to tailor GRE to avoid ambiguities which are likely to cause misunderstanding. The problem is how to determine which ambiguities are liable to misunderstanding and which ones are not. In this paper, we investigate the use of language corpora to answer this question. The core idea is that misunderstanding is probable when the most likely interpretation of a phrase (statistically) is not the intended interpretation of that phrase.

Here we report on three investigative studies. The first study asks *can corpus data be used to find the likelihood, for readers, of different interpretations of a given* NP. Since the least ambiguous expression might not always be the one preferred by most readers, perhaps because it is very lengthy, or very disfluent, our second study asks *which of several* NP*s are preferred by a reader* (for a given referent set). Finally, we measured the reading and comprehension times for expressions which would be classified differently according to the corpus-based model we have developed.

We also formulated a GRE algorithm based on these findings.

Related work

A framework for GRE

Earlier GRE research is dominated by algorithms that generate referring expressions whose intended referent set is a single object, for example, *the small dog*. Most GRE algorithms start by building semantic descriptions which contain enough information to identify the referent, so called Distinguishing Descriptions (DDs). DDs are made up of properties available in a *knowledge base*. Thus, given a domain D of entities and an intended referent $r \in D$, GRE algorithms find a subset S of properties $p_1, p_2, ..., p_n$ such that $[\![p_1]\!] \cap [\![p_2]\!] ... \cap [\![p_n]\!] = \{r\}$, where $[\![p]\!]$, the extension of a property p, stands for the set of objects which have the property p. Perhaps the best known algorithm of this kind is the Incremental Algorithm, which iterates through a list of properties, adding properties to the description one by one, until a DD is found or the list of properties is exhausted (Dale & Reiter, 1995).

In recent years, a number of proposals have been made for allowing GRE algorithms to produce plural referring expressions (i.e. reference to arbitrary sets of objects) (Stone, 2000; van Deemter, 2002; Gardent, 2002; Horacek, 2004; Gatt, 2007), and this is the type of algorithm that we are focussing on in this article.

Surface structure in GRE

Most GRE algorithms produce abstract, semantic DDs. Only a few produce actual words (Stone & Webber, 1998; Krahmer & Theune, 2002; Siddharthan & Copestake, 2004). Siddharthan and Copestake address the need for (sometimes) avoiding lexical ambiguities, but no GRE algorithm to date addresses structural ambiguities (although they are mentioned in Gardent (2002)). Yet ignoring Linguistic Realization is hazardous: linguistic ambiguities can be introduced when a DD is realized using words and phrases. Such surface ambiguities can cause confusion concerning the intended referent of the

description.

To see this, consider a meadow with various animals, and the text generator's task is to single out the black sheep and black goats from the rest of the animals. We shall represent DDs using "⊓" for conjunction, and "⊔" for disjunctions. Suppose a GRE algorithm has generated the DD *(black* ⊓ *sheep)* ⊔ *(black* ⊓ *goats)*. This could be realized as: *the black sheep and the black goats*, or *the black sheep and goats*. The former NP is structurally unambiguous, but lengthy and disfluent; the latter is potentially ambiguous between *(black* ⊓ *sheep)* ⊔ *goats*, and *black* ⊓ *(sheep* ⊔ *goats)*; only the latter is logically equivalent to the DD to be conveyed. This example highlights the possible tension between brevity and avoidance of potential ambiguity. The question facing us is how to balance these two conflicting factors.

GRE *evaluation*

In recent years, the NLG community has seen a substantial number of studies to evaluate GRE algorithms (see, for example, Belz and Gatt (2007); Gatt, Belz, and Kow (2008, 2009)). Most evaluations have focussed on the semantic content of the generated descriptions, as produced by the Content Determination stage of a GRE algorithm; this means that linguistic realization (i.e. the choice of words and linguistic constructions) is seldom addressed. Secondly, most existing evaluations are speaker-oriented, focussing on the degree of "humanlikeness" of the generated descriptions, disregarding their effect on the hearer or reader. In this paper, we are exploring how "effective" a GRE algorithm is for a reader.

Disambiguating coordination ambiguities

Language corpora have been used to resolve ambiguities in natural language utterances (Binot & Jensen, 1987; Wu & Furugori, 1998; Chantree, Kilgarriff, De Roeck, & Willis, 2005; Chantree, 2006; Willis, Chantree, & De Roeck, 2008), and this approach is

supported by psycholinguistic studies which found that word frequencies play a positive role in resolving syntactic ambiguities (see, for example, Trueswell (1996); Merlo and Stevenson (1999)). A few studies have taken these considerations to language generation (Inui, Tokunaga, & Tanaka, 1992; Neumann, 1994; van Deemter, 2004).

Our own use of corpora for language generation is closest in spirit to Chantree et al. (2005); Chantree (2006) and Willis et al. (2008). Like these authors, we use corpus-based heuristics, based on information obtained from the the British National Corpus[1] (BNC) via Kilgarriff's Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004), to identify those ambiguities which are liable to be misunderstood. The Sketch Engine generates summaries of words' grammatical behavior, known as Word Sketches. The Word Sketches give information about the frequency with which words are linked by a given grammatical relation. Rather than looking at an arbitrary window of text around a given word, the correct collocations are found by use of grammatical patterns (Kilgarriff et al., 2004). Suppose we want to generate the Word Sketches for the word *old* (node word). Upon receiving this word along with its part-of-speech, the Sketch Engine provides one list of collocates (argument words) for each grammatical relation *old* participates in, along with a salience score, which is calculated from the overall frequencies of the node word and the argument word, in the BNC. For example, for the *modifies* relation, a truncated list is [<lady,43.01>, <age,38.33>, <man, 36.43>,...]. This example suggests that *old* modifies *lady* (*old lady*) more often than *man* (*old man*), because the former receives a higher salience score.

Given a potentially ambiguous NP, corpus-based information offers a resource through which it may be possible to estimate the probabilities of each of the different interpretations of this NP. If these probabilities indicate the interpretation(s) most likely to be perceived by hearers/readers, it may be possible to separate ambiguities that are likely to lead to misunderstanding from those which are not. Working in a language

generation context, for example, van Deemter (2004) considered an expression as *viciously* ambiguous if the intended interpretation failed to be considerably more probable than all other interpretations (some of which might even be more likely than the intended interpretation). Following an experiment in which human judgments were gathered, Chantree and colleagues coined the closely related notion of a *nocuous* ambiguity, which they defined as cases where there is a lack of consensus amongst readers on a single reading (Chantree et al., 2005; Chantree, 2006; Willis et al., 2008). With this concept in mind, they proposed the following three heuristics, each of which involves a class of expressions where there is no nocuous ambiguity because a "preferred" interpretation (of an NP of the form *Adj* $Noun_1$ *and* $Noun_2$) can be predicted with reasonable reliability.

First, the *Coordination-Matches* Heuristic: if the two head words appear frequently within the corpus as a coordination, then a wide-scope reading is likely. For example, the coordination *cats and dogs* occurs frequently; therefore, in *the black cats and dogs*, *black* is predicted to have a wide scope. Second, the *Distributional-Similarity* Heuristic: if the two head words in a coordination show a high distributional similarity, then a wide-scope reading is likely. For example, the nouns *boots* and *shoes* have a high distributional similarity, therefore in *the brown boots and shoes*, *brown* is predicted to have wide scope. Third, the *Collocation-Frequency* Heuristic: if a modifier is collocated more frequently with the nearest head word than with the head word further away, then a narrow-scope reading is likely. For example, in *the bald men and women*, *bald* is predicted to have narrow scope because *bald* rarely modifies *women*.

In our own investigations, we decided to follow Chantree et al. in their use of the Word Sketches, but we also made a number of changes. We used Word Sketch information in a way which captures essentially the same patterns as the Coordination-Matches and Collocation-Frequency heuristics, but we made no use of the Distributional Similarity heuristic, which Chantree et al. describe as being a weak indicator to predict wide scope.

Moreover, the classification of ambiguity as nocuous/innocuous is insufficiently fine-grained for our purposes, because it does not take the intended interpretation into account. For example, if Chantree et al.'s assessments are correct, then *bald men and women* will have a consensus narrow-scope reading, and hence be "innocuous"; however, if the intended interpretation actually involves bald men and bald women (i.e. the wide-scope reading), then that will lead to misunderstanding. Van Deemter's notion of "vicious" ambiguity covers these cases as well as Chantree et al.'s: an utterance is viciously ambiguous if the intended interpretation cannot be predicted safely (van Deemter, 2004).

An outline of our model

Before describing our experimental studies, we first sketch the concepts we shall be using.

Where an interpretation corresponds to an NP of the sort we are investigating, it may be referred to as a "narrow-scope" or "wide-scope" interpretation, depending on which scope of the Adjective it corresponds to. Thus $(old \sqcap men) \sqcup (old \sqcap women)$ will be said to be a wide-scope interpretation of *the old men and women*, and $(old \sqcap men) \sqcup women$ would be the narrow-scope interpretation. That is, we generalize the notion of "scope" from the textual NP to the underlying representation of the interpretation.

In scopally ambiguous expressions, there is a tension between wide- and narrow-scope interpretations. This can be viewed in terms of two competing forces: a *Coordination Force*, whereby $Noun_1$ and $Noun_2$ attract each other to form a syntactic unit, and a *Modification Force*, whereby *Adj* and $Noun_1$ attract each other to form a syntactic unit. We define that there is a Strong Coordination Force (SCF) if the collocational frequency between the two nouns is *high*, and a Weak Coordination Force (WCF) if the collocational frequency is *low*. Similarly, there is a Strong Modification

Force (SMF) if the collocational frequency of *Adj* is *high* with $Noun_1$ and *low* with $Noun_2$, and a Weak Modification Force (WMF) otherwise. After a preliminary investigation of the data, we decided to operationalize high collocational frequency between two words as meaning that either of the two words appears among the top 30% collocates of the other word in the grammatical relation of interest; low collocational frequency means that neither of the two words appears among the top 70% collocates of the other word in the grammatical relation. Importantly, this means that between 30% and 70% frequency is considered to be neither low nor high; hence, a phrase could manifest neither "strong" nor "weak" Coordination/Modification Force.

Experiment 1: Interpreting NPs

We aim to build a generator which should avoid noun phrases that are liable to misunderstanding. But misunderstandings cannot be ruled out, and if a hearer misunderstands a noun phrase then secondary aspects such as reading (and/or comprehension) speed are of little consequence. We therefore plan first to find out the likelihood of misunderstanding.

Hypotheses. We formulated four hypotheses which represent all four possible combinations of high and low coordination and modification forces to predict an interpretation of a scopally ambiguous NP.

Hypothesis 1: If there is an SCF and an SMF, then a narrow-scope reading is the most likely.

Hypothesis 2: If there is an SCF and a WMF, then a wide-scope reading is the most likely.

Hypothesis 3: If there is a WCF and an SMF, then a narrow-scope reading is the most likely.

Hypothesis 4: If there is a WCF and a WMF, then a wide-scope reading is the most likely.

Hypotheses 2 and 3 are intuitively obvious, because both forces operate in the same direction. Hypotheses 1 and 4 are based on some small preliminary studies that we carried out. In these studies we observed that participants showed a strong tendency towards wide scope when they encountered the NPs involving both WCF and WMF (Hypothesis 4), and a strong tendency towards narrow scope when the NPs involved SCF and SMF combination (Hypothesis 1). For Hypothesis 1, we also take into account the results of Chantree (2006) and Willis et al. (2008), who found that if a modifier is collocated more frequently with the nearest head word than with the head word further away, then a narrow-scope reading is likely, no matter how frequent the two head words are.

As noted above, a phrase could – with our definitions – manifest neither SCF nor WCF, neither SMF nor WMF. Hence there will be phrases where our hypotheses make no prediction about scope.

Materials and Design. In this experiment, referential domains were depicted using Euler diagrams. In our version of Euler diagrams, convex contours represent sets of things sharing common characteristics: we *shaded* an area of the diagram to indicate the set of things to which the diagram was referring. Scopally ambiguous NPs can be associated with pairs of Euler diagrams as shown in Fig. 1, where each diagram represents a separate interpretation of the NP. We take the diagram on the left to mean the NP *the young lions and (all) the horses*, and the diagram on the right to mean the NP *the young lions and the young horses*.

A trial in this experiment consisted of 2 to 4 Euler diagrams and an English NP displayed underneath these diagrams. One diagram corresponds to a wide-scope reading; one corresponds to a narrow-scope reading. We varied the number of figures from two to

four to minimize the risk that the participants might figure out the purpose of the study. We also included sixteen filler items, containing NPs that do not contain a coordination, for instance, *the dogs on the left*.

The nouns and adjectives used in the scopally ambiguous NPs were chosen from the BNC as follows: we first extracted a sample of words which met a certain criterion (collocational frequency value), then selected randomly from that sample. Four pairs of nouns were used: two with SCF, and two with WCF. For each pair of nouns, four different adjectives were used: two with SMF, and two with WMF.[2]

This gave us a total of eight different nouns and sixteen different adjectives, i.e., four cases per hypothesis. Each participant was presented with all sixteen experimental items, together with sixteen filler items, a total of thirty-two trials.

Procedure and Participants. The experiment was carried out over the Web. Participants were students from various UK universities who were approached via email. Before the experiment, participants received a mini-tutorial on our version of Euler diagrams. Items were ordered for presentation in such a way that there was at least one filler item between two experimental items, but otherwise randomly. For each item, participants removed (by a mouse click) the figure that they thought was referred to by the NP. Participants were allowed to withdraw from the experiment at any stage. Data were gathered from sixty-five self-reported native or fluent speakers of English. Sixty participants completed the experiment.

Results and Discussion. Results were recorded according to whether a participant opted for a wide- or narrow-scope reading. If a participant selected a filler diagram in an experimental trial, we assigned wide/narrow-scope reading randomly to the corresponding NP, to avoid missing data points. Only 23 data points (2.39%) of a total 960 were treated in this manner. The participants' responses are shown in Table 1. The data show that a

reasonably high proportion of participants' judgements are in favor of our hypotheses. We take this to indicate that the Word Sketches can contribute to predicting the most likely reading of scopally ambiguous coordinated NPs. A one-tailed sign binomial test further revealed that the results are statistically highly significant[3] [$p < 0.001$]. We also observed, however, that in these NPs, a narrow-scope reading tends to be particularly frequent in the *extreme* case where *Adj* has a zero co-occurrence with $Noun_2$ in the BNC. Therefore, we shall use a modified version of Strong Modification Force (SMF): SMF$'$ will mean that *Adj* and $Noun_2$ have *zero* (rather than below 30%) co-occurrence in the BNC.

The fact that all four hypotheses were confirmed allows us to summarize our findings using the following Prediction Rules: (As before, WS is Wide Scope, NS is Narrow Scope, SMF$'$ is Strong Modification Force, and WMF is Weak Modification Force.)

1. WMF $\rightarrow$ WS
2. SMF$'$ $\rightarrow$ NS

It is worth mentioning here that we used a small and engineered dataset. On the one hand, this allows us to focus on specific and manageable phenomena in a simple experimental design in which every participant is presented with every item. On the other hand, a small dataset can cast doubts on the generalizations which we drew from our sample. However, since the sample NPs were randomly selected from a diverse corpus, and our findings corroborate those of Chantree (2006) and Willis et al. (2008), we are confident that our generalizations are on the right track.

We have seen that the Word Sketches can offer helpful predictions to a generator concerning the likelihood of interpretations. But an NP that is not likely to be misunderstood may have other disadvantages. For example, it may lack fluency or it may be perceived as unnecessarily lengthy. In such cases, the question comes up which factor should weigh more heavily in a generator's choice between noun phrases: the length of the expression or the lack of ambiguity. For this reason, we conducted a study in which

readers' preferences were tested.

Experiment 2: Readers' Preferences

The question of how to choose between different NPs could be approached in a number of different ways: asking hearers which of several descriptions they prefer, asking hearers to rate several descriptions, measuring hearers' processing effort, measuring hearers' errors etc. Here we report on a forced-choice readers' preference experiment in which participants were asked to compare pairs of natural language descriptions of one and the same target set, selecting the one they found more appropriate.

Two main factors, *brevity* and *clarity*, are manipulated. 'Brief' descriptions took the form *the Adj Noun*$_1$ *and Noun*$_2$. 'Non-brief' descriptions took the forms *the Adj Noun*$_1$ *and the Noun*$_2$ (for narrow scope) and *the Adj Noun*$_1$ *and the Adj Noun*$_2$ (for wide scope). That is, 'brevity' has a specialized sense involving the presence/absence of the determiner (*the*) and possibly *Adj* before the second noun. Importantly, the 'non-brief' expressions are always syntactically unambiguous, but the 'brief' NPs are potentially ambiguous. We call an NP *clear* if it is syntactically unambiguous or the scope of its intended interpretation is the same as the one predicted by our rules based on WMF and SMF$'$, otherwise the NP is *unclear*. Fig. 2 displays this classification. (The figure includes annotations indicating *roughly* how Chantree and van Deemter's terminology is related, but their formalizations are not exactly the same as ours.) We hypothesize that:

Hypothesis 5: Keeping clarity the same, people will prefer brief expressions over non-brief ones.

Hypothesis 6: Where only one of clarity and brevity can be achieved, people will prefer clarity over brevity. (In other words, clarity is more important than brevity.)

Materials, Design and Procedure. Once again, referential domains were represented using Euler diagrams. In each trial, participants were shown an Euler diagram, with some of its area filled to indicate the target referent (and hence the intended interpretation). They were also shown two English NPs, which attempted to identify the filled area. The competing NPs were either clear brief $(+c, +b)$ and clear non-brief $(+c, -b)$, or unclear brief $(-c, +b)$ and clear non-brief $(+c, -b)$; one of the pairs was always clear non-brief, and thus unambiguous. (The combination unclear and non-brief $(-c, -b)$ is ruled out by our technical sense of 'non-brief': as noted earlier, 'non-brief' NPs do not have scope ambiguity.) Two sample trials are shown in Fig. 3. The intended interpretation on the left is narrow scope, and on the right it is wide scope. The diagrams make the intended interpretation obvious.

The nouns and adjectives used in the materials were drawn from the BNC using the same procedure as for Experiment 1. Four pairs of nouns (two with SCF, and two with WCF) and sixteen different adjectives (four different adjectives for each noun pair: two with SMF′, and two with WMF) were used. From these adjectives and nouns, sixteen NPs of the form *the Adj Noun*$_1$ *and Noun*$_2$ were constructed. These sixteen NPs were then used to construct thirty-two experimental trials: in sixteen experimental trials the non-brief NP (the counterpart of the brief NP) took the form *the Adj Noun*$_1$ *and the Adj Noun*$_2$ (wide scope), and in the remaining sixteen experimental trials the non-brief NP took the form *the Adj Noun*$_1$ *and the Noun*$_2$ (narrow scope).

For presentation, the items were ordered so that after every two experimental items there was a filler, otherwise randomly. Each participant was presented (after the instructions) with all thirty-two experimental trials, together with sixteen fillers, a total of forty-eight trials. The same recruitment procedure as for Experiment 1 yielded 60 participants, of whom 46 completed the experiment.

Results and Discussion. Results were coded according to whether a participant preferred the $(+c, +b)$ over the $(+c, -b)$ NP, or the $(+c, -b)$ over the $(-c, +b)$ NP. More than 79% participants preferred $(+c, +b)$ NPs over $(+c, -b)$ ones. Similarly, more than 81% participants preferred $(+c, -b)$ NPs over $(-c, +b)$ ones.

A one-way ANOVA was used to test for preference differences among three types of expressions (clear brief, unclear brief, and non-brief). Preferences for expressions differed significantly across the three types: $[F(2, 61) = 28.69, p < .001]$. To further analyze the preference for clear brief NPs over clear non-brief ones, and clear non-brief NPs over unclear brief ones, we report pairwise comparisons using a t-test. Pairwise comparisons revealed that participants preferred clear brief expressions over clear non-brief ones $[t = 7.94, p < 0.01]$, which confirms Hypothesis 5. Similarly, participants preferences for clear non-brief over unclear brief expressions $[t = 7.13, p < 0.01]$ are also statistically significant, hence confirming Hypothesis 6.

Our data set shows an unavoidable "gap". Only three types of situations are considered: (a) a description can be brief and clear (e.g. using 'the old men and women' to convey wide scope), (b) brief and unclear (e.g. 'the rowing boats and ships' for wide scope, given a prediction of narrow scope), or (c) non-brief and clear (e.g. 'the old men and the old women' for wide scope). It might be thought that there exists a fourth option: non-brief and unclear. But this is ruled out by our technical sense of 'non-brief': as noted earlier, 'non-brief' NPs do not have scope ambiguity. Because of this "missing cell", it was not possible to analyze our data using a two-way ANOVA test, which would have automatically taken care of all possible interactions between clarity and brevity.

Interim Summary and Outlook

We found evidence suggesting that Kilgarriff's Word Sketches can be used to predict the most likely reading of a scopally ambiguous coordinated noun phrase. Modification

Force emerged as the deciding factor. Moreover, *ceteris paribus*, brief descriptions were preferred over non-brief ones. These results suggest a model which (a) predicts the most likely reading of an expression using the Word Sketches, and (b) prefers clear expressions to unclear ones, but if several of the expressions are clear then brief expressions are preferred over non-brief ones. These ideas, however, needed to be tested further, because the experiments which led to this model considered only certain aspects of the hearer's reaction to NPs (e.g. meta-linguistic judgements about a participant's preferences). While this approach has the advantage that participants can directly compare expressions, the method does not tell us how difficult to process various types of expressions would actually be for hearers. We therefore conducted one final experiment, designed to tap more directly into the reading/comprehension process.

Experiment 3: Reading and Comprehension Times

To assess the readability of coordinated noun phrases, we use two indicators of hearers' benefits: *reading time* and *comprehension time*. These indicators form the basis of automatic readability metrics (Flesch, 1948; Kincaid, Fishburne, Rogers, & Chissom, 1975), and are mostly used in language processing studies (Dubinsky, Egan, Schmauder, & Traxler, 2000; Angwin, Chenery, Copland, Murdoch, & Silburn, 2006; Swets, Desmet, Clifton, & Ferreira, 2008). To measure reading times, we used self-paced reading – one of the most popular method amongst psycholinguists who study various aspects of language including readability and intelligibility (Dubinsky et al., 2000; Swets et al., 2008).

Hypotheses

We explore the hypothesis that brief expressions have an advantage over non-brief ones. In what follows, the term *predictable* is used as defined in Fig. 2, as the subclass of those brief NPs where our rules make a prediction. Taking readability and intelligibility together as 'processing', we hypothesize that:

Hypothesis 7: Participants process predictable brief expressions more quickly than non-brief expressions.

Confirmation of this hypothesis would be an indication that processing time accords with participants' explicit preference (Experiment 2) for brief rather than non-brief NPs, even though the brief ones are syntactically ambiguous and the non-brief ones are not. To gain more detailed insight into the outcome of this experiment, we separated out reading time and comprehension time (which are combined in Hypothesis 7), and framed the following additional hypotheses:

Reading Time:

RS1: Participants read predictable (brief) expressions more quickly than unpredictable (brief) ones.

RS2: Participants read unpredictable (brief) expressions more quickly than non-brief ones.

RS3: Participants read predictable (brief) expressions more quickly than non-brief ones.

Comprehension Time:

CS1: Participants comprehend predictable (brief) expressions more quickly than unpredictable (brief) ones.

CS2: Participants comprehend non-brief expressions more quickly than unpredictable (brief) ones.

CS3: Participants do not comprehend non-brief expressions more quickly than predictable (brief) ones.

The Study

Participants were presented with a sequence of trials, each of which consists of a lead-in sentence followed by a target sentence and a comprehension question. The target sentences took the form NP VP, where NP denotes a target NP. The comprehension questions were chosen to force participants to think about the meaning of sentences that contain coordination ambiguities, and the procedure (below) allowed us to separate reading time and comprehension time for the target sentences. For example, 'There is small and large furniture in the room. *The small tables and chairs* were brought from Italy'. The comprehension question in this case was 'Were the large chairs brought from Italy?'.

Materials and Design

The choice of nouns and adjectives to construct 'base NPs' (the NPs of the form *Adj* $Noun_1$ *and* $Noun_2$) is motivated by the need to have a balanced distribution of NPs in each of the following three classes. The **wide scope class** is the class of NPs for which our Prediction Rules predict a wide-scope reading; the **narrow scope class** is the one for which our Prediction Rules predict a narrow-scope reading; the **ambiguous class** is the one for which our rules do not predict a reading.

Four wide-scope class base NPs and four narrow-scope class base NPs were constructed. To balance this, eight base NPs in the ambiguous class were constructed, making a total of sixteen base NPs. Sixteen different adjectives and thirty-two different nouns (i.e. sixteen noun-pairs) were used for this purpose; they were drawn from the BNC in the same way as for Experiments 1 and 2. The sixteen base NPs were then used to construct sixteen target sentences. To allow participants to answer the questions, the scene was set by a lead-in sentence before each of the target sentences, of the general form 'There was/were/are NP_1 PREP NP_2', where NP_1 describes the entities (to be used in the

target sentence), and NP_2 denotes the location of the entities. The length of the lead-in sentence varied between nine and fourteen words; to avoid this variation in length affecting reading (or comprehension) time for the target sentences, the same lead-in sentence was used in all variants (below) of each discourse.

Four variants of each discourse were constructed, where variation is made in the base NP to represent four possible phrasings as shown in Table 2. Four lists, each containing sixteen experimental items, were constructed. Each list consisted of four groups of four items, where each group contained items like those in one column of Table 2; that is, four different phrases with the same lexical items. Of these groups, two were derived from a base NP from the wide-scope class (like column 1 of Table 2), two were derived from a base NP from the narrow-scope class (like column 2 of Table 2). This gave us sixty-four (64 = 4 x 4 x 4) discourses in total. Finally, a comprehension question was attached to each discourse item.

Each participant was presented with all sixteen experimental items (from one of the four lists) and twenty-four filler items, a total of forty trials. A trial in this experiment consisted of a two-sentence discourse followed by a comprehension question.

Procedure and Participants. Fifty-five self-reported native speakers of English took part in the experiment, and each participant was given a £5 voucher. Participants were students or employees at the University of Aberdeen who had no background knowledge of linguistics. The experiment, which lasted for approximately 30 minutes, was carried out in an experimental laboratory room at the University of Aberdeen. The participants were briefed about the purpose and format of the experiment, and given instructions followed by four practice trials; the practice data are not included in the analysis.

After the briefing and practice, participants encountered the trials, one at a time, in a different pseudo-random order. They were presented with a two-sentence discourse - one sentence at a time - followed by a comprehension question. First, they saw a row of

dashes, which covered the words in the sentences. A participant had to: press the space bar on the computer to reveal the first sentence (in the discourse), read the sentence at normal speed, press the space bar again to reveal the next sentence, read it in the same way as the first one, and press the space bar when finished reading it. The computer then presented a *yes/no* comprehension question, which participants answered by pressing one of two keys. The computer recorded: the length of time taken to read each sentence in the discourse, the response to the question, and the time taken to answer the question. After every ten trials, participants were given the option to take a break; they could withdraw from the experiment at any stage.

Data Analysis and Results. The data was analyzed against various conditions, such as comprehension time in predictable versus unpredictable cases. 76 (8.6%) out of a total of 880 data points were outlying, defined as lying at least 2 standard deviations above the mean processing time. We discarded the data from 14 participants (25.4% of the total number of participants) who either showed more than 25% (processing time) outliers, or gave responses which indicated an "incorrect" interpretation for more than 50% of the structurally unambiguous fillers (cf. (Ratcliff, 1993)). The remaining data from 41 participants were analyzed.

Results (main hypothesis):

The mean times for the experimental items appear in Table 3. The data indicate that participants processed more quickly ($\mu = 5964.08$) predictable (brief) expressions than non-brief ones ($\mu = 6362.24$). A t-test revealed that the processing time difference is statistically significant [$t = -1.77, p = 0.03$]. This confirms Hypothesis 7.

Results (additional hypotheses)

A one-way ANOVA revealed that the reading time differs significantly across the three types of expressions (predictable, unpredictable and non-brief):
[$F(2, 120) = 4.23, p = 0.02$]. Pairwise comparisons using a t-test showed that participants

read brief expressions (both predictable and unpredictable) more quickly than non-brief ones [predictable vs non-brief: $t = -.2.78$, $p < 0.01$; unpredictable vs non-brief: $t = -3.95$, $p < 0.01$]. This confirms both RS2 and RS3. However, we could not find a significant difference between predictable and unpredictable ones [$t = -.68$, $p = 2.3$], so RS1 is not confirmed.

Similarly, for comprehension time, a one-way ANOVA showed a significant difference across the expressions involved: $[F(2, 120) = 6.52, p < 0.01]$. Pairwise comparisons revealed that participants comprehended more quickly both predictable and non-brief expressions than unpredictable ones [predictable vs unpredictable: $t = -3.24$, $p < 0.01$; non-brief vs unpredictable: $t = -3.87$, $p < 0.01$]. This confirms both CS1 and CS2. Although the mean comprehension time for non-brief expressions was less than that for predictable expressions, the difference was not statistically significant [$t = -0.70$, $p = 0.25$], which still leaves open the possibility that CS3 could be correct.

When a number of different hypotheses are tested simultaneously on the same data set, a Bonferroni correction is often applied to counter data "fishing". We applied a Bonferroni correction to our six additional propositions by adjusting significance level α (i.e. dividing α by 6). In this case, the results of 4 of the 6 propositions (CS1, CS2, RS2, RS3) were still statistically significant.

The contrast between RS1 and CS1 is worth noting. We detected no difference in reading speed between "predictable" and "unpredictable" expressions (RS1), but "predictable" expressions were comprehended more quickly than the "unpredictable" ones (CS1). We can speculate that a difficult text does not slow down the reader, but the difficulties may show up in comprehension. This perspective is broadly in line with the finding of Ferreira, Ferraro, and Bailey (2002): the language comprehension system creates syntactic and semantic representations that are merely "good enough" (see also Ferreira and Patson (2007)). Ferreira et al. (2002) argued that people often obtain a shallow

understanding of an utterance meaning, or even sometimes misunderstand utterances.

Second, in *reading* time (RS2) there is a significant advantage for "unpredictable" expressions over "non-brief" ones. This could be interpreted as: ambiguities do not affect reading time, but the reading time is significantly affected by text length. This interpretation is consistent with other findings that ambiguous sentences are read faster than disambiguated ones (Traxler, Pickering, & Clifton, 1998; van Gompel, Pickering, B, & Liversedge, 2005; Swets et al., 2008). The fact that in *comprehension* (CS2) there is a significant advantage for "non-brief" expressions over "unpredictable" ones shows the trade-off between the factors of brevity and clarity.

Overall the results confirmed our main hypothesis. Additionally, four of our six additional hypotheses (RS2, RS3, CS1, CS2) were also confirmed.

Discussion

The experiment presented here reveals two primary results. First, participants read brief expressions more quickly than non-brief ones. This confirms that our model should prefer brief to non-brief NPs. There was a tendency for participants to comprehend predictable expressions more quickly than unpredictable ones. This suggests that a generation model which preferred predictable to unpredictable expressions might make gains in speed as well as reducing the chance of confusion.

If both reading and understanding are addressed, this raises the question of how these two dimensions should be traded off against each other. If one algorithm's output was read more quickly than that of another, but understood more slowly, which of the two should be preferred? Perhaps there is a legitimate role here for meta-linguistic judgments after all, in which participants are asked to express their preference between expressions (see Paraboni, Masthoff, and van Deemter (2006) for discussion). An alternative point of view is that these questions are impossible to answer independent of a realistic setting in

which participants utter sentences with a concrete communicative purpose in mind, which would allow task-based evaluation.

An algorithm for generating optimal coordinated NPs

We have formulated and implemented an algorithm whose aim is to generate optimal coordinated noun phrases in each situation, and which is based on the hypotheses tested in our experiments. The algorithm has been described in more detail in Khan, van Deemter, and Ritchie (2008), but we summarize it here.

Following van Deemter and Krahmer (2006) and Gatt (2007), we start by generating formulas in Disjunctive Normal Form (DNF). DNFs are set-denoting formulas whose overall structure is that of a set union (or, equivalently, logical disjunction). The algorithm, which builds on Gatt (2007) and is henceforth referred to as GAP, uses a divide-and-conquer approach to break the intended referent set into smaller components (subsets) and builds a DD for each such component using a conjunction of properties, using the incremental strategy of Dale and Reiter (1995). These DDs are then grouped together to form a single description that uniquely identifies the *whole* intended referent set. Our DNF formulas are like conventional set theoretic expressions, except that their building blocks are English *words* rather than names of sets. For example, we use formulas such as $man \sqcup (big \sqcap dog)$ to denote the set of domain objects that are either men or big dogs. The algorithm consists of five stages, as exemplified below. The first stage (1) is *Construction of Initial Description.* During this stage, we use GAP to build an initial description in DNF which denotes the intended referent set. The second stage is the *Transformation* stage (2). Here, the initial DNF formula is transformed in various ways, to produce a variety of set-denoting formulas, each of which is logically equivalent to the initial formula, but using a different logical structure. (Transformations could have been applied later, but we chose to apply them at the logical level, where the necessary structural manipulations are easiest

to perform.) The following transformation rules are used:

a. $((A \sqcap N_1) \sqcup (A \sqcap N_2)) \Rightarrow (A \sqcap (N_1 \sqcup N_2))$

b. $(X \sqcup Y) \Rightarrow (Y \sqcup X)$

These rules apply as often as possible, producing new formulas. The third stage of the algorithm (3) is the linguistic *Realization* stage. During this step, the formulas produced by the transformation stage are converted into strings of words. An example of a realization rule is

$((Adj \sqcap Noun_1) \sqcup (Adj \sqcap Noun_2)) \rightarrow$ *the Adj* $Noun_1$ *and the Adj* $Noun_2$. The Realization stage produces a set of noun phrases (one for each valid sequence of Rule applications), each of which could be used to refer to the target set. The problem now is to select the *best* noun phrase, which is done during the last two stages, where the results of our experiments will be utilized.

First, the algorithm enters a *Clarity Assessment* stage (4). This stage makes use of the Prediction Rules (Experiment 1), which stated that WMF $\rightarrow$ WS, and SMF$'$ $\rightarrow$ NS.

Finally, the algorithm enters its *Selection* stage (5). In accordance with our experiments, the algorithm prefers clear NPs over unclear ones. If several NPs are clear then the choice between them is made on the basis of brevity.

Example. The following example illustrates the working of the algorithm, letting R abbreviate the word *radical*, S *student*, and T *teacher*:

1. Construction of Initial Description: For this illustration, suppose that the output of description building is the DNF formula (a) $(R \sqcap S) \sqcup (R \sqcap T)$.

2. Transformation: The transformation rules generate three additional formulas: (b) $(R \sqcap T) \sqcup (R \sqcap S)$, (c) $R \sqcap (S \sqcup T)$, (d) $R \sqcap (T \sqcup S)$.

3. Realization of these formulas results in the following noun phrases: (a) *The radical students and the radical teachers*, (b) *The radical teachers and the radical students*,

(c) *The radical students and teachers*, and (d) *The radical teachers and students*.

4. Clarity Assessment: Each of these linguistic realizations is tested for clarity. The Prediction Rules predict wide scope for (d), because the relation between *radical* and *teachers* is one of WMF (according to the BNC data), hence (d) is clear (because the original DD was a wide-scope interpretation for this phrase). They do not predict a scope for (c) (because the relation between *radical* and *students* is neither WMF nor SMF′), hence (c) is unclear. The other two noun phrases, (a) and (b), are clear (because they are unambiguous), but less brief than (d).

5. Selection: Based on these assessments, the noun phrase (d) (clear and brief) is selected as the winner, and generated.

Conclusion and Future Work

This paper has described experiments with human participants, the outcomes of which informed the design and implementation of an algorithm for the generation of referring expressions. In a nutshell, the algorithm seeks to optimize generation output in light of possible surface structure ambiguities (as in the noun phrase "old men and women"). This work opens several avenues for future research. For example, our approach might help NLG systems handle other surface ambiguities, for instance involving prepositional phrase (PP) attachment, which have a similar "conjunctive" aspect as the coordinations that we have studied. That would depend on whether the likelihood (for readers) of different attachment possibilities could be predicted from corpus statistics.

Since our model makes predictions about the most likely interpretation of phrases (essentially imposes a classification on the set of phrases), it would be natural to seek an assessment of the accuracy of these predictions. It might seem that it would be possible to extract this data from Experiment 3, but a limitation[4] in the materials used means that this is not feasible. A new experiment would have to be designed to investigate this issue.

The tasks carried out by the participants in our studies are, of course, rather artificial, as we were attempting to control the variables involved. It would be interesting to measure experimentally the extent to which our model's predictions are an accurate reflection of a human reader's choices in a more realistic situation than Experiment 1. Also, we realize that contextual factors are likely to affect both interpretation and generation. It would therefore be interesting to explore the effect of preceding context upon the interpretation of NPs occurring later in that same text, since context could conceivably overwhelm the generic likelihoods based on the Word Sketches. Consider a two-sentence discourse: "There were old men and young women in the room. The old men and women were mourning." Our rules predict wide scope for the adjective "old", but the preceding context suggests otherwise.

Even though the decisions implemented in our generation algorithm are based on extensive experiments, this does not mean that these decisions are *always* the right ones. The situations where readers' preferences were examined in Experiment 2, for example, did not emphasize the risks that might be associated if incorrect interpretations were to arise. In highly fault-critical situations (e.g. Cushing (1994)) it seems likely that the likelihood of misunderstandings needs to be always minimized, and never traded off against brevity or fluency. The use of ambiguity-avoiding *controlled language* (e.g. Danlos, Jussieu, Lapalme, and Lux (2000)) would seem to be preferable here. The benefits of a corpus-based approach, advocated in the present article, would therefore seem to be largely absent if the algorithms were to be applied in such situations. Our experiments suggest, however, that in normal (i.e., less critical) situations, a little bit of ambiguity is not lethal.

References

Abney, S. (1996). Statistical methods and linguistics. In J. Klavans & P. Resnik (Eds.), *The balancing act: Combining symbolic and statistical approaches to language* (pp. 1–26). Cambridge, Massachusetts: The MIT Press.

Angwin, A., Chenery, H., Copland, D., Murdoch, B., & Silburn, P. (2006). Self-paced reading and sentence comprehension in Parkinson's disease. *Journal of Neurolinguistics*, *19*(3), 239–252.

Belz, A., & Gatt, A. (2007). Attribute selection for referring expression generation: New algorithms and evaluation methods. In *Proceedings of the 2nd UCNLG workshop: Language generation and machine translation (UCNLG + MT)* (pp. 75–83).

Binot, J.-I., & Jensen, K. (1987). A semantic expert using an online standard dictionary. In *Proceedings of international joint conference on artificial intelligence (IJCAI-87)* (pp. 709–714).

Chantree, F. (2006). *Identifying nocuous ambiguity in natural language requirements.* Unpublished doctoral dissertation, The Open Univeristy, Milton Keynes, England.

Chantree, F., Kilgarriff, A., De Roeck, A., & Willis, A. (2005). Disambiguating coordinations using word distribution information. In *Proceedings of the recent advances in natural language processing.* Borovets, Bulgaria.

Cushing, S. (1994). *Fatal words: Communication clashes and aircraft crashes.* Chicago: University of Chicago Press.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *18*, 233–263.

Danlos, L., Jussieu, P., Lapalme, G., & Lux, V. (2000). Generating a controlled language. In *Proceedings of the first international conference on natural language generation.* Mitzpe Ramon, Israel.

Dubinsky, S., Egan, M., Schmauder, A. R., & Traxler, M. J. (2000). Functional

projections of predicates: experimental evidence from coordinate structure processing. *Syntax*, *3*(3), 182–214.

Ferreira, F., Ferraro, V., & Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*, 11–15.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233.

Gardent, C. (2002). Generating minimal definite descriptions. In *Proceedings of the 40th annual meeting of the* ACL. Philadelphia, USA.

Gatt, A. (2007). *Generating coherent references to multiple entities.* Unpublished doctoral dissertation, University of Aberdeen, Aberdeen, Scotland.

Gatt, A., Belz, A., & Kow, E. (2008). The TUNA challenge 2008: Overview and evaluation results. In *Proceedings of the 5th international conference on natural language generation.*

Gatt, A., Belz, A., & Kow, E. (2009). The TUNA-REG challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European workshop on natural language generation* (pp. 174–182). Athens, Greece.

Horacek, H. (2004). On referring to sets of objects naturally. In *Proceedings of the 3rd international conference on* NLG (pp. 70–79). UK.

Inui, K., Tokunaga, T., & Tanaka, H. (1992). Text revision: A model and its implementation. In *Proceedings of the 6th international workshop on* NLG (pp. 215–230). Berlin, Heidelberg.

Khan, I. H., van Deemter, K., & Ritchie, G. (2008). Generation of referring expressions: Managing structural ambiguities. In *Proceedings of the 22nd international conference on computational linguistics (*COLING*-8)* (pp. 433–440). Manchester.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of EURALEX* (pp. 105–116). Lorient, France.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, Fog count, and Flesch reading ease formula) for navy enlisted personnel. *Navy Training Command Research Branch Report*, 8–75.

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In K. van Deemter & R. Kibble (Eds.), *Information sharing: Reference and presupposition in language generation and interpretation* (pp. 223–264). Stanford, CA: CSLI Publications.

Merlo, P., & Stevenson, S. (1999). Language acquisition and ambiguity resolution: The role of frequency distributions. In *Proceedings of the 21st annual conference of the cognitive science society* (pp. 399–404).

Neumann, G. (1994). *A uniform computational model for natural language parsing and generation.* Unpublished doctoral dissertation, University of the Saarland, Saarland, Saarbrücken.

Paraboni, I., Masthoff, J., & van Deemter, K. (2006). Overspecified reference in hierarchical domain: measuring the benefits for readers. In *Proceedings of the fourth international natural language generation conference* (pp. 55–62). Sydney, Australia.

Ratcliff, R. (1993). Methods for dealing with reaction-time outliers. *Psychological Bulletin*, *114*, 510–532.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems.* Cambridge, UK: Cambridge University Press.

Siddharthan, A., & Copestake, A. (2004). Generating referring expressions in open domains. In *Proceedings of the 42nd annual meeting of the* ACL. Barcelona, Spain.

Stone, M. (2000). On identifying sets. In *Proceedings of the 1st* INLG *conference* (pp.

116–123). Mitzpe Ramon.

Stone, M., & Webber, B. (1998). Textual economy through close coupling of syntax and semantics. In *Proceedings of the 9th international workshop on* NLG (pp. 178–187). New Brunswick, New Jersey.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216.

Traxler, M., Pickering, M., & Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(35), 558–592.

Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, *35*, 566–585.

van Deemter, K. (2002). Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, *28*(1), 37–52.

van Deemter, K. (2004). Towards a probabilistic version of bidirectional OT syntax and semantics. *Journal of Semantics*, *21*(3), 251–281.

van Deemter, K., & Krahmer, E. (2006). Graphs and Booleans: On the generation of referring expressions. In H.Bunt & R.Muskens (Eds.), *Computing meaning, vol. iii, studies in linguistics and philosophy.* Dordrecht: Kluwer.

van Gompel, R. P. G., Pickering, M. J., B, J. P., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory & Language*, *52*, 284–307.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. In O. Orgun & P. Sells (Eds.), *Morphology and the Web of grammar: Essays in memory of Steven G. Lapointe.* Stanford, CA: CSLI Publications.

Willis, A., Chantree, F., & De Roeck, A. (2008). Automatic identification of nocuous ambiguity. *Research on Language and Computation*, *6*(4), 355–374.

Wu, H., & Furugori, T. (1998). A computational method for resolving ambiguities in coordinate structures. In *Proceedings of PACLIC-12* (pp. 263–270). National University of Singapore.

Author Note

This work is supported by a University of Aberdeen Sixth Century Studentship, and EPSRC grant EP/E011764/1. We thank Albert Gatt, Alexandra A. Cleland and the reviewers for their valuable comments.

Footnotes

[1]http://www.natcorp.ox.ac.uk/

[2]More details of the materials used in the experiments reported in this paper can be found at http://www.csd.abdn.ac.uk/∼kvdeemte/Experimental-Materials.pdf.

[3]In all the experiments reported in this paper, $p < 0.05$ is our standard for statistical significance.

[4]In some cases both wide- and narrow-scope readings were plausible.

Table 1

Response proportions: Experiment 1

	Force	**Predicted Reading**	**Participants' Judgement**	**p-value**
Hypothesis 1	SCF & SMF	NS	NS (51/60)	< 0.001
Hypothesis 2	SCF & WMF	WS	WS (55/60)	< 0.001
Hypothesis 3	WCF & SMF	NS	NS (46/60)	< 0.001
Hypothesis 4	WCF & WMF	WS	WS (54/60)	< 0.001

Table 2

Possible phrasings for wide- and narrow-scope meaning

	Phrasings for wide-scope meaning	**Phrasings for narrow-scope meaning**
1.	*the Adj Noun*$_1$ *and the Adj Noun*$_2$	*the Adj Noun*$_1$ *and the Noun*$_2$
2.	*the Adj Noun*$_2$ *and the Adj Noun*$_1$	*the Noun*$_2$ *and the Adj Noun*$_1$
3.	*the Adj Noun*$_1$ *and Noun*$_2$	*the Adj Noun*$_1$ *and Noun*$_2$
4.	*the Adj Noun*$_2$ *and Noun*$_1$	*the Noun*$_2$ *and Adj Noun*$_1$

Table 3

Mean Reaction Time (ms)

Expression Type	**Reading Time (a)**	**Comprehension Time (b)**	**Processing Time (c = a + b)**
Predictable (brief)	2919.07	3045.01	5964.08
Non-brief	3421.63	2940.61	6362.24
Unpredictable	3014.23	3616.15	6630.38

Figure Captions

Figure 1. Interpreting an NP in a referential domain, using Euler diagrams

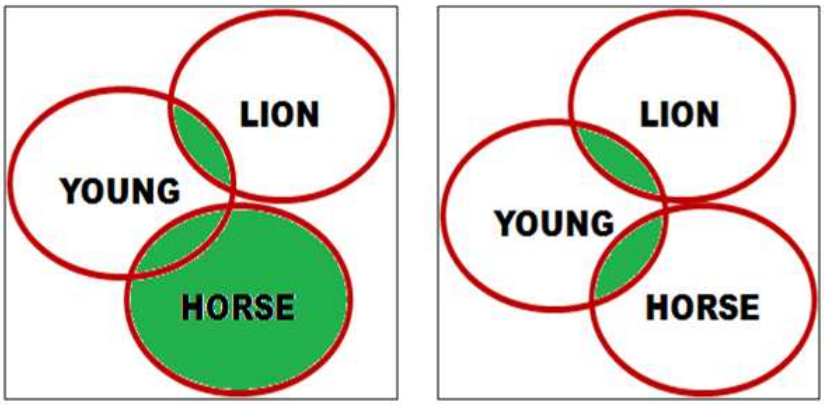
Figure 2. Clear versus unclear NP

Figure 3. Sample Trials: Choosing the best NP (the intended reading is narrow scope on the left, and wide scope on the right)

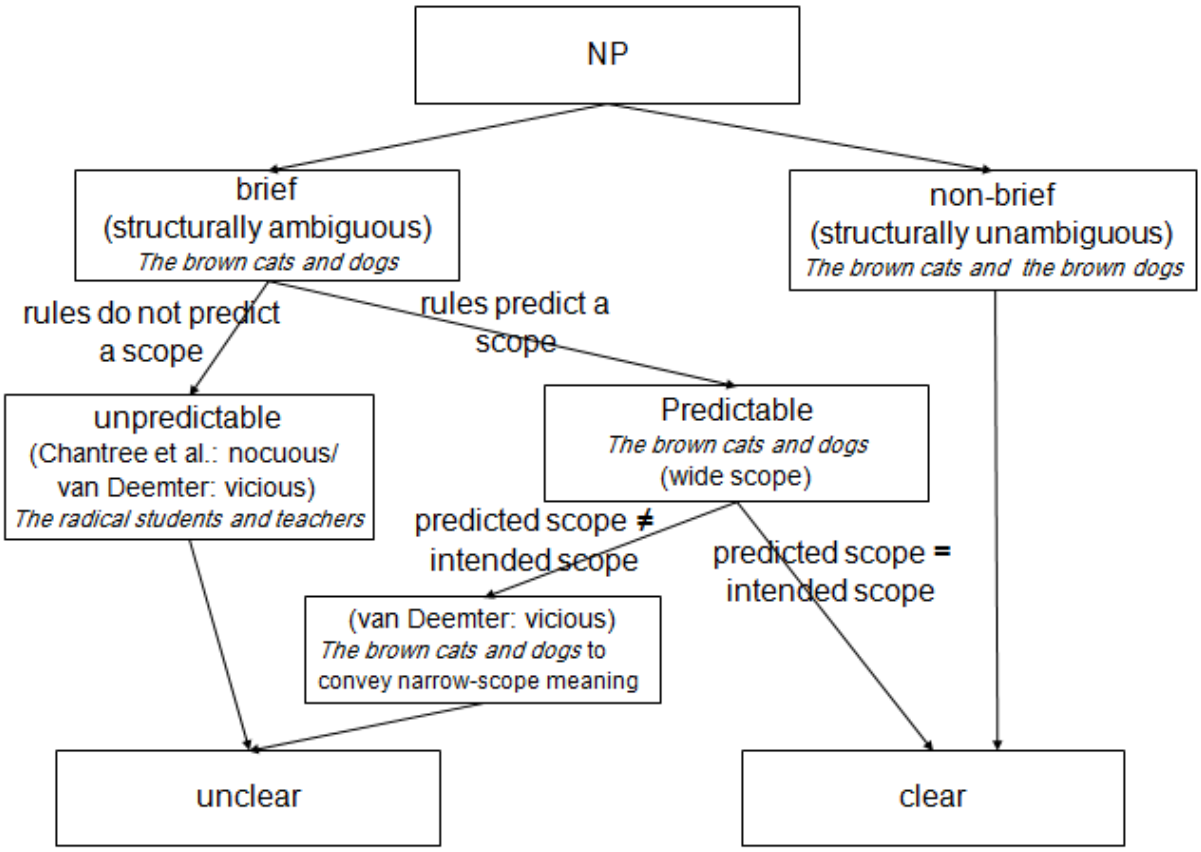
HORSE: Set of all animals that are horses

LION: Set of all animals that are lions

YOUNG: Set of all young animals



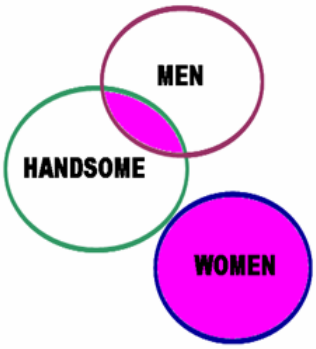
Please, remove the figure containing the young lions and horses.


NP
brief
(structurally ambiguous)
The brown cats and dogs
non-brief
(structurally unambiguous)
The brown cats and the brown dogs
rules do not predict a scope
rules predict a scope
unpredictable
(Chantree et al.: nocuous/
van Deemter: vicious)
The radical students and teachers
Predictable
The brown cats and dogs
(wide scope)
predicted scope ≠ intended scope
predicted scope = intended scope
(van Deemter: vicious)
The brown cats and dogs to convey narrow-scope meaning
unclear
clear

MEN = Set of all people who are men

WOMEN = Set of all people who are women

HANDSOME = Set of all people who are handsome



Which phrase works best to identify the filled area?

1. **The handsome men and women**
2. **The handsome men and the women**

MEN = Set of all people who are men

WOMEN = Set of all people who are women

HANDSOME = Set of all people who are handsome

Which phrase works best to identify the filled area?

1. **The handsome men and women**
2. **The handsome men and the handsome women**