Lexical Choice and Conceptual Perspective in the Generation of Plural Referring Expressions

Albert Gatt and Kees van Deemter
University of Aberdeen

3 May, 2007

Abstract. A fundamental part of the process of referring to an entity is to *categorise* it (for instance, as *the woman*). Where multiple categorisations exist, this implicitly involves the adoption of a conceptual perspective. A challenge for the automatic Generation of Referring Expressions is to identify a *set* of referents *coherently*, adopting the same conceptual perspective. We describe and evaluate an algorithm to achieve this. The design of the algorithm is motivated by the results of psycholinguistic experiments.

Keywords: natural language generation, generation of referring expressions, plurals, conceptual coherence, semantic similarity

1. Introduction

Generation of Referring Expressions (GRE) is an area of research that has attracted considerable attention from semanticists, psycholinguists and computationalists. In this paper, we take a *computational* stance, asking what the most appropriate referring expression is in a given situation, and presenting an algorithm, which is motivated and evaluated through extensive experiments with human subjects, that seeks to produce such expressions. We begin by introducing some fundamental assumptions in computational GRE.

Most GRE algorithms seek a felicitous linguistic description that distinguishes an intended referent from its distractors (i.e., all other entities) in a Knowledge Base (KB), whose content is assumed to be shared between speaker and hearer. Much of the GRE literature has focused on the semantic heart of the problem, known as Content Determination (CD), which involves finding a set of properties that later modules can 'translate' into an appropriate string of words. CD algorithms are typically informed by some interpretation of the Gricean maxims (Dale and Reiter, 1995), especially various versions of the Maxim of Brevity, which has been interpreted in this context as a constraint on saying no more than is absolutely required to distinguish an intended referent (Dale, 1989; Gardent, 2002).

Recent work in GRE has proposed various generalisations of the original problem. One significant development has been the study of

Table I. Example domain

	GENDER	OCCUPATION	NATIONALITY
e_1	man	postgraduate	Maltese
e_2	man	undergraduate	Greek
e_3	man	chef	Italian

reference to sets (van Deemter, 2000; Stone, 2000; van Deemter, 2002; Gardent, 2002; Horacek, 2004). After an initial focus on logical and algorithmic aspects of this problem, the main questions are now arguably of an empirical kind. Given a certain shared Knowledge Base, a set can usually be referred to in many different ways. *Which of these many descriptions are linguistically (most) acceptable?*

A fundamental part of the process of referring to an entity is to *categorise* it (for instance, as *the woman*). Where multiple categorisations exist, this implicitly involves the adoption of a conceptual perspective. The listener can infer the perspective adopted from a speaker's lexical choice (e.g. Clark, 1997). A challenge for the automatic generation of plural references is that of *conceptual coherence*, whereby a set is covered using related properties to categorise its elements. To explain our initial intuition, consider a reference to $\{e_1, e_3\}$ in Table I, a KB where entities are specified for values of three attributes. Suppose we use the Incremental Algorithm (IA) (Dale and Reiter, 1995), often considered the gold standard in the area. IA searches along an ordered list of attributes, selecting properties of the intended referents that remove some distractors. Assuming the ordering in the top row of the table, IA would yield *the postgraduate and the chef*[1], which is fine in case OCCUPATION is the *relevant* attribute in the discourse, but otherwise is arguably worse than an alternative like *the Italian and the Maltese*, because it is more difficult to see what a postgraduate and a chef have in common. What this example suggests is that the conceptual relatedness of elements of a set affects the felicitousness of a description, leading us to hypothesise the following constraint, which is vaguely worded for the time being:

> **Conceptual Coherence Constraint** (CC): As far as possible, conceptualise elements of a plurality in similar ways.

Related issues have been raised in the formal semantics literature. Building on a long tradition of work on the semantics of questions,

[1] The property *man* is disregarded because it is true of all entities and fails to remove any distractors.

Aloni (2002) argues that an appropriate answer to a question of the form *'Wh x?'* must conceptualise the different instantiations of x using a perspective which is relevant given the hearer's information state and the context. Kronfeld (1989) distinguishes a description's *functional relevance* – i.e. its success in distinguishing a referent – from its *conversational relevance*, which arises in part from implicatures. In our example, describing e_1 as *the postgraduate* carries the implicature that the entity's academic role is relevant. When two entities are described using dissimilar properties, say *the student and the Italian*, the contrast may be misleading for the listener. The hypothesised Conceptual Coherence Constraint therefore has some pragmatic implications. A description of a set may trigger an inference process in the listener, whose assumption is that the properties ascribed to the referents, and which constitute a cover of the intended set, have some value in 'binding' these elements together to form a coherent plurality.

Useful though these insights are, it seems clear that they can only be rules of thumb. An algorithm that took Aloni's position literally, for example, would fail to be *logically complete*, because it would sometimes fail to find a distinguishing description even though one exists (van Deemter, 2002). This would happen if the elements of a set could not be distinguished from the same conceptual perspective, given the state of the KB. In other words, although a GRE algorithm should attempt to find a coherent description, it should not fail in the absence of one.

We will show that the CC can be explained and modelled in terms of lexical semantic forces within a description, a claim supported by the results of three experiments. Our starting point was the observation, frequently made in psycholinguistic work, that plural references in discourse are easier for listeners to resolve when they are 'similar'. Similarity has been operationalised in different ways as, for example, the ontological homogeneity of a set of referents, or the extent to which they share properties in a discourse (Eschenbach et al., 1989; Sanford and Moxey, 1995; Kaup et al., 2002; Koh and Clifton, 2002). These experiments have often focused on anaphoric plurals (prototypically realised as pronouns). Our work expands these results in two directions, by seeking a precise notion of 'similarity' which can explain the initial intuitions, and by focusing not on pronouns, but on definite plural references.

Though our focus on 'low-level', lexical, determinants of adequacy in our experiments constitutes a departure from the standard Gricean view in GRE, our position need not be considered opposed to more pragmatically-oriented accounts. Rather, the pragmatic 'effect' that is incorporated in the CC could be an emergent property of a plural description whose basis lies in a speaker's lexical choice.

The second aim of this paper is to describe an algorithm motivated by the experimental findings (§3) which seeks to find the most coherent description available in a domain according to CC; this algorithm is evaluated using an experiment with human subjects, where the model that it incorporates is explicitly compared to the predictions of the more standard, brevity-oriented view that has dominated computational GRE. The outcomes of this evaluation study are remarkable not only because they lend strong support to our hypotheses concerning coherence, but also because they unexpectedly fail to lend support to longstanding ideas about the importance of brevity in reference.

2. Empirical evidence

We take as paradigmatic the case where a plural reference involves disjunction/union, that is, has the logical form $\lambda x\,(p(x) \vee q(x))$, realised as a description of the form *the* N_1 *and the* N_2. By hypothesis, the case where all referents can be described using identical properties (logically, a conjunction), is a limiting case of CC.

Three experiments are reported below. The first, based on a phrasal judgment paradigm, aimed to make the relevant definition of similarity more precise. Two further experiments tested the Conceptual Coherence hypothesis in domains involving referent identification, bringing the experimental situation closer to that in which GRE algorithms operate.

2.1. Definitions of similarity

Intuitively, the similarity or relatedness of two words or concepts is a function of the things they have in common. Consider, for example, the two words *master* and *pupil*. To a native speaker, these two words might be perceived as quite related, perhaps because the entities they denote tend to co-occur in the same situations. More precisely, they are ways of conceptualising humans in terms of roles that have some relationship between them (both, for example, are roles associated with education). As a result of this, they might also tend to be spoken about in the same contexts, in relation to a number of things that they have in common.

If similarity were operationally defined with reference to an ontology or taxonomic hierarchy, then 'what two words (or concepts) have in common' could be defined in terms of the relationships that hold between them in that taxonomy. This view characterises our first definition of similarity, called ontological relatedness (OR), estimated

using the WordNet taxonomy. OR is characterised as the multiplicative inverse of the number of edges between concepts in the WordNet IS-A nominal hierarchy, normalised to account for possible zero values (Pederson et al., 2004). To continue with our example, the fifth nominal sense of *master* in WordNet 2.1 is *schoolmaster*, while the first sense of *pupil* is that of a *learner*. We would expect these two concepts to be highly related: both, for example, are hyponyms of *person* or *individual*.

This is arguably a partial story, insofar as the two words (resp. concepts) have more in common than this. For instance, both might be talked about in relation to *classroom* and *school*. The OR definition of similarity is compared in our first experiment to a distributional, corpus-based, measure (Distributional Similarity, abbreviated DS), based on the work of Lin (1998b). The similarity of two arbitrary objects a and b is a function of the information gained by giving a joint description of a and b in terms of what they have in common, compared to describing a and b separately. Applied to corpora, DS focuses on the grammatical relations in which two words occur (Lin, 1998a). Such relations are formalised as triples $\langle rel, w, w' \rangle$, where rel is a grammatical relation, w the word of interest and w' its co-argument in rel. For instance, some of the grammatical triples associated with *master*, obtained from the British National Corpus, are $\langle$*subject-of*, *master*, *attend*$\rangle$, and $\langle$*subject-of*, *master*, *write*$\rangle$. Both of these are also relations in which *pupil* is attested in the corpus. However, the two words will not be attested in these contexts to the same extent; nor will they always occur with the same co-arguments in the same contexts. For example $\langle$*modifies*, *strict*, *master*$\rangle$ might occur reasonably frequently, but the corresponding triple for *pupil* ('strict pupil') is presumably not so frequent.

To quantify the degree of association between a word and a co-argument in a grammatical relation, DS takes into account the mutual information of w and w' in that relation, expanding on previous work by Church and Hanks (1990), using the following equation:

$$I(rel, w, w') = \log \left(\frac{\|\langle rel, *, * \rangle\| \times \|\langle rel, w, w' \rangle\|}{\|\langle rel, w, * \rangle\| \times \|\langle rel, *, w' \rangle\|} \right) \quad (1)$$

where $\|\langle x, y, z \rangle\|$ is the frequency of the triple $\langle x, y, z \rangle$, and $*$ indicates any argument. The estimate of mutual information therefore takes into account (a) the overall frequency of the relation in question and (b) the overall frequency of w in that relation with w', scaling this by the frequency with which w and w' occur in that relation overall (Lin, 1998a). To estimate similarity between two words, we take into account their co-arguments in specific grammatical relations, weighted by their mutual information. Let $\sigma(w_1, w_2)$ denote the similarity estimate of two

words w_1 and w_2, and let $F(w)$ be the set of words and relations which, together with w form an attested grammatical triple. For example, $\langle$*subject-of*, *attend*$\rangle$ is an element of both F(*master*) and F(*student*). Lin's formula to estimate similarity is as follows:

$$\sigma(w_1, w_2) = \frac{2 \times I(F(w_1) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))} \quad (2)$$

Under this definition, the extent to which two words have common features is a function of the extent to which they are used in the same contexts, or talked about in the same way. Thus, DS emphasises language *use*.

Throughout the rest of this paper, the corpus-derived heuristics for estimating DS are obtained from SketchEngine[2] (Kilgarriff, 2003; Kilgarriff et al., 2004), which contains information about word similarity and the mutual information of grammatical triples, based on estimates from the British National Corpus (BNC)[3]. In the experiments reported below, similarity between pairs of nouns was estimated on the basis of the three grammatical relations of (a) *Subjecthood*, the likelihood of two nouns occurring as subjects of the same verb; (b) *Objecthood*, the likelihood of two nouns occurring as objects of the same verb; (c) *Modification*, the likelihood of two nouns being pre- or post-modified by the same adjectives.

We contrasted the DS measure in (2) to OR. A previous experiment (Gatt and van Deemter, 2005) found significant correlations between the similarity of a pair of nouns in a plural NP, and the extent to which human subjects perceived that plural NP as likely to be used in some situation. Correlations were highest for DS, as compared to a number of WordNet-based measures that combined the definition of OR given above with various corpus-derived heuristics. Experiment 1 sought to replicate this finding in an experimental design.

2.2. Experiment 1

To substantiate the Conceptual Coherence hypothesis, participants in the first experiment were asked to judge definite plural NPs, in terms of their *perceived likelihood of usage in some situation*. The similarity between head nouns in the NPs was the main factor manipulated.

2.2.1. *Method*

As in Gatt and van Deemter (2005), we used Magnitude Estimation (ME) (Stevens, 1957), a technique developed in psychophysics. In ME,

[2] http://www.sketchengine.co.uk
[3] http://www.natcorp.ox.ac.uk

participants are asked to rate physical stimuli (e.g. loudness) by assigning them a number on a scale of their own choice. All stimuli are compared to an initially judged *modulus* item. Scores are normalised to enable comparison across participants. Taking m to be the rating assigned by a person to the initial modulus item, and t to be the rating assigned to a subsequent stimulus, the normalised score t_n is calculated as follows:

$$t_n = \log\left(\frac{t}{m}\right) \tag{3}$$

This method has been applied successfully to linguistic judgments, ranging from ratings of the grammaticality of different sentences (Bard et al., 1996; Keller, 2003), to the acceptability of adjective-noun combinations (Lapata et al., 1999).

If the relationship between a subjective judgment of a stimulus and its real magnitude is systematic, the plot of real magnitudes against subjective judgments in log-log coordinates should fall on a straight line (that is, a regression line should have $R^2 \approx 1$, where R^2 is the proportion of variance in the data that the regression equation covers). Given the lack of an objective measure against which to compare subjective magnitudes in the present experiments, we used *Cross-Modality Matching*, a variant of the ME task in which participants are asked to rate items in two completely different modalities. If participants are self-consistent in their judgments, normalised scores for one modality, regressed on the normalised scores for the other, should fall on a straight line with R^2 approaching 1, indicating that the task has some psychological validity.

In the experiment, participants rated items using a numeric scale, and a visual method, which involved moving a slider from left to right. The slider position returned a real value in $(1, 100)$, though participants were not aware of this. It was made clear to participants that how far to the right they moved the slider would reflect how much they thought a phrase was likely to be used. As is standard in the ME paradigm, participants went through a calibration phase prior to the experiment, in which they were introduced to the concept of proportion, and practiced using sliders and numeric scales to express it.

2.2.2. *Materials and design*

Twelve pairs of nouns were manually selected from word lists generated from the BNC. From each pair, a description of the form *the* N_1 *and the* N_2 was constructed. The materials represented all combinations of the following within-subjects factors:

Table II. Materials used in Experiment 1

DS	OR	**Example**
high	high	the leader and the chairman
high	low	the manager and the council
low	high	the department and the resource
low	low	the garden and the police

1. *Frequency* (FR; 3 levels): Noun pairs were matched for frequency, which was either *High* ($f \geq 500$ per million), *Medium* ($500 > f \geq 300$ per million) or *Low* ($f \leq 100$ per million).

2. *Distributional Similarity* (DS; 2 levels): A pair of nouns n_1 and n_2 in a disjunctive description had *High* DS if $\sigma(n_1, n_2) \geq 0.2$. The pair had *Low* DS if $\sigma(n_1, n_2) \leq 0.05$.

3. *Ontological Relatedness* (OR; 2 levels): *High* OR meant that the multiplicative inverse of the shortest path length between (the most highly related senses of) n_1 and n_2 was greater than or equal to 0.3. *Low* OR was defined as a minimum path value less than 0.01.

Some example phrases are shown in Table II. As the examples show, it was possible to find pairs of words, such as *manager* and *council*, which had a high DS value, but did not have a high OR value. Nouns such as these belong to different ontological categories according to WordNet, whose IS-A taxonomy does not have a common root. For example, while *manager* is subsumed by *person*, the three WordNet senses of *council* are hyponyms of *administrative unit*, *assembly* or *meeting*, all of which have *social group* as their least common subsumer. The high DS value of these nouns is due to their tendency to occur in several similar contexts (e.g. both are modified by *senior*, *general*, *technical*, and so on). In addition, logical metonymy is frequently found with group nouns such as *council*, so that the word stands in for its members in the context of a sentence.

2.2.3. *Participants and procedure*
27 self-reported native or fluent speakers of English did the experiment on the web. They first rated a modulus item, itself a definite plural NP, using both the numeric and slider modalities, then were exposed to each of the 12 trials in random order, in two modalities (displayed at different points). Thus, each participant made 24 ratings, judging each

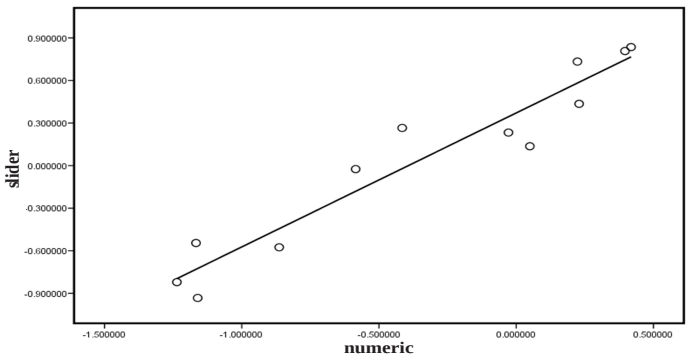

Figure 1. Cross-modality plot (log-log coordinates) for Experiment 1

phrase in relation to the modulus, whose original rating was shown in the relevant modality.

2.2.4. *Results and Discussion*

Figure 1 displays the regression plot of mean numeric and slider magnitudes for each trial in log-log coordinates. It indicates near-perfect self-consistency in ratings across modalities ($R^2 = .92$, $\beta = 0.96$, $p < 0.001$), suggesting that the task was valid, and made sense to individuals.

A 3(FR)×2(DS)×2(OR) ANOVA was conducted on the normalised ratings, using both participants (F_1) and items (F_2) as sources of variance. A significant main effect of DS was observed ($F_1(1, 26) = 47.909, p < 0.001$, $F_2(2, 11) = 53.505, p < 0.001$). The main effect of FR was also significant ($F_1(2, 26) = 16.083, p < 0.001$; $F_2(2, 11) = 7.272$, $p < .001$). No reliable main effect of OR was obtained ($F_1(1, 26) = 2.617, ns.$, $F_2(2, 11) = 1.081, p > 0.6$), but there was a reliable interaction of this variable with FR ($F_1(2, 26) = 9.414, p = .001$; $F_2(2, 11) = 3.472$, $p = .03$). The overall interaction of the three factors was also significant, but only by subjects ($F_1(2, 26) = 6.145$, $p = .004$; $F_2(2, 11) = 2.278$, ns). No other interactions were significant.

Post-hoc Tukey's comparisons of different levels of FR showed that the main effect was due exclusively to a difference between *High* and *Low* frequency levels (Tukey's $HSD = 3.558$, $p < .05$). This also helps to explain the FR × OR interaction. While *High* DS items were rated more highly at all levels of FR, *High* OR items were only rated as more likely to be used when FR was very high or very low.

The main effect of DS shows that people's judgment of plural noun phrases is strongly determined by the extent to which the nominal constituents of those phrases tend to be used in the same linguistic

Table III. Conditions in Experiment 2

	a	b	c	distractor
High DS	spanner	chisel	plug	thimble
Low DS	toothbrush	knife	ashtray	clock

contexts, supporting earlier results (Gatt and van Deemter, 2005). The lack of a main effect of Ontological Relatedness, and the lack of an interaction between OR and DS, is surprising, given previous psycholinguistic work which showed that ontologically homogeneous nouns tended to increase the likelihood of plural reference, and reduce the processing effort in reading (Koh and Clifton, 2002). The measure of distributional similarity used here will reflect ontological similarity to the extent that ontologically homogeneous entities are talked about in the same context. However, word pairs which belong to different ontological categories were still judged as perfectly likely to be used, if they had high distributional similarity. This may have been a result of OR having been defined in terms of the Minimum Path measure, which our earlier experiment found to have a significant, though lower correlation to people's judgments. Moreover, ontologically unrelated words were often pairs consisting of an animate, human noun and a group noun that permitted an interpretation, via logical metonymy, that made it compatible with a 'human agent' reading, possibly overriding an effect of ontological heterogeneity.

2.3. Experiment 2

In Experiment 2, participants were placed in a situation where they were buying objects from an online store. They saw scenarios containing four pictures of objects, three of which (the targets) were identically priced. Participants referred to them by completing a 2-sentence discourse:

S1 The *object 1* and the *object 2* cost *amount*.

S2 The *object 3* also costs *amount*.

If similarity is a constraint on referential coherence, then the participants should prefer a plural reference in S1 if *object 1* and *object 2* can be categorised using similar nouns.

2.3.1. *Materials and design*
All the pictures were artefacts selected from a set of drawings normed in a picture-naming task with British English speakers (Barry et al., 1997). Of the three targets (a, b, c), c was always an object whose name in the norms was *dis*similar to that of a and b. The semantic similarity of (nouns denoting) a and b was manipulated using the same definitions of *High* and *Low* Distributional similarity (DS) as in the previous experiment. Examples of the objects used in the conditions are shown in Table III.

Visual Similarity (VS) of a and b was also manipulated, to avoid a possible bias for referring to a set of two items that are similar in appearance. Pairs of pictures were first normed with a group who rated them on a 10-point scale based on their visual properties. *High* VS (HVS) pairs had a mean rating ≥ 6; Low VS (LVS) pairs had mean ratings ≤ 2. Two sets of materials were constructed, for a total of 2 (DS) $\times$ 2 (VS) $\times$ 2 = 8 trials.

2.3.2. *Participants and procedure*
29 self-reported native or fluent speakers of English completed the experiment over the web. Trials were displayed on a screen displaying the four domain objects in a 2×2 array. The two targets on which DS and VS were manipulated, a and b, were never adjacent. Participants completed each discourse by clicking on the picture which they wanted to refer to in the next available sentence slot. They had the option of resetting the display and changing their references.

2.3.3. *Results and discussion*
Responses were coded according to whether objects a and b were referred to in the plural subject of S1 ($a + b$ responses) or not ($a - b$ responses). If our hypothesis is correct, there should be a higher proportion of $a + b$ responses in the HDS condition. In an initial analysis, the visual similarity of pictures turned out to play no role at all in people's selection of content. Thus, from different Visual Similarity conditions are combined in what follows. Analysis is carried out on response proportions using pairwise Signed Rank Tests by participants (Z_1) and items (Z_2). We also report an initial χ^2 test on response frequencies.

Participants referred to the designated targets 72% of the time in the High DS condition, compared to 20.2% in the Low DS condition. The difference in response frequencies across the two conditions was highly significant ($\chi^2 = 41.371$, $p < .001$). By participants, the proportion of $a + b$ responses was reliably higher in the High DS, compared to the Low DS condition ($Z_1 = 4.313$, $p < .001$), though it only approached

significance by items ($Z_2 = 1.826$, $p = .06$). The same pattern was observed in comparing proportions of $a - b$ responses in the two conditions, with a significantly greater proportion of these in the Low DS condition ($Z_1 = 4.411$, $p < .001$; $Z_2 = 1.826$, $p = .06$).

Given the choice, participants prefer to describe similar entities in a plural description. Although the results showed that people referred to dissimilar entities roughly 30% of the time in the first sentence of a discourse overall, the trend is clearly and reliably in the predicted direction, with more references to the designated targets when they were similar[4].

The main conclusion that can be drawn from this experiment is that participants show a strong preference for entities with similar types or head nouns in plurals. This is predicted by the Conceptual Coherence Hypothesis, and suggests that distributional similarity at the lexical level is playing a role in determining people's choices. What the experiment does not address is the question of Content Determination. At the outset of this chapter, some motivating examples were given of discourses and referential domains in which it was clear that entities could be referred to in different ways and that by hypothesis, reference to plurals would be constrained by the availability of similar properties. This aspect of the CC is perhaps the most crucial, since it has a direct bearing on the content determination strategy of a GRE algorithm that seeks to satisfy it. Experiment 3 addressed this hypothesis directly.

2.4. Experiment 3

Experiment 3 was a sentence continuation task, designed to approximate content determination in GRE. Participants saw a series of discourses, whose function was to evoke a concrete domain of discourse involving three entities (e_1, e_2, e_3). Each of these was introduced as having two different distinguishing properties. The final sentence in each discourse had a missing plural subject NP referring to two of these. (The third entity played the role of a distractor, without which the reference task would have been uninteresting.) The context made it clear which of the three entities had to be referred to. Our hypothesis was that participants would prefer to use semantically similar properties for the plural reference.

2.4.1. *Materials and design*

Materials consisted of 16 discourses, such as those in Figure 2. After an initial introductory sentence, the three entities were introduced in separate sentences. In all discourses, the pairs $\{e_1, e_2\}$ and $\{e_2, e_3\}$

[4] The reliability of these results is strengthened by a two previous replications.

Three of the richest men in Europe were spotted last night dining at a London restaurant. All three are millionaires with a passion for fine arts and antiques.

(e_1) One of the men, a Rumanian, is a dealer$_i$.

(e_2) The second, a prince$_j$, is a collector$_i$.

(e_3) The third, a duke$_j$, is a bachelor.

Continuation:
The XXXXXXXXXX were both accompanied by servants, but the bachelor wasn't.

Figure 2. Example discourse in Experiment 2

could be described using either pairwise similar or dissimilar properties (similar pairs are co-indexed in the figure). This experiment consisted of two conditions, of which only one is reported here. In the relevant condition, based on 8 materials, the distinguishing properties of each entity were *nouns* (e.g. *duke*, *prince*, *bachelor*)[5] For counterbalancing, two versions of each discourse were constructed, such that, if $\{e_1, e_2\}$ was the target set in Version 1, then $\{e_2, e_3\}$ was the target in Version 2. Twelve filler items requiring singular reference in the continuation were also included. The order in which the entities were introduced was randomised across participants, as was the order of trials.

2.4.2. *Participants and procedure*
18 native speakers of English, from the Aberdeen NLG Group database of experimental participants, completed the experiment. Items were presented in random order. Participants completed all discourses, and were randomly assigned to Version 1 or 2 so that for any discourse, there were roughly equal numbers of participants who referred to two different pairs of entities.

2.4.3. *Results and discussion*
Errors, consisting of references to a non-target entity, were omitted from analysis. The other responses were categorised as follows:

1. *Similar*: These were plural responses in which the two target referents were correctly identified using the similar properties provided in the discourse. There were three sub-categories of this response type:

[5] The other condition, with a further 8 discourses, contained adjectival distinguishing properties. Detailed consideration of adjectives is beyond the scope of the present paper.

(a) *Disjunctive*: The plural reference consisted of a disjunctive NP with the two similar properties. E.g. *the duke and the prince* in Figure 2.

(b) *Superordinate*: The plural reference consisted of a superordinate term that subsumed the two similar properties. E.g. *the noblemen*, where *noblemen* subsumes *prince and duke*.

(c) *Include similar*: The two similar properties were used in a disjunctive NP, together with other properties. E.g. *the bachelor duke and the prince*.

2. *Dissimilar/other*: All other references were classified in this category.

Statistical results are reported comparing proportions of *Similar* responses overall (i.e. collapsing over response categories 1*a*–1*c*), to *Dissimilar* (2) responses, and also comparing the *disjunctive* (1*a*) responses to Dissimilar responses.

Overall, *Similar* responses accounted for 66% of plural descriptions in the Nominal condition. Proportions of *Similar* descriptions overall (category 1*a*–*c*) differed significantly from *Dissimilar* ($Z_1 = 2.719$, $p = .03$; $Z_2 = 1.997$, $p = .05$). Restricting attention only to those descriptions consisting of disjunctive NPs (1*a*) does not change the picture by participants ($Z_1 = 2.337$, $p = .01$), though the result is weaker by items ($Z_2 = 1.680$, $p = .09$).

The results support the hypothesis that a constraint on similarity in the categorisation of elements of a set is operative in content determination for plural references. Our story so far has focused on nouns, in line with the motivating arguments in §1, where the emphasis was on the way entities are categorised. The implicit assumption in these experiments has been that nouns typically represent *types*, which denote the conceptual category of an entity. What of non-type properties, such as *red* or *clever*? While it is premature to suggest that CC plays no role in modifier selection, it is likely that modifiers play a different role from nouns, namely to add information to an already-represented entity. Previous work has shown that restrictions on the plausibility of adjective-noun combinations exist, suggesting a dependency between the initial categorisation of an object, and what other properties can subsequently be predicated of it (Lapata et al., 1999). Unlikely combinations (e.g. *the immaculate kitchen* rather than *the spotless kitchen*) impact processing in online tasks (Murphy, 1984). A possible explanation, offered in some lexical semantic theories (Pustejovsky, 1995), is that nominals have structured lexical entries with slots or roles to which modifiers attach selectively, and the composition of

Table IV. An example knowledge base

	TYPE	OCCUPATION	SPECIALISATION	GIRTH
e_1	woman	professor	physicist	plump
e_2	woman	lecturer	geologist	thin
e_3	man	lecturer	biologist	plump
e_4	man		chemist	thin

noun-modifier combinations results in some aspect of the nominal semantics being foregrounded. Though the algorithm presented below is aimed primarily at coherence in categorisation/conceptualisation, it also makes use of mutual information values between nouns and adjectives to take modifier-noun combinations into account.

3. An algorithm for referring to sets

Our next task is to port the results to GRE. The main ingredient to achieve conceptual coherence will be the definition of semantic similarity. In what follows, all examples will be drawn from the domain in Table IV. We assume a distinction between *types*, that is, any property that can be realised as a noun; and *modifiers*, or non-types. Given a set of target referents $R \subseteq U$, where U is the set of domain entities, the algorithm described below generates a description D in Disjunctive Normal Form (DNF) with the following properties:

1. Any disjunct in D contains a 'type' property, i.e. a property realisable as a head noun.

2. If D has two or more disjuncts, each a conjunction containing at least one type, then the disjoined types should be as similar as possible, given the information in the KB and the *completeness* requirement: that the algorithm find a distinguishing description whenever one exists.

The algorithm achieves conceptual coherence by first grouping the available types in the KB into *conceptual perspectives*, sets of nouns with a high pairwise similarity. Perspectives are related to each other via a well-defined notion of semantic distance. The Content Determination procedure attempts to distinguish R by selecting lexical items from

the same perspective, failing which, it tries to minimise the semantic distance between the conceptual perspectives represented in the description.

3.1. Finding perspectives

Our algorithm makes use of the SketchEngine database as its primary knowledge source. In addition to the weighted grammatical triples used to estimate similarity (see equations 1 & 2), the database also contains a thesaurus, wherein a given word is accompanied by an ordered list of semantically similar words.

Since the definition of similarity applies to words, the first step is to generate all possible lexicalisations of the available attribute-value pairs in the domain. In practice, this is carried out using WordNet: for every value of an attribute, the set of lexicalisations is defined as the elements of its set of synonyms. We distinguish between type properties (the set T), and non-types or modifiers (M)[6]. The thesaurus is used to find the pairwise similarity of types in order to group them into related clusters. We also use information about grammatical triples to identify, for each type in the KB, those modifiers that can felicitously combine with it. This is based on the mutual information estimate (see equation 1) between a noun and a modifier in the relevant modification relation. For example, in Table IV, e_3 has *plump* as the value for GIRTH, which combines more felicitously with *man* than with *biologist*. Respecting type-modifier combinations in this way avoids the negative impact of counterintuitive combinations described by some authors (cf. §2.4).

We now make the notion of a perspective more precise. Let T be the set of types in the KB, and let $\sigma(t, t')$ be the (symmetrical) similarity between any two types t and t'. These determine a semantic space $\mathbb{S} = \langle T, \sigma \rangle$. We define the notion of a perspective as follows.

Definition 1. **Perspective**
A perspective $\mathcal{P}$ is a convex subset of $\mathbb{S}$, i.e.:
$\forall t, t', t'' \in T : ((t, t' \in \mathcal{P} \wedge \sigma(t, t'') \geq \sigma(t, t')) \rightarrow t'' \in \mathcal{P})$

Types are clustered using the nearest-neighbour search algorithm described in Gatt (2006), which takes as input a representation of the semantic space. For each type t, the algorithm finds its nearest semantic neighbour in $\mathbb{S}$. The procedure to merge such pairs into clusters involves taking the transitive closure of the nearest neighbour relation. Thus, if

[6] This is determined on the basis of the BNC corpus, as follows: if a word occurs as a noun then it is a member of T; if it occurs as another syntactic category then it is a member of M. Note that T and M need not be disjoint. Note also that entities can have more than one type property (see e.g. Table IV).

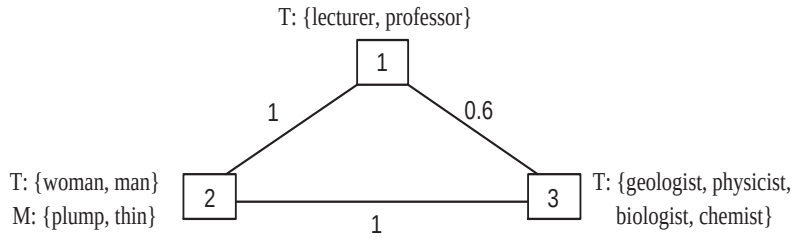

Figure 3. Perspective Graph

t is the nearest neighbour of t', and t' is the nearest neighbour of t'', then $\{t, t', t''\}$ is a cluster. Clearly, the resulting sets are convex in the sense of Definition 1.

Once types are clustered, modifiers are assigned to a cluster as follows. Let m be a modifier, and let $I(t, m)$ be the corpus-derived mutual information holding between m and any type (equation 1). The modifier is included in the cluster containing the type t such that $I(t, m) = \max_{t' \in T} I(t', m)$. Thus, a cluster is a pair $\langle \mathcal{P}, M_{\mathcal{P}} \rangle$ where $\mathcal{P}$ is a perspective (a set of types), and $M_{\mathcal{P}} \subseteq M$. The distance $\delta(A, B)$ between two clusters A and B is defined straightforwardly in terms of the semantic distance between the perspectives they contain, $\mathcal{P}_A$ and $\mathcal{P}_B$:

$$\delta(A, B) = \frac{1}{1 + \frac{\sum_{x \in \mathcal{P}_A, y \in \mathcal{P}_B} \sigma(x,y)}{|\mathcal{P}_A \times \mathcal{P}_B|}} \tag{4}$$

Finally, a weighted, connected graph $\mathcal{G} = \langle V, E, \delta \rangle$ is created, where V is the set of clusters, E a set of edges where each edge is weighted by the semantic distance between two perspectives. As an example, Figure 3 shows the graph constructed for the domain in Table IV[7].

We now define the coherence of a description more precisely. Given a description D, we shall say that a perspective $\mathcal{P}$ is *represented in* D if there is at least one type $t \in \mathcal{P}$ which is in D. Let $\mathbb{P}_D$ be the set of perspectives represented in D. Since $\mathcal{G}$ is connected, $\mathbb{P}_D$ determines a connected subgraph of $\mathcal{G}$. The *total weight* of D, $w(D)$, is the sum of weights of the edges in $\mathbb{P}_D$.

Definition 2. **Maximal coherence**
A description D is *maximally coherent* iff there is no description D' coextensive with D such that $w(D) > w(D')$.

[7] We simplify the presentation by showing only one lexicalisation per KB property.

3.2. Content determination

The Content Determination procedure takes as input a set of intended referents R and the perspective graph $\mathcal{G} = \langle V, E, \delta \rangle$. It traverses the nodes of the graph, searching within each node (cluster) in V and selecting a word w if (a) w is true of at least one referent and (b) w has *discriminatory value*, that is, there are some distractors of which it is not true. In searching through a cluster, types are prioritised over modifiers (so that nouns are always selected first). The procedure terminates as soon as the description is distinguishing or all the available words have been exhausted. These criteria are essentially those used in Dale and Reiter's (1995) Incremental Algorithm.

Definition 2 requires that $w(D)$ be minimised. Let $\mathbb{P}_D$ be the set of perspectives represented in D on termination. *Maximal coherence* would require $\mathbb{P}_D$ to be the subgraph of $\mathcal{G}$ with the lowest total cost from which a distinguishing description could be constructed. Finding such Shortest Connection (Steiner) Networks is a known intractable problem. Therefore, we adopt a greedy interpretation of the coherence constraint, whereby the algorithm aims to maximise *local coherence*. Note that any maximally coherent description trivially satisfies the new definition:

Definition 3. **Definition: Local coherence.**
A description D is *locally coherent* iff there is no D' coextensive with D, obtained by replacing types from some perspective in $\mathbb{P}_D$ with types from another perspective such that $w(D) > w(D')$.

To achieve local coherence, the algorithm maintains a set $Nodes$ of those perspectives in $\mathcal{G}$ represented in the description D (that is, those clusters from which a word has been selected). At any stage, the decision regarding which cluster in the graph to visit next is determined in relation to this set. Let $next(\mathcal{G}) \in V$ be the node (cluster) of the graph to be visited next. There are two cases to be accounted for. The first is the case where $Nodes$ is empty. This is the case right at the beginning of the Content Determination procedure, for example. Here, the initial (root) node for search is defined as the vertex of $\mathcal{G}$ with the greatest number of referents in its extension. Such a decision is aimed at reducing potential processing overhead even further, since the procedure begins at the node most likely to satisfy maximal coherence. More precisely, this is the node such that the union of the extensions of all the words it contains has the largest intersection with R of all the available nodes, making it a likely candidate for distinguishing R from a single perspective. For the second case, where $Nodes$ is not empty, the next node to be visited is defined as the node in V which results in

Algorithm 1 Content Determination procedure

```
1: while D is not distinguishing do
2:     if no nodes have been visited then
3:         next(G) = the node covering the greatest number of referents
4:     else
5:         next(G) = the node which minimises the total cost of D
6:     end if
7:     for word w in the current node do
8:         if w is true of some referent, and removes some distractors then
9:             add w to D
10:            add the current node to Nodes
11:        end if
12:    end for
13:    remove the current node from V
14: end while
```

the least increase in the total cost $w(D)$. This is estimated with respect to $Nodes$, since this set contains all the perspectives from which some words have been selected for D up to a given point in the process. The two cases are defined as follows:

$$next(\mathcal{G}) = \begin{cases} \max_{v \in V} \left| \bigcup_{w \in v} [\![w]\!] \cap R \right| & \text{if } Nodes = \emptyset \\ \min_{n \in V} \sum_{u \in Nodes} \delta(u, n) & \text{otherwise.} \end{cases} \quad (5)$$

A summary of the steps in the procedure is given in Algorithm 1. The results of this procedure closely approximate maximal coherence, because the algorithm starts with the vertex most likely to distinguish the referents, and then greedily proceeds to those nodes which minimise $w(D)$ given the current state, that is, taking all previously used nodes into account. As an example of the output, we will take $R = \{e_1, e_3, e_4\}$ as the intended referents in Table IV. In selecting the initial node, there is a tie between clusters 2 and 3 in Figure 3, since all three entities have type properties in these clusters. In either case, the entities are distinguishable from a single cluster. If cluster 3 is selected as the root, the output is $\lambda x\,[physicist(x) \vee biologist(x) \vee chemist(x)]$. In case the algorithm selects cluster 2 as the root node the final output is the logical form $\lambda x\,[man(x) \vee (woman(x) \wedge plump(x))]$. There is an alternative description that the algorithm does not consider. An algorithm that aimed for conciseness would generate
$\lambda x\,[professor(x) \vee man(x)]$ (*the professor and the men*), which does not satisfy local coherence. These examples therefore highlight the possible tension between the avoidance of redundancy and achieving coherence. It is to an investigation of this tension that we now turn.

4. Evaluation

It has been known at least since Dale and Reiter (1995) that the best distinguishing description is not always the shortest one. Yet, brevity plays a part in all GRE algorithms, sometimes in a strict form (Dale, 1989), or by letting the algorithm *approximate* the shortest description (Dale and Reiter, 1995). This is also true of references to sets, the clearest example being Gardent's (2002) constraint based approach, which always finds the description with the smallest number of logical operators. Such proposals do not take coherence (in our sense of the word) into account.

Our evaluation took the form of an experiment to compare the output of our *Coherence Model* with the family of algorithms that have placed brevity at the centre of content determination. The evaluation compared readers' preference for coextensive descriptions which were optimally brief or not ($\pm b$) and also either optimally coherent or not ($\pm c$). Non-brief descriptions took the form *the A, the B and the C*. Brief descriptions 'aggregated' two disjuncts into one (e.g. *the A and the D's* where D comprises the union of B and C). We expected to find that:

H1 $+c$ descriptions are preferred over $-c$.

H2 $(+c, -b)$ descriptions are preferred over ones that are $(-c, +b)$.

H3 $+b$ descriptions are preferred over $-b$.

Confirmation of H1 would be interpreted as evidence that, by taking coherence into account, our algorithm is on the right track. If H3 were confirmed, then earlier algorithms were (also) on the right track by taking brevity into account. Confirmation of H2 would suggest that, in references to sets, conceptual coherence is more important than brevity (defined as the number of disjuncts in a disjunctive reference to a set).

4.1. Materials, design and procedure

Six discourses were constructed, each introducing three entities. Each set of three could be described using all 4 possible combinations of $\pm b \times \pm c$ (see Figure 4). Entities were *people* in two of the discourses, and *artefacts* of various kinds in the remainder. Properties of entities were introduced textually; the order of presentation was randomised. A forced-choice task was used. Each discourse was presented with 2 possible continuations consisting of a sentence with a plural subject NP, and participants were asked to indicate the one they found most natural. The 6 comparisons corresponded to 6 sub-conditions:

Three old manuscripts were auctioned at Sotheby's.

e_1 One of them is a book, a biography of a composer.

e_2 The second, a sailor's journal, was published in the form of a pamphlet. It is a record of a voyage.

e_3 The third, another pamphlet, is an essay by Hume.

Continuations:

$(+c, -b)$ The biography, the journal and the essay were sold to a collector.

$(+c, +b)$ The book and the pamphlets were sold to a collector.

$(-c, +b)$ The biography and the pamphlets were sold to a collector.

$(-c, -b)$ The book, the record and the essay were sold to a collector.

Figure 4. Example domain in the evaluation

C1. **Coherence constant**
- a. $(+c, -b)$ vs. $(+c, +b)$
- b. $(-c, -b)$ vs. $(-c, +b)$

C2. **Brevity constant**
- a. $(+c, -b)$ vs. $(-c, -b)$
- b. $(+c, +b)$ vs. $(-c, +b)$

C3. **Tradeoff/control**
- a. $(+c, -b)$ vs. $(-c, +b)$
- b. $(-c, -b)$ vs. $(+c, +b)$

Participants saw each discourse in a single condition. They were randomly divided into six groups, so that each discourse was used for a different condition in each group. 39 native English speakers, all undergraduates at the University of Aberdeen, took part in the study.

4.2. Results and discussion

Results were coded according to whether a participant's choice was $\pm b$ and/or $\pm c$. Table V displays response proportions. In what follows, we report the results of a Friedman ANOVA to compare response proportions across conditions, and χ^2 tests to compare proportions of $\pm b$ and $\pm c$ responses within conditions.

Conditions differed significantly both by subjects (Friedman $\chi^2 = 107.3, p < .001$) and by items ($\chi^2 = 30.2, p < .001$). When coherence

Table V. Response proportions (%)

	C1a	C1b	C2a	C2b	C3a	C3b
$+b$	51.3	43.6	–	–	30.8	76.9
$+c$	–	–	82.1	79.5	69.2	76.9

was kept constant (C1a and C1b), the likelihood of a response being $+b$ was no different from $-b$ (C1a: $\chi^2 = .023, p = .8$; C1b: $\chi^2 = .64, p = .4$); the conditions C1a and C1b did not differ significantly ($\chi^2 = .46, p = .5$). By contrast, conditions where brevity was kept constant (C2a and C2b) resulted in significantly higher proportions of $+c$ choices (C2a: $\chi^2 = 16.03, p < .001$; C2b: $\chi^2 = 13.56, p < .001$). No difference was observed between C2a and C2b ($\chi^2 = .08, p = .8$). In the trade off case (C3a), participants were much more likely to select a $+c$ description than a $+b$ one ($\chi^2 = 39.0, p < .001$); a majority opted for the $(+b, +c)$ description in the control case ($\chi^2 = 39.0, p < .001$).

The results show that readers' choices are strongly impacted by Coherence. They do not indicate a preference for brief descriptions. To an extent, this echoes previous findings that speakers may relinquish brevity in favour of task or discourse constraints (Jordan, 2000). However, it is remarkable that the experiment fails to show any brevity effect in situations where it is unclear that any purpose was served by being non-brief. In the two conditions where coherence was kept constant, were speakers concerned with brevity, they would be expected to opt for the $+b$ descriptions.

Since this experiment compared our algorithm against the current state of the art in references to sets, these results do not necessarily warrant the affirmation of the null hypothesis in the case of H3. We limited Brevity to number of disjuncts, omitting negation, and varying only between length 2 or 3. Longer or more complex descriptions may evince different tendencies. Nevertheless, the results show a strong impact of Coherence, compared to (a kind of) brevity, in strong support of the algorithm presented above, as a realisation of the Coherence Model.

5. Conclusions and future work

This paper started with an empirical investigation of conceptual coherence in plural reference, leading to a definition of *local* coherence as the basis for a new algorithm. Our evaluation strongly supports our Coherence Model and casts doubt on the importance of brevity as a

criterion for the felicity of plural descriptions, echoing an earlier view expressed with respect to singular reference (Dale and Reiter, 1995).

We are currently extending this work in two directions. First, we are investigating similarity effects *across noun phrases*, and their impact on text readability. Finding an impact of such factors would make this model a useful complement to current theories of discourse, which usually interpret coherence in terms of discourse/sentential structure. Second, we intend to study the effects that lexical ambiguity and polysemy may have in this area. Consistent with most of the state of the art in GRE (with the exception of Siddharthan and Copestake, 2004), the work presented here does not account for these issues, though ambiguous words may result in unclarities in generated descriptions. Here we just note that words are often disambiguated by nearby words that are similar: *the river and its bank* is probably not made unclear by the fact that 'bank' could denote a financial institution.

As is always the case in Natural Language Generation (NLG), our empirical findings afford several algorithmic interpretations. One aspect of Algorithm 1 which is under-determined by data is worth highlighting, because it contrasts with received wisdom in linguistic pragmatics. Recall that the algorithm starts from the perspective that contains the greatest number of referents in its extension. This procedure is motivated by considerations of discriminatory power (i.e., greediness). An alternative model would choose its starting point on the basis of whatever perspective is pragmatically most opportune, perhaps because of its compatibility with a speaker's goal or the perspective adopted earlier in the discourse. An algorithm along these lines may well offer an even better match to human behaviour.

It is only in recent years that NLG has started to follow other areas of Natural Language Processing by taking empirical information into account. A dominant methodology uses corpus-based metrics, either for ranking possible outputs in an over-generation architecture (Langkilde, 2000; Varges and Mellish, 2001), or to compare outputs to a corpus 'gold standard' (Papineni et al., 2002). The present paper is similar in spirit, insofar as our algorithm uses corpus-derived similarity estimates (following Lin, 1998b, and Kilgarriff, 2003). However, the empirical grounding for the algorithm relied on psycholinguistic experiments, as did the evaluation. The relative cost of such experiments is offset by their allowing the falsification of precise hypotheses, whereas corpora tend to be sources of exclusively positive evidence, making it hard to judge whether rare or unattested expressions are unacceptable. Indeed, previous work (Reiter and Sripada, 2002) has shown that the opposite can be the case.

Acknowledgements

Thanks to Ielka van der Sluis, Imtiaz Khan, Ehud Reiter, Chris Mellish, Graeme Ritchie and Judith Masthoff for useful comments. This work forms part of the TUNA project, supported by EPSRC grant no. GR/S13330/01 (`http://www.csd.abdn.ac.uk/research/tuna`).

References

Aloni, M.: 2002, 'Questions under cover'. In: D. Barker-Plummer, D. Beaver, J. van Benthem, and P. S. de Luzio (eds.): *Words, Proofs, and Diagrams.* Stanford, Ca.: CSLI.

Bard, E. G., D. Robertson, and A. Sorace: 1996, 'Magnitude estimation of linguistic acceptability'. *Language* **72**(1), 32–68.

Barry, C., C. M. Morrison, and A. W. Ellis: 1997, 'Naming the Snodgrass and Vanderwart pictures'. *Quarterly Journal of Experimental Psychology* **50A**(3), 560–585.

Church, K. W. and P. Hanks: 1990, 'Word association norms, mutual information and lexicography'. *Computational Linguistics* **16**(1), 22–29.

Clark, E.: 1997, 'Conceptual perspective and lexical choice in acquisition'. *Cognition* **64**, 1–37.

Dale, R.: 1989, 'Cooking up referring expressions'. In: *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL-89.*

Dale, R. and E. Reiter: 1995, 'Computational interpretation of the Gricean maxims in the generation of referring expressions'. *Cognitive Science* **19**(8), 233–263.

Eschenbach, C., C. Habel, M. Herweg, and K. Rehkamper: 1989, 'Remarks on plural anaphora'. In: *Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics, EACL-89.*

Gardent, C.: 2002, 'Generating minimal definite descriptions'. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-02.*

Gatt, A.: 2006, 'Structuring knowledge for reference generation: A clustering algorithm'. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL-06.*

Gatt, A. and K. van Deemter: 2005, 'Semantic similarity and the generation of referring expressions: A first report'. In: *Proceedings of the 6th International Workshop on Computational Semantics, IWCS-6.*

Horacek, H.: 2004, 'On referring to sets of objects naturally'. In: *Proceedings of the 3rd International Conference on Natural Language Generation, INLG-04.*

Jordan, P. W.: 2000, 'Can nominal expressions achieve multiple goals? A corpus study'. In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL-00.*

Kaup, B., S. Kelter, and C. Habel: 2002, 'Representing referents of plural expressions and resolving plural anaphors'. *Language and Cognitive Processes* **17**(4), 405–450.

Keller, F.: 2003, 'A psychophysical law for linguistic judgments'. In: *Proceedings of the 25th Annual Conference of the Cognitive Science Society, CogSci-03.*

Kilgarriff, A.: 2003, 'Thesauruses for natural language processing'. In: *Proceedings of the Conference on Natural Language Processing and Knowledge Engineering, NLP/KE-03.*

Kilgarriff, A., P. Rychly, P. Smrz, and D. Tugwell: 2004, 'The Sketch Engine'. In: *Proceedings of the 11th International Congress of the European Association for Lexicography, EURALEX-04.*

Koh, S. and C. Clifton: 2002, 'Resolution of the antecedent of a plural pronoun: Ontological categories and predicate symmetry'. *Journal of Memory and Language* **46**, 830–844.

Kronfeld, A.: 1989, 'Conversationally relevant descriptions'. In: *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL-89.*

Langkilde, I.: 2000, 'Forest-based statistical language generation'. In: *Proceedings of the 1st Meeting of the North Americal Chapter of the Association for Computational Linguistics, NAACL-00.*

Lapata, M., S. McDonald, and F. Keller: 1999, 'Determinants of Adjective-Noun plausibility'. In: *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, EACL-99.*

Lin, D.: 1998a, 'Automatic retrieval and clustering of similar words'. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, COLING-ACL-98.*

Lin, D.: 1998b, 'An information-theoretic definition of similarity'. In: *Proceedings of the International Conference on Machine Learning.*

Murphy, G.: 1984, 'Establishing and accessing referents in discourse'. *Memory and Cognition* **12**, 489–497.

Papineni, S., T. Roukos, W. Ward, and W. Zhu.: 2002, 'BLEU: a. method for automatic evaluation of machine translation'. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-02.*

Pederson, T., S. Patwardhan, and J. Michelizzi: 2004, 'WordNet::Similarity — Measuring the relatedness of concepts'. In: *Proceedings of the Nineteenth National Conference on Artificial Intelligence, AAAI-04.*

Pustejovsky, J.: 1995, *The Generative Lexicon.* Cambridge, Ma.: MIT Press.

Reiter, E. and S. Sripada: 2002, 'Human variation and lexical choice'. *Computational Linguistics* **28**, 545–553.

Sanford, A. and L. Moxey: 1995, 'Notes on plural reference and the scenario-mapping principle in comprehension'. In: C.Habel and G.Rickheit (eds.): *Focus and cohesion in discourse.* Berlin: de Gruyter.

Siddharthan, A. and A. Copestake: 2004, 'Generating referring expressions in open domains'. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, ACL-04.*

Stevens, S. S.: 1957, 'On the psychophysical law'. *Psychological Review* **64**, 153–181.

Stone, M.: 2000, 'On identifying sets'. In: *Proceedings of the 1st International Conference on Natural Language Generation, INLG-00.*

van Deemter, K.: 2000, 'Generating vague descriptions'. In: *Proceedings of the First International Conference on Natural Language Generation, INLG-00.*

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

Varges, S. and C. Mellish: 2001, 'Instance-based natural language generation'. In: *Proceedings of the 2nd Meeting of the North Americal Chapter of the Association for Computational Linguistics, NAACL-01.*