Conceptual Coherence in the Generation of Referring Expressions

Albert Gatt
Department of Computing Science
University of Aberdeen
agatt@csd.abdn.ac.uk

Kees van Deemter
Department of Computing Science
University of Aberdeen
kvdeemte@csd.abdn.ac.uk

Abstract

One of the challenges in the automatic generation of referring expressions is to identify a set of domain entities coherently, that is, from the same conceptual perspective. We describe and evaluate an algorithm that generates a conceptually coherent description of a target set. The design of the algorithm is motivated by the results of psycholinguistic experiments.

1 Introduction

Algorithms for the Generation of Referring Expressions (GRE) seek a set of properties that distinguish an intended referent from its distractors in a knowledge base. Much of the GRE literature has focused on developing efficient content determination strategies that output the best available description according to some interpretation of the Gricean maxims (Dale and Reiter, 1995), especially Brevity. Work on reference to sets has also proceeded within this general framework (van Deemter, 2002; Gardent, 2002; Horacek, 2004).

One problem that has not received much attention is that of *conceptual coherence* in the generation of plural references, i.e. the ascription of related properties to elements of a set, so that the resulting description constitutes a coherent cover for the plurality. As an example, consider a reference to $\{e_1, e_3\}$ in Table 1 using the Incremental Algorithm (IA) (Dale and Reiter, 1995). 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, IA would yield *the postgraduate and the chef*, 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.

	type	**occupation**	**nationality**
e_1	man	postgraduate	maltese
e_2	man	undergraduate	greek
e_3	man	chef	italian

Table 1: Example domain

Such examples lead us to hypothesise the following constraint:

> **Conceptual Coherence Constraint (**CC**)**: As far as possible, describe objects using related properties.

Related issues have been raised in the formal semantics literature. 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 contrasting properties, say *the student and the italian*, the contrast may be misleading for the listener.

Any attempt to port these observations to the GRE scenario must do so without sacrificing logical completeness. While a GRE algorithm should attempt to find the most coherent description available, it should not fail in the absence of a coherent set of properties. This paper aims to achieve a dual goal. First (§2), 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 two experiments. Our focus on 'low-level', lexical, determinants of adequacy constitutes a departure from the standard Gricean view. Second, we describe an algorithm

motivated by the experimental findings (§3) which seeks to find the most coherent description available in a domain according to CC.

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.

Previous work on plural anaphor processing has shown that pronoun resolution is easier when antecedents are ontologically similar (e.g. all humans) (Kaup et al., 2002; Koh and Clifton, 2002). Reference to a heterogeneous set increases processing difficulty.

Our experiments extended these findings to full definite NP reference. Throughout, we used a *distributional* definition of similarity, as defined by Lin (1998), which was found to be highly correlated to people's preferences for disjunctive descriptions (Gatt and van Deemter, 2005). 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. The relevant data in the lexical domain is the grammatical environment in which words occur. This information is represented as a set of triples $\langle rel, w, w' \rangle$, where rel is a grammatical relation, w the word of interest and w' its co-argument in rel (e.g. $\langle$ *premodifies, dog, domestic* $\rangle$). Let $F(w)$ be a list of such triples. The information content of this set is defined as mutual information $I(F(w))$ (Church and Hanks, 1990). The similarity of two words w_1 and w_2, of the same grammatical category, is:

$$\sigma(w_1, w_2) = \frac{2 \times I(F(w_1) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))} \quad (1)$$

For example, if *premodifies* is one of the relevant grammatical relations, then *dog* and *cat* might occur several times in a corpus with the same premodifiers (*tame*, *domestic*, etc). Thus, $\sigma(dog, cat)$ is large because in a corpus, they often occur in the same contexts and there is considerable information gain in a description of their common data.

Rather than using a hand-crafted ontology to infer similarity, this definition looks at real language use. It covers ontological similarity to the extent that ontologically similar objects are talked about in the same contexts, but also cuts across ontological distinctions (for example *newspaper* and *journalist* might turn out to be very similar).

We use the information contained in the SketchEngine database[1] (Kilgarriff, 2003), a largescale implementation of Lin's theory based on the BNC, which contains grammatical triples in the form of *Word Sketches* for each word, with each triple accompanied by a salience value indicating the likelihood of occurrence of the word with its argument in a grammatical relation. Each word also has a thesaurus entry, containing a ranked list of words of the same category, ordered by their similarity to the head word.

Condition	a	b	c	distractor
HDS	spanner	chisel	plug	thimble
LDS	toothbrush	knife	ashtray	clock

Figure 1: Conditions in Experiment 1

2.1 Experiment 1

In Experiment 1, 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 *object1* and the *object 2* cost *amount*.

S2 The *object3* also costs *amount*.

If similarity is a constraint on referential coherence in plural references, then if two targets are similar (and dissimilar to the third), a plural reference to them in S1 should be more likely, with the third entity referred to in S2.

Materials, design and procedure 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).

Each trial consisted of the four pictures arranged in an array on a screen. 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 as a factor with two levels: **High Distributional Similarity (HDS)** meant that b occurred among the top 50 most similar items to a in its Sketchengine thesaurus entry. **Low DS (LDS)**)

[1] `http://www.sketchengine.co.uk`

meant that b did not occur in the top 500 entries for a. Examples are shown in Figure 2.1.

Visual Similarity (VS) of a and b was also controlled. 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.

29 self-reported native or fluent speakers of English completed the experiment over the web. To complete the sentences, participants clicked on the objects in the order they wished to refer to them. Nouns appeared in the next available space[2].

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. We did not expect an effect of VS. In what follows, we report by-subjects Friedman analyses (χ^2_1); by-items analyses (χ^2_2); and by-subjects sign tests (Z) on proportions of responses for pairwise comparisons.

Response frequencies across conditions differed reliably by subjects ($\chi^2_1 = 46.124, p < .001$). The frequency of $a + b$ responses in S1 was reliably higher than that of $a - b$ in the HDS condition ($\chi^2_2 = 41.371, p < .001$), but not the HVS condition ($\chi^2_2 = 1.755, ns$). Pairwise comparisons between HDS and LDS showed a significantly higher proportion of $a + b$ responses in the former ($Z = 4.48, p < .001$); the difference was barely significant across VS conditions ($Z = 1.9, p = .06$).

The results show that, given a clear choice of entities to refer to in a plurality, people are more likely to describe similar entities in a plural description. However, these results raise two further questions. First, given a choice of distinguishing properties for individuals making up a target set, will participants follow the predictions of the CC? (In other words, is distributional similarity relevant for content determination?) Second, does the similarity effect carry over to modifiers, such as adjectives, or is the CC exclusively a constraint on types?

[2]Earler replications involving typing yielded parallel results and high conformity between the words used and those predicted by the picture norms.

	Three millionaires with a passion for antiques were spotted dining at a London restaurant.
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.
	The XXXX were both accompanied by servants, but the bachelor wasn't .

Figure 2: Example discourses

2.2 Experiment 2

Experiment 2 was a sentence continuation task, designed to closely approximate content determination in GRE. Participants saw a series of discourses, in which three entities (e_1, e_2, e_3) were introduced, each with two distinguishing properties. The final sentence in each discourse had a missing plural subject NP referring to two of these. 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, *even if* dissimilar properties were also available.

Materials, design and procedure Materials consisted of 24 discourses, such as those in Figure 2.2. After an initial introductory sentence, the 3 entities were introduced in separate sentences. In all discourses, the pairs $\{e_1, e_2\}$ and $\{e_2, e_3\}$ could be described using either pairwise similar or dissimilar properties (similar pairs are coindexed in the figure). In half the discourses, the distinguishing properties of each entity were *nouns*; thus, although all three entities belonged to the same ontological category (e.g. all human), they had distinct types (e.g. *duke*, *prince*, *bachelor*). In the other half, entities were of the same type, that is the NPs introducing them had the same nominal head, but had distinguishing adjectival modifiers. 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. The experiment was completed by 18 native speakers of English, selected from the Aberdeen NLG Group database. They were randomly assigned to either Version 1 or 2.

Results and discussion Responses were coded 1 if the semantically similar properties were used (e.g. *the prince and the duke* in Fig. 2.2); 2 if the

similar properties were used together with other properties (e.g. *the prince and the bachelor duke*); 3 if a superordinate term was used to replace the similar properties (e.g. *the noblemen*); 4 otherwise (e.g. *The duke and the collector*).

Response types differed significantly in the nominal condition both by subjects ($\chi^2_1 = 45.89, p < .001$) and by items ($\chi^2_2 = 287.9, p < .001$). Differences were also reliable in the modifier condition ($\chi^2_1 = 36.3, p < .001$, $\chi^2_2 = 199.2, p < .001$). However, the trends across conditions were opposed, with more items in the 1 response category in the nominal condition (53.7%) and more in the 4 category in the modifier condition (47.2%). Recoding responses as binary ('similar' = 1,2,3; 'dissimilar' = 4) showed a significant difference in proportions for the nominal category ($\chi^2 = 4.78, p = .03$), but not the modifier category. Pairwise comparisons showed a significantly larger proportion of 1 ($Z = 2.7, p = .007$) and 2 responses ($Z = 2.54, p = .01$) in the nominal compared to the modifier condition.

The results suggest that in a referential task, participants are likely to conform to the CC, but that the CC operates mainly on nouns, and less so on (adjectival) modifiers. Nouns (or types, as we shall sometimes call them) have the function of categorising objects; thus similar types facilitate the mental representation of a plurality in a conceptually coherent way. According to the definition in (1), this is because similarity of two types implies a greater likelihood of their being used in the same predicate-argument structures. As a result, it is easier to map the elements of a plurality to a common role in a sentence. A related proposal has been made by Moxey and Sanford (1995), whose *Scenario Mapping Principle* holds that a plural reference is licensed to the extent that the elements of the plurality can be mapped to a common role in the discourse. This is influenced by how easy it is to conceive of such a role for the referents. Our results can be viewed as providing a handle on the notion of 'ease of conception of a common role'; in particular we propose that likelihood of occurrence in the same linguistic contexts directly reflects the extent to which two types can be mapped to a single plural role.

As regards modifiers, while it is probably premature to suggest that CC plays no role in modifier selection, it is likely that modifiers play a different role from nouns. Previous work has shown that restrictions on the plausibility of adjective-noun combinations exist (Lapata et al., 1999), and that using unlikely combinations (e.g. *the immaculate kitchen* rather than *the spotless kitchen*) impacts processing in online tasks (Murphy, 1984). Unlike types, which have a categorisation function, modifiers have the role of adding information about an element of a category. This would partially explain the experimental results: When elements of a plurality have identical types (as in the modifier version of our experiment), the CC is already satisfied, and selection of modifiers would presumably depend on respecting adjective-noun combination restrictions. Further research is required to verify this, although the algorithm presented below makes use of the Sketch Engine database to take modifier-noun combinations into account.

id	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

Table 2: An example knowledge base

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 3.

We make the following assumptions. There is a set U of domain entities, properties of which are specified in a KB as attribute-value pairs. 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 $R \subseteq U$, the algorithm described below generates a description D in Disjunctive Normal Form (DNF), having 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.

We first make our interpretation of the CC 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.:

$$\forall t, t', t'' \in T :$$
$$\{t, t'\} \subseteq \mathcal{P} \wedge \sigma(t, t'') \geq \sigma(t, t') \rightarrow t'' \in \mathcal{P}$$

The aims of the algorithm are to describe elements of R using types from the same perspective, failing which, it attempts to minimise the distance between the perspectives from which types are selected in the disjunctions of D. Distance between perspectives is defined below.

3.1 Finding perspectives

The system makes use of the SketchEngine database as its primary knowledge source. Since the definition of similarity applies to words, rather than properties, the first step is to generate all possible lexicalisations of the available attribute-value pairs in the domain. In this paper, we simplify by assuming a one-to-one mapping between properties and words.

Another requirement is to distinguish between type properties (the set T), and non-types (M)[3]. The Thesaurus is used to find pairwise similarity of types in order to group them into related clusters. Word Sketches are used to find, for each type, the modifiers in the KB that are appropriate to the type, on the basis of the associated salience values. For example, in Table 3, e_3 has *plump* as the value for **girth**, which combines more felicitously with *man*, than with *biologist*.

Types are clustered using the algorithm described in Gatt (2006). For each type t, the algorithm finds its nearest neighbour n_t in semantic space. Clusters are then found by recursively grouping elements with their nearest neighbours. If t, t' have a common nearest neighbour n, then $\{t, t', n\}$ is a cluster. Clearly, the resulting sets are convex in the sense of Definition 1. Each modifier is assigned to a cluster by finding in its Word Sketch the type with which it co-occurs with the greatest salience value. Thus, a cluster is a pair

[3]This is determined using corpus-derived information. Note that T and M need not be disjoint, and entities can have more than one type property

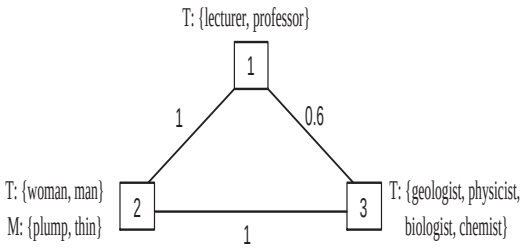

Figure 3: Perspective Graph

$\langle \mathcal{P}, M' \rangle$ where $\mathcal{P}$ is a perspective, and $M' \subseteq M$. The distance $\delta(A, B)$ between two clusters A and B is defined straightforwardly in terms of the distance between their perspectives $\mathcal{P}_A$ and $\mathcal{P}_B$:

$$\delta(A, B) = \frac{1}{1 + \frac{\sum_{x \in \mathcal{P}_A, y \in \mathcal{P}_B} \sigma(x,y)}{|\mathcal{P}_A \times \mathcal{P}_B|}} \qquad (2)$$

Finally, a weighted, connected graph $\mathcal{G} = \langle V, E, \delta \rangle$ is created, where V is the set of clusters, and E is the set of edges with edge weights defined as the semantic distance between perspectives. Figure 3.1 shows the graph constructed for the domain in Table 3.

We now define the coherence of a description more precisely. Given a DNF description D, we shall say that a perspective $\mathcal{P}$ is *realised 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 realised 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 $w(D) > w(D')$.

(Note that several descriptions of the same referent may all be maximally coherent.)

3.2 Content determination

The core of the content determination procedure maintains the DNF description D as an associative array, such that for any $r \in R$, $D[r]$ is a conjunction of properties true of r. Given a cluster $\langle \mathcal{P}, M \rangle$, the procedure searches incrementally first through $\mathcal{P}$, and then M, selecting properties that are true of at least one referent and exclude some distractors, as in the IA (Dale and Reiter, 1995).

By Definition 2, the task of the algorithm is to minimise the total weight $w(D)$. If $\mathbb{P}_D$ is the

set of perspectives represented in D on termination, then *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. Under this interpretation, $\mathbb{P}_D$ corresponds to a Shortest Connection, or Steiner, Network. Finding such networks is known to be NP-Hard. Therefore, we adopt a weaker (greedy) interpretation. Under the new definition, if D is the only description for R, then it trivially satisfies maximal coherence. Otherwise, the algorithm aims to maximise *local coherence*.

Definition 3. Local coherence
A description D is *locally coherent* iff:

a. **either** D is maximally coherent **or**

b. 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 $w(D) > w(D')$.

Our implementation of this idea begins the search for distinguishing properties by identifying the vertex of $\mathcal{G}$ which contains the greatest number of referents in its extension. This constitutes the root node of the search path. For each node of the graph it visits, the algorithm searches for properties that are true of some subset of R, and removes some distractors, maintaining a set N of the perspectives which are represented in D up to the current point. The crucial choice points arise when a new node (perspective) needs to be visited in the graph. At each such point, the next node n to be visited is the one which minimises the total weight of N, that is:

$$\min_{n \in V} \sum_{u \in N} w(u, n) \tag{3}$$

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 3. First, the algorithm determines the cluster with the greatest number of referents in its extension. In this case, there is a tie between clusters 2 and 3 in Figure 3.1, 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 (for example, in the Dale and Reiter's IA). This is also true of references to sets, the clearest example being Gardent's 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. This raises obvious questions about the relative importance of brevity and coherence in reference to sets.

The 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. Participants were asked to compare pairs of descriptions of one and the same target set, selecting the one they found most natural. Each description could either be 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

	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.
$(+c, -b)$	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

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 be interpreted as meaning 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).

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 human 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:

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.

Results and discussion Results were coded according to whether a participant's choice was $\pm b$ and/or $\pm c$. Table 4 displays response proportions. Overall, the conditions had a significant impact on responses, both by subjects (Friedman $\chi^2 = 107.3, p < .001$) and by items ($\chi^2 = 30.2, p < .001$). When coherence 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 very 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 tradeoff 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 ($\chi^2 = 39.0, p < .001$).

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

Table 3: Response proportions (%)

The results strongly support H1 and H2, since participants' choices are impacted by Coherence. They do not indicate a preference for brief descriptions, a finding that echoes Jordan's (2000), to the effect that speakers often relinquish brevity in favour of observing task or discourse constraints. 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 might 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 reference, which led to a definition of *local* coherence as the basis for a new greedy algorithm that tries to minimise the semantic distance between the perspectives repre-

sented in a description. The evaluation strongly supports our Coherence Model.

We are 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 relinquish the assumption of a one-to-one correspondence between properties and words (cf. Siddharthan and Copestake (2004)), making use of the fact that words can be disambiguated by nearby words that are similar. To use a well-worn example: the 'financial institution' sense of *bank* might not make *the river and its bank* lexically incoherent as a description of a piece of scenery, since the word *river* might cause the hearer to focus on the aquatic reading of the word anyway.

6 Acknowledgements

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

References

M. Aloni. 2002. Questions under cover. In D. Barker-Plummer, D. Beaver, J. van Benthem, and P. Scotto de Luzio, editors, *Words, Proofs, and Diagrams*. CSLI, Stanford, Ca.

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

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

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

Robert Dale. 1989. Cooking up referring expressions. In *Proc. 27th Annual Meeting of the Association for Computational Linguistics.*

C. Gardent. 2002. Generating minimal definite descriptions. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics.*

A. Gatt 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.*

A. Gatt. 2006. Structuring knowledge for reference generation: A clustering algorithm. In *Proc. 11th Conference of the European Chapter of the Association for Computational Linguistics.*

H. Horacek. 2004. On referring to sets of objects naturally. In *Proc. 3rd International Conference on Natural Language Generation.*

P. W. Jordan. 2000. Can nominal expressions achieve multiple goals? In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.*

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

A. Kilgarriff. 2003. Thesauruses for natural language processing. In *Proc. NLP-KE, Beijing.*

S. Koh 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.

A. Kronfeld. 1989. Conversationally relevant descriptions. In *Proc. 27th Annual Meeting of the Association for Computational Linguistics.*

M. Lapata, S. McDonald, and F. Keller. 1999. Determinants of adjective-noun plausibility. In *Proc. 9th Conference of the European Chapter of the Association for Computational Linguistics.*

D. Lin. 1998. An information-theoretic definition of similarity. In *Proc. International Conference on Machine Learning.*

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

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

A. Siddharthan and A. Copestake. 2004. Generating referring expressions in open domains. In *Proc. 42nd Annual Meeting of the Association for Computational Linguistics.*

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