Towards a psycholinguistically motivated algorithm for referring to sets: The role of semantic similarity. *

Albert Gatt and **Kees van Deemter**
Department of Computing Science
University of Aberdeen
{agatt, kvdeemte}@csd.abdn.ac.uk

Abstract

This paper explores the role of semantic similarity in content selection and aggregation of expressions referring to sets. Similarity plays a role in ensuring that a referring expression corresponds to a *coherent conceptual gestalt*. On the basis of corpus-based and experimental evidence we propose an algorithm which (a) separates content selection and aggregation to avoid a combinatorial explosion; (b) uses similarity between entities to prioritise among search alternatives.

1 Introduction

The problem of Generating Referring Expressions (GRE) can be summed up as a search for properties in a database whose combination uniquely distinguishes an intended referent. We shall refer to the latter as the *Uniqueness Criterion*. Dale and Reiter's [1995] Incremental Algorithm (IA), has emerged as a gold standard, modelling search as hillclimbing along a pre-determined preference order of properties, where a property is selected if it excludes at least one distractor. Thus, the IA disregards the Gricean exhortation to maintain brevity. This approach is justifiable on psycholinguistic grounds, since adult speakers do not respect the brevity maxim, but tend to introduce redundant information in their descriptions [Ford and Olson, 1975; Sonnenschein, 1982; Pechmann, 1989; Arts, 2004]. However, the reliance on an a priori definition of a preference order in the IA has two consequences. First, the incremental approach does not do justice to human flexibility in content selection, where the choice of attributes may depend on communicative and linguistic context and intention [Arts, 2004]. A partial solution to these problems has been proposed by Siddharthan and Copestake [2004], whose version of the IA dynamically orders properties depending on their discriminatory potential prior to starting search. This approach, which bears some resemblance to the one proposed here, is discussed further below. A second consequence is the potential for highly redundant descriptions in the worst case.

The problems of redundancy and lack of flexibility are especially evident once GRE is extended to sets [van Deemter, 2002]. In this case, properties are combined not only via conjunction/intersection, but also disjunction/union. Hence, not only does the problem of efficiency again rear its head (the algorithm now searches through $n-$level disjunctions of properties in the preference order), but the adequacy of the outcome becomes a crucial consideration. A highly redundant description, or one which is logically complex, may be of little use to the listener/reader. Strategies to deal with this problem have ranged from best-first search and optimisation of a logical form [Horacek, 2003; 2004] to constraints on description length [Gardent, 2002]. Unfortunately, none of these strategies is backed by empirical data on how humans process references to sets. There is in fact a surprising gap in the psycholinguistic literature, where studies on reference have generally focused on the singleton case (cf. [Levelt, 1989] for a review).

An additional problem that arises in reference to sets, and one which is a focus of the present paper, is the problem of choice. While in the singleton case, the notion of preference may be fairly clear-cut, when disjunctions of the form $p_i \vee p_j$ are being considered, the problem is not only whether the disjunction contributes to satisfying Uniqueness, but also whether the description so generated is a *coherent cover* for the set in question. This paper tackles this issue by defining 'coherence' of a description in terms of the semantic similarity of the properties (sometimes called *descriptors*) used to identify each referent r within a set of intended referents $\mathcal{R}$.

1.1 Sets and the problem of gestalts

One psycholinguistic explanation for the tendency to produce redundant descriptions is that we tend to conceptualise objects in the world as gestalts, rather than as loose collections of attributes. For instance, a speaker might utter *the black triangle*, where *the triangle* would suffice as a distinguishing description. According to Schriefers and Pechmann [1988], 'this expresses the notion that there is a basic conceptual unit *black triangle*'. In other words, while analysing separate attributes of objectrs may be a low-level perceptual process, at the level at which conceptualisation for message formulation

*This work forms part of the TUNA project. See `http://www.itri.brighton.ac.uk/projects/tuna`. Part of this work was conducted at the ITRI, University of Brighton. Thanks are due to Adam Kilgarriff for making the WASPS thesaurus available, and Ehud Reiter for useful comments on previous drafts of this paper. The support of EPSRC Grant No. GR/S13330/01 is gratefully acknowledged.

takes place, the speaker considers the entire unit *black triangle* for inclusion. This is taken into account in the IA by ranking certain preferred properties early in the preference order, increasing the likelihood (though not guaranteeing) their inclusion.

When there is a *set* of intended referents, the selection of such 'basic units' is more difficult. The simplest case is perhaps the one where all entities have properties in common. Shared properties may give rise to a salient perceptual group, along the lines suggested by the Gestalt psychologists (cf. [Wertheimer, 1938; Rock, 1983]), resulting in an inclusion of the properties identifying this group in a speaker's description. For instance, suppose there are three referents, and they all happen to have the properties *chair* and *black*, and the conjunction $chair(x) \wedge black(x)$ is distinguishing for each of these referents. Here, *the black chairs* would suffice to satisfy Uniqueness and would result in a coherent description, in a way that parallels Schriefers and Pechmann's example. On the other hand, reference to sets also involves cases where entities do not have properties in common. The choice of properties may have an impact on the perceived acceptability and ease of comprehension of the description, as well as on the extent to which it forms a 'coherent conceptual unit'. As an example, consider the domain in Table 1.

e_1	e_2	e_3
postgraduate	undergraduate	man
physicist	italian	greek

Suppose the set of intended referents were $\{e_1, e_2\}$. The following are among the options: *the postgraduate and the italian*, *the postgraduate and the undergraduate*. Intuitively, the latter is a better description, since the properties selected are more related. Similarly, if the referents were $\{e_2, e_3\}$, the best description seems to be *the italian and the greek*. In a preliminary investigation, [Gatt and van Deemter, 2005], we have shown that speakers find referring expressions of the form *the* N_1 *and (the)* N_2 more acceptable the greater the semantic similarity between the descriptors N_1 and N_2. We proposed an exaplanation based on recent findings in psycholinguistics. Models of lexical access in speech production [Dell, 1986; Roelofs, 2000] have converged on the view that the mental lexicon is an associative memory, where concepts which are semantically related are adjacent (cf. [Levelt, 1998] for a review). Evidence for this comes, for example, from the *semantic interference* effect, where the production of a target word is inhibited by the presence of a semantically similar distractor [Meyer, 1996; Vigliocco *et al.*, 2002; Damian, 2000; Damian *et al.*, 2001]. This is often explained with reference to spreading activation: Semantic similarity results in concepts (and hence, lemmas) for both target and distractor being activated. We proposed that, in the case of references to sets realised as linguistic conjunctions, the selection of an initial descriptor (say, N_1) makes it easier to retrieve a second descriptor which is closely associated and include it in the description. Given our finding that such descriptions are preferred, it may also represent a better gestalt or 'conceptual unit' since it refers using related properties. The rest of this paper is concerned with making these hypotheses more precise, describing the empirical evidence for them, and proposing a GRE algorithm based on the empirical data. In what follows, we focus on the generation of references to sets using *distributive* properties [1]

2 Empirical evidence

We describe three experiments, one using corpus data and two with human participants, that attempt to make precise the hypothesis outlined in Section 1 – namely, that semantic similarity plays a role in making descriptions correspond to a coherent conceptual gestalt – and motivate the approach to GRE described in Section 3. Before doing so, it is worth recapitulating the findings of [Gatt and van Deemter, 2005], as these provide the motivation for the present work. In that experiment, participants were asked to rate their preferences for descriptions of the form *the* N_1 *and (the)* N_2 using magnitude estimation. In this method, participants first rate a *modulus* item, i.e. an initial description, using a numeric scale of their own choice. Subsequently, each phrase is compared to the modulus and rated, using the same numeric scale, to reflect its relative acceptability given the modulus (see [Bard *et al.*, 1996]). The advantage of this approach is that it places acceptability ratings on a ratio scale, making the comparison of ratings of different phrases meaningful. We gave participants 32 different phrases to rate. Each of these involved a pairing of two nouns denoting animate or inanimate entities. Noun pairs were matched for frequency in the BNC. The semantic similarity of each noun pair was calculated using four similarity measures, as well as a random number in $(0, 1)$. We correlated the geometric means of the ratings of phrases to the semantic similarity estimates. All similarity correlations were significant, while the random measure had no correlation with acceptability. The two measures which were most highly rated were those defined by Kilgarriff and Tugwell [2001] for the WASPS thesaurus ($r = .576$, $p < .01$), and by Resnik [1995] for WordNet[2] ($r = .538$, $p < .01$).

The WASPS measure is an adaptation of a Lin's [1998] measure of mutual information, where the significance of the collocation of two words is based on their occurrence in dependency triples in corpora. A dependency triple always has the form W_1, R, W_2, for any two words W_1, W_2 and grammatical relation R. The *salience* of a collocation of two words is calculated as the product of log frequency and mutual information I, where

$$I = \log\Big(\frac{\|*, R, *\| \times \|W_1, R, W_2\|}{\|W_1, R, *\| \times \|*, R, W_2\|}\Big)$$

where $\| \ldots \|$ is the frequency of some dependency triple, and $(*)$ is a wildcard. The WASPS thesaurus, where entries are words accompanied by their significant collocates belonging to the same grammatical category, was compiled from the significant patterns in the BNC.

Given the results, we have adopted the WASPS measure as the best correlated measure with humans intuitions, and

[1] *Distributive* properties are those that are true of *each* element of the set, e.g. *red*, while *collective* properties are true of the set as a whole, e.g. *clustered*.

[2] The Resnik measure was computed using the WordNet::Similarity package for Perl [Pederson *et al.*, 2004]

have used it in most subsequent experiments, as well as our implementation. The measure is attractive for another reason: it does not depend on a hand-crafted taxonomy such as WordNet for estimates of similarity, as does the Resnik measure. The high correlation obtained between preferences and WASPS measures can be interpreted as implying that, the more associated nouns are in a corpus, the higher the preference for a plural noun phrase containing both. However, a correlational study is hardly sufficient to justify the argument that semantic similarity is a factor in predicting human preferences during referential tasks; rather, our purpose in the study was to identify a working definition of similarity, obtaining a preliminary result indicating whether the hypothesis was on the right track. We now report on three further experiments, which focus on the similarity of nominal descriptors in definite descriptions.

2.1 Experiment 1. Evidence from corpora

We collected a sample A of definite descriptions from the BNC, all of which had the form *the* N_1 *and (the)* N_2, and occurred as either subject or object NPs. The head nouns of these NPs were extracted, resulting in a set of noun-noun pairs. Following morphological processing[3], the similarity of each noun-noun pair was found in the WASPS thesaurus. Noun pairs for which no similarity value was found were removed from the sample, leaving a total of 4617 descriptions (i.e. noun pairs). A second sample R of noun-pairs was produced from the first by taking the first noun in each pair and randomly pairing it with one of the second nouns. Once more, similarity values were found for the randomly generated noun pairs, removing those pairs for which no value had been found. This left a total of 1433 descriptions. In itself, this is already an indicative value, for it implies that far fewer noun pairs in the random sample had a significant value in the WASPS database.

Three tests were carried out. Each time, the similarity values in the two samples were split into two intervals using a different threshold s, and a t-test for collocations conducted to attempt to falsify the following hypotheses:

- H_1: The probability P_{σ_A} of finding a value of $\sigma(N_1, N_2) \geq s$ in sample A is no different than the corresponding probability P_{σ_R} in sample R.
- H_2: P_{σ_A} such that $\sigma(N_1, N_2) < s$ is no different from the corresponding P_{σ_R}

The values of s were $\sigma = 0.04$, $\sigma = 0.06$ and $\sigma = 0.08$. The formula for the t-value in such cases is

$$\frac{P(\sigma_A) + P(\sigma_R)}{\sqrt{\frac{P(\sigma_A)}{n} + \frac{P(\sigma_R)}{m}}}$$

where n is the total number of cases in sample A, m is the total in sample R. The results of the tests are given in the following table, which displays probabilities for the two intervals in the two samples, and the corresponding t-value.

[3]Conducted using morph [Minnen *et al.*, 2001]

	Attested	Random	
interval	$P(\sigma_A)$	$P(\sigma_R)$	t
$\sigma < 0.04$	0.4	0.8	16.09
$\sigma \geq 0.04$	0.6	0.1	25.48
$\sigma < 0.06$	0.7	0.9	8.98
$\sigma \geq 0.06$	0.2	0.03	27.99
$\sigma < 0.08$	0.8	0.9	4.51
$\sigma \geq 0.08$	0.1	0.01	21.15

It is evident from the above table that the randomly generated model results in a gross over-estimation of the probability of finding a noun pair such that $\sigma(N_1, N_2) < s$, while underestimating the probability of finding pairs such that $\sigma(N_1, N_2) \geq s$. The lowest t-values are found for the comparison between $P(\sigma_A)$ and $P_(\sigma_R)$ when $\sigma < 0.06$ and $\sigma < 0.08$. This reflects a more general trend, that values above 0.04 are comparatively rare. This is likely due to a combination of three factors. First, given the limited amount of post-processing of the corpus data, considerable noise may have been present. Second, several noun pairs, particularly those with low frequency, were not found in the database, thus potentially skewing the probability distribution. Third, that high similarity values in WASPS tend to be less frequent as a whole. This may be due to the frequency of the words themselves (since the wasps value is calculated as a function of frequency and I). For instance, in the entry for *woman*, with a BNC frequency per million of ca. 631, the highest-valued collocate is *man*, with a value of 0.1098. By contrast, the highest-ranked collocate for *professor*, with a frequency per million of 55 is *lecturer*, with a value of 0.0894.

Despite these caveats, the conclusion can still be drawn that relatively high similarity values of noun pairs in definite NP conjunctions have a significantly high probability, and that this pattern is non-random. Nevertheless, there are two problems with corpus-derived data. Our interest is primarily in the generation of referring expressions, but NPs in the sample may or may not have been referential. It is also impossible to ascertain that the tendency observed here is due to a preference for using descriptors which are semantically similar, or due to an artefact of the situation in which texts are authored, wherein context and theme may restrict the range of options. To address these issues, an experiment was designed placing human participants in a situation where they had to refer to three entities, two of which could be described using semantically similar descriptors.

2.2 Experiment 2. Similarity in a reference task.

Design

The experiment was designed to test the hypothesis that the semantic similarity of nouns influences the reference choices of speakers. To this end, we designed a task in which participants were placed in a situation where they were buying objects from an online store. A trial consisted of a scenario in which four pictures of objects were presented. All objects were artifacts and prices were indicated for each; three were identically priced. Participants had to refer to the three identically priced objects in the context of the following two sentences:

1. The *object1* and the *object 2* cost *amount*.

2. The *object3* also costs *amount*.

The choice of interest was the objects they would choose to refer to in sentence 1, i.e. the entities they would aggregate together in a single plural referential NP. We hypothesised that, the more semantically similar the names of two objects, the greater the likelihood that they would be aggregated, given that they were of the right price. This would corroborate the findings presented in the previous section, but would do so in the context of a reference task, where the problems outlined above for corpora would be controlled for. In the experimental design, the three identically priced entities in each trial were called the *targets*. One of these was denoted by a noun that was always semantically *dis*similar to the nouns denoting the other two. In the remaining pair, called the *designated targets* we manipulated semantic similarity as a within-subjects factor with two levels:

1. Semantically Similar (SS+): $\sigma(N_1, N_2) \geq 0.05$
2. Semantically Dissimilar (SS-): $\sigma(N_1, N_2) < 0.05$

where similarity values are obtained from the WASPS thesaurus. Our hypothesis can therefore be rephrased as follows: In case the designated targets are semantically similar (SS+), they have a high likelihood of being referred to together in Sentence 1. In case they are not, the likelihood of the two referents in Sentence 1 being the designated targets is no different from the likelihood of some other combination involving the non-designated target. As an example, the three targets in one SS+ trial were *spanner, chisel, plug* where $\sigma(spanner, chisel) = 0.0747$, $\sigma(spanner, plug) = 0.0346$ and $\sigma(chisel, plug) = 0.0228$. Since the objects were presented graphically, we wanted to control for the possible effect of visual similarity, which might bias speakers towards referring to two entities in Sentence 1 because they were visually similar. Pairs of pictures were presented to a group of participants in the Department of Computing Science at Aberdeen, who were asked to rate them on a scale from 1 (not similar at all) to 10 (highly similar). Participants were instructed to consider only the visual properties of the pictures (all of which were line drawings), and compare them on this basis. Each pair was rated by five people. Mean ratings for each picture pair were calculated. A further factor, *Visual* Similarity, again with two levels, was defined on the basis of these:

1. Visually Similar (VS+): mean judgment ≥ 6
2. Visually Dissimilar (VS-): mean judgment ≤ 2

Thus, the experiment had a full-factorial 2×2 within-subjects design with 2 replications, for a total of 4 trials. We only considered visual similarity for the designated targets. Thus, the possible effect of visual similarity of the third target to one of the designated set is not excluded.

Materials

Pictures were selected from the Snodgrass and Vanderwart set [Snodgrass and Vanderwart, 1980]. This is a set of line drawings of objects of various kinds, which have been normed for use with picture-naming experiments. Each picture is accompanied by its most frequently given name and the percentage agreement among subjects in the norming study. In our experiment, we used the norms published for British English speakers by Barry *et al.* [1997]. Semantic similarity was calculated for picture pairs in the experimental trials on the basis of the most frequent word used for each picture. All the names for the targets selected in the experiment had a percentage agreement over 85%, mostly falling in the 95% to 100% interval.

Procedure

The experiment was conducted over the internet. Participants were instructed about the task and the scenarios they would be exposed to. They were simply told that their task was to select which objects to refer to in the two sentences, given the price indicated. No emphasis was laid on naturalness. At any point, a participant could revise their options. When this happened, a record was kept of each repeitition; only the last repetition is used in the present analysis. In each trial, the four objects were presented arranged horizontally in a table in pseudo-random order. The incomplete sentences were just below the table. In order to complete the sentences, participants had to click on the picture whose name they would like to see entered in the next available sentence slot. Upon clicking, the name of the object appeared in the next available slot. They did not type the names themselves. This measure was taken to avoid divergences among subjects in the selection of head nouns. We reasoned that retrieval of object names with a high rate of naming agreement would be unproblematic.

Participants

Prior to participation, subjects had to fill in a questionnaire with basic details, including their self-rated fluency in English (native, non-native/fluent, not fluent). Participants who started the experiment but failed to complete it were omitted from analysis. A total of 36 self-rated native or fluent speakers of English completed the experiment.

Results and discussion

In the analysis of results, the two objects selected for reference in Sentence 1 were considered. The data was coded according to whether the two designated targets had been selected, or whether one of the designated targets was referred to with the third identically-priced object. The percentage frequency of each type of response was calculated separately for each condition. These results are shown in the following table.

	SS+	SS-	VS+	VS-
designated	66%	35%	53%	48%
other	34%	65%	47%	52%

The frequencies of each type of response when semantic similarity is taken into account show exactly opposite trends: approximately 65% of responses are of the designated type in the SS+ condition; the opposite is the case in the SS- condition. By contrast, the frequencies of response when visual similarity is taken into account do not seem to differ significantly. These impressions are confirmed by chi-squared tests. There is a significant difference in the number of 'designated' versus 'other' responses when semantic similarity is taken into account ($\chi^2 = 28.394$, $p < 0.001$), whereas the difference is not significant when they are compared in the visual

similarity condition ($\chi^2 = .608$, $p = .436$). The importance of semantic similarity in predicting choice of referents in the plural referring expression was further confirmed by a stepwise binary logistic regression, in which the importance of each of the two independent variables, as well as their interaction, was assessed. The model with Semantic and Visual Similarity as the parameters turned out to eb a significant improvement over the null model (with only a constant) (model $\chi^2_1 = 29.607$, $p < 0.001$), with an overall percentage of correctly predicted responses of 65.9%. However, this model did not mark a significant improvement over one with only Semantic Similarity, as shown by the difference in model χ^2 values ($\chi^2_1 - \chi^2_2 = 0.711$).

In summary, the results of this experiment support the hypothesis that the choice of descriptors in plural referring expressions is mediated by the semantic similarity of descriptors. In terms of the hypothesis presented in Section 1, speakers seem to prefer semantically similar descriptors when the set of intended referents has no common properties that can be used to group the set into a coherent gestalt. However, there is a potential confounding factor in Experiment 2 in that participants may simply have been making a selection by clicking from left to right on the identically-priced pictures. Since there were only four objects in the display for each trial, the three targets were adjacent in a number of trials, raising the possibility that targets were simply selected on the basis of their adjacency. This would not explain the significantly higher percentage of designated targets chosen in the Semantically Similar condition; however, to ensure that the results were reliable, the experiment was replicated.

2.3 Experiment 3. Replication.

We ran an experiment identical to Experiment 2. This time, participants were asked to type in the nouns of the entities they wished to refer to in the sentence slots. This was done so as to avoid the order effect alluded to above, whereby someone could simply be clicking on pictures from left to right. Apart from this change, the design, procedure and materials were identical to those in Experiment 2.

Participants

Participants who did not complete the experiment, or who reported themselves as not fluent in English, were omitted from the sample. This left 48 self-reported native or fluent speakers of English.

Coding of results

Since the setup in this experiment allowed for some freedom in lexical selection, some disagreement with the predicted words was observed. Of the total number of trials, there were only 8% of cases in which a participant did not use one of the words predicted by the picture naming norms. On some other trials, participants used compound nouns with the predicted word as head (e.g. *picnic basket* instead of the predicted *basket*), in which case, the head noun of the compound was used in the analysis. A conservative strategy was undertaken in coding the results. In case a participant did not produce a predicted word, the response to Sentence 1 was coded as 'other', unless the word used had a high similarity to the predicted word in the WASPS thesaurus.

Results and discussion

The table below shows frequencies for 'designated' versus 'other' responses in the two conditions.

	SS+	SS-	VS+	VS-
designated	54%	38%	45%	47%
other	45%	62%	54%	53%

As before, a chi-square test was conducted, comparing the different response categories within each condition. Responses differed significantly acros the two levels of Semantic Similarity ($\chi^2 = 10.336$, $p = .001$), but not across levels of Visual Similarity ($chi^2 = .046$, $p > .8$). Once again, a binary logistic regression was carried out to test the significance of each independent variable. The regression model with Semantic Similarity as the only parameter was significantly better than the null model (model $\chi^2 = 10.385$, $p = 0.001$), and had a 58.3% precision rate in predicting the data in the sample. There was no significant increase in model chi^2 values between this first model and Model 2, which included Visual Similarity ($\chi^2_2 - \chi^1_1 = 0.048$), and no increase in accuracy. The difference between Model 1 and Model 3, which contained the interaction, was slightly higher than in Experiment 2 ($\chi^2_3 - \chi^2_1 = 3.17$) and had a slightly higher prediction accuracy of 58.6% over Model 1. The results of this experiment largely corroborate the findings in Experiment 2. Once again, semantic similarity is the main predictor of which descriptors speakers will include in a reference of the form *the* N_1 *and the* N_2.

2.4 General discussion

In summary, results from corpora and experiments with human speakers illustrate the role played by similarity in the choice of descriptors for a set of intended referents. In terms of their relevance to GRE, the results also raise a second point. In the GRE literature, there is often the tacit assumption that reference is a 'one-shot' process, that is, if a referring function is called during the execution of an NLG program, the output is a single referring expression. However, this limits the options considerably. In the present context, if two or more entities have to be referred to and the constraint on semantic similarity is violated, the output may be a description which does not represent a coherent cover for the set in question. In this case, one option would be to generate separate referring expressions for different subsets of the set of intended referents. We explore this possibility further Section 4, after a description of the Content Determination procedure in Section 3.

3 An algorithm for referring to sets

This section describes a GRE algorithm based on our experimental findings. A formalisation of the content-selection procedure is given in the Appendix. In what follows, steps in the relevant sections of the Appendix are cross-referenced in square brackets. As noted in Section 1, we shall be focusing exclusively on distributive reference.

The algorithm requires a nonempty set $\mathcal{DOM}$, the domain of entities [7.1.1] and a nonempty set $\mathcal{PROP}$, the set of all the properties in the domain [7.1.3], from which $\mathcal{P}$, the set of

relevant properties (those pertaining to $\mathcal{R}$) is derived [7.1.4]. In addition, the set of distractors $\mathcal{D}$ is defined as $\mathcal{DOM} - \mathcal{R}$ [7.1.5], i.e. all those entities *not* in $\mathcal{R}$. For each referent $r_i \in \mathcal{R}$, three further data structures are initialised: C_{r_i} is the set of properties of r_i [7.1.7], D_{r_i} is the set of *immediate distractors* of r_i, i.e. those elements of $\mathcal{D}$ which share at least one property with r_i [7.1.8], and UD_{r_i} is the as yet empty description for r_i [7.1.9]. A preference order $\mathcal{PO}$ [7.1.10] and a global description $\mathcal{UD}$ [7.1.11], are initialised as empty. The process of constructing a description for $\mathcal{R}$ is composed of three steps:

1. The construction of a preference order $\mathcal{PO}$ of pairs of relevant properties by the function `orderBySimilarity`.
2. The construction of a distinguishing description $\mathcal{UD}$ by the function `distinguishReferents`, if one can be found. We call $\mathcal{UD}$ an *underspecified description*, for reasons made clear below.
3. An aggregation function which constructs a logical form from $\mathcal{UD}$. This function once again uses similarity between properties in $\mathcal{UD}$ to determine how best to express the description, whether in the form of a single disjunctive logical form, or as separate LFs.

Our approach departs from standard assumptions in that properties are simply represented as words, with entitiy identifiers in the database having a pointer to a wordlist. This facilitates the calculation of similarity values, and also significantly reduces the amount of pre-coded knowledge required (for instance, in specifying properties as attribute-value pairs)[4]. Moreover distinguishing content-selection proper from aggregation avoids the combinatorial explosion in search space, which arises when the IA is generalised to Booleans [van Deemter, 2002]. In general, our algorithm aims to *maximise the internal coherence of a description*; we are not aiming primarily at the generation of *minimal* descriptions, although we discuss an extension in Section 3.4 whereby minimality is guaranteed in case $|\mathcal{R}| = 1$, and aproximated in case $|\mathcal{R}| > 1$.

3.1 Function `orderBySimilarity`

The first step in the generation process is to create a preference order. This function takes as input the set $\mathcal{R}$ and the set of properties C_{r_i} of each $r_i \in \mathcal{R}$. The preference order is constructed by taking the pairwise product of $\mathcal{R}$, denoted $\mathcal{R}^2$, and calculating for each pair $\{r_i, r_j\}$, the pairwise similarity of properties in $\mathcal{C}_{r_i}$ and $\mathcal{C}_{r_j}$. Thus, if $\mathcal{R} = \{r_1, r_2, r_3\}$, the preference order is constructed by comparing each property in C_{r_1} to each property in C_{r_2}, C_{r_3}, etc. The pairwise comparison is necessary since the similarity relation $\sigma(p_i, p_j)$ is binary. The comparison is carried out across pairs of referents (as opposed to just taking pairs of relevant properties) to ensure that, in the subsequent search for distinguishing properties, the preferred property pairs will be those that maximise the similarity between referents. Similarity, for a pair $\langle p_i, p_j \rangle : r_i \in \|p_i\| \wedge r_j \in \|p_j\|$ is calculated as follows:

$$\sigma(p_i, p_j) = \begin{cases} 1 \; if \; p_i = p_j & [7.2.4 - 7.2.4] \\ wasps(p_i, p_j) \; otherwise & [7.2.5 - 7.2.6] \end{cases}$$

where $wasps(p_i, p_j)$ is the similarity value in the wasps database, which returns 0 if no value exists. The function returns a partially ordered set $\mathcal{PO}$ of *pairs* of properties, such that for any two pairs P_i and P_j, $P_i \leq P_j \leftrightarrow \sigma(P_i) \leq \sigma(P_j)$[7.2.11]. In case $|\mathcal{R}| = 1$, the creation of the preference order cannot be based on similarity of referents. Instead, the algorithm calculates, for each property $p \in \mathcal{P}$, the discriminatory value of p. We discuss this more fully in Section 4.

3.2 Function `distinguishReferents`

The content-selection function takes as input $\mathcal{PO}$ and $\mathcal{R}$, and has available C_{r_i}, D_{r_i} and UD_{r_i} for each $r_i \in \mathcal{R}$. For each element p of each pair of properties $\langle p_i, p_j \rangle \in \mathcal{PO}$, the function checks, for each $r_i \in \mathcal{R}$, whether $p \in C_{r_i}$. If so, it checks further whether p removes at least one distractor for r_i, i.e. $\|p\| - D_{r_i} \neq \emptyset$ [7.3.3]. If so, it updates D_{r_i} [7.3.4] and intersects $\{p\}$ with UD_{r_i} [7.3.5]. Since $\mathcal{PO}$ contains properties of pairs of referents, at any given step in this process, at least two referents are being treated in succession. Moreover, the process attempts to maximise coherence by looking at highly similar pairs first. At each step, the function checks whether each referent has been distinguished, i.e. $D_{r_i} = \emptyset$, in which case UD_{r_i} is unified to the global description $\mathcal{UD}$ [7.3.8]. The process continues until all referents have been distinguished, when the function returns $\mathcal{UD} = UD_{r_1} \cup UD_{r_2} \cup \ldots \cup UD_{r_n}$, a description in disjunctive normal form. In case not all referents are distinguishable, $\mathcal{UD}$ is returned as the best approximation given the available information. We call $\mathcal{UD}$ an *underspecified description*, in that it distinguishes the set of intended referents, but does not specify precisely what form the final description will take, a decision which is deferred to the aggregation step.

In summary, reference to a set is modelled as a process of distinguishing each entity in the set by constructing a conjunction (intersection) of properties. Note that, since $\mathcal{D}$ is defined as the set of all entities *not* in $\mathcal{R}$, this process also covers the case where two referents can only be distinguished as a set, but not separately. This occurs, for instance, where two referents have identical properties which distinguish them from the rest of the elements in $\mathcal{DOM}$.

3.3 A worked example

As an example, we will consider the following simple domain

e_1	postgraduate	physicist
e_2	undergraduate	greek
e_3	italian	postgraduate

Assume that $\mathcal{R} = \{e_1, e_2\}$. The sets C_{e_1} and C_{e_2} are as indicated by the corresponding rows in the above matrix, while $D_{e_1} = \{e_3\}$ and $D_{e_2} = \emptyset$, since the only distractor, e_3, does not share any properties with e_2. The function

[4] Cf. [Siddharthan and Copestake, 2004] for a related approach. The Siddharthan and Copestake algorithm is motivated by the problem of tasks involving re-generation (e.g. text summarisation) where the input is open text, focusing on generating minimal singular descriptions. As a result, they too use 'words', rather than $\langle A, V \rangle$ pairs.

`orderBySimilarity` compares each property of e_1 and e_2, resulting in the following preference order (corresponding values are indicated):

1	postgraduate	undergruaduate	0.0767
2	physicist	undergraduate	0.0436
3	physicist	greek	0.0253
4	postgraduate	greek	0

The algorithm now proceeds along the preference ordering, considering each pair of properties until each referent is distinguished from e_3. From the first pair of properties in $\mathcal{PO}$, *postgraduate* is not useful for e_1, as it does not eliminate e_3, nor for e_2, which doesn't have the property. The next property in the pair, *undergraduate* belongs to e_2 and eliminates e_3 for this referent; it is added to UD_{e_2}. At this point, only e_1 remains to be distinguished. The algorithm now moves to the next pair in $\mathcal{PO}$. Here, *physicist* belongs to e_1 and excludes e_3; hence, it is added to UD_{e_1}. This terminates the search process, as both entities are uniquely distinguished. At this point $\mathcal{UD} = \Big\{\{undergraduate\}_{e_2}, \{physicist\}_{e1}\Big\}$. The resulting description could be realised as *the undergraduate and the physicist*.

The preference order would be quite different for a different $\mathcal{R}$, say $\{e_2, e_3\}$. Here, the properties *greek* and *italian* would be ranked highest. This is one of the strengths of a dynamic ordering approach. A consequence of using the WASPS thesaurus is that only properties belonging to the same grammatical category can be compared, since the thesaurus separates nouns, verbs and adjectives. If, in addition to the properties used in the above domain, there were, say, the properties e_1 : *tall*, e_2 : *short* and e_3 : *chubby*, it is guaranteed that the pair $\{tall, short\}$ would be ranked as a related pair much higher than, say, the pair $\{tall, undergraduate\}$. As a result, the algorithm would consider $\{tall, short\}$ first, increasing the likelihood that the properties selected for the set of referents would be related.

3.4 Aggregation

As shown above, $\mathcal{UD}$ is a description in disjunctive normal form. It can be represented as a a sparse matrix in which each column corresponds to one of the properties, and each row corresponds to some $r_i \in \mathcal{R}$. For example, suppose that $\mathcal{R} = \{e_1, e_2, e_3\}$ and that, at the end of the content selection procedure, $\mathcal{UD} = \Big\{\{p_1, p_2, p_3\}_{e_1}, \{p_1, p_3\}_{e_2}, \{p_3\}_{e_3}\Big\}$. This can be represented as follows:

e_1	p_1	p_2	p_3
e_2	p_1		p_3
e_3			p_3

The output could be $(p_1 \wedge p_2 \wedge p_3) \vee (p_1 \wedge p_3) \vee p_3$, which can be further simplified, for instance using the Quine-McKluskey procedure. Here, we leave aside the logical problem of simplification, taking a perspective which emphasises coherence. Once again, we use semantic similarity to guide aggregation decisions. As shown in Experiments 2 and 3, speakers evinced a preference for plural referring expressions containing descriptors with a similarity value above a threshold. Here, we use this insight, proposing that UD_{r_i} and UD_{r_j} should be unified if there is at least one pair of properties $\langle p_i, p_j \rangle$ $p_i \in UD_{r_i} \wedge p_j \in UD_{r_j}$ which are sufficiently similar to eachother to guarantee a coherent description.

The aggregation function uses conditional probabilities to make the decision of whether all or a part of the content in $\mathcal{UD}$ can be realised as a single plural referring expression, or whether it is more appropriate to break up the logical form into different sub-expressions. This is implemented as a direct application of the data gathered in Experiments 2 and 3, since the experiment was based on an explicit choice of what to aggregate in a linguistic conjunction versus what to refer to separately. The conditional probability we are interested in is $P(aggr|\sigma)$, the probability of producing a plural conjoined referring expression given the similarity value σ, relative to $P(\neg aggr|\sigma)$, the likelihood of choosing to refer using separate NPs. From each experimental trial in experiments 2 and 3, we used the similarity value of the two nouns used in the plural referring expression of Sentence 1. Subsequently, we calculated the similarity between one of the nouns (randomly selected) in Sentence 1 and the noun in Sentence 2, taking this to be the similarity value for a pair of non-aggregated descriptors. This resulted in a corpus of noun pairs and their corresponding similarity values ($N = 675$), of which approximately half had been aggregated as definite NP conjunctions and half had not. The similarity values were split into five intervals, with $0.05 \leq \sigma$ the highest interval. For each interval, the conditional probability $P(aggr|\sigma)$ and $P(\neg aggr|\sigma)$ was calculated.

The aggregation function recurses through the rows of $\mathcal{UD}$, checking whether for at least one property in row_n and another in row_{n+1}, $P(aggr|\sigma) > P(\neg aggr|\sigma)$[5]. If so, the two rows are aggregated (i.e. properties within rows are conjoined and the two rows disjoined), and the algorithm proceeds to the next row, comparing this to the previously aggregated logical form. In case $P(aggr|\sigma) < P(\neg aggr|\sigma)$, the properties within a row are conjoined but not aggregated with the previous rows. The process can be repeated to create a separate aggregate for the remaining rows, until all rows have been accounted for. Rows which have not been aggregated, i.e. for which $P(aggr|\sigma) < P(\neg aggr|\sigma)$, are realised as separate sub-expressions.

As an example, suppose the underspecified description returned by `distinguishReferents` were the following.

e_1	postgraduate
e_2	academic
e_3	undergraduate

The algorithm begins by comparing row 1 and row 2. The similarity between the two descriptors *postgraduate* and *academic* is 0.0389. According to the experimental data, $P(aggr|\sigma)$ for this value (actually, for values falling in the interval $0.03 \leq \sigma \leq 0.04$) is 0.017, whereas $P(\neg aggr|\sigma) > 0.8$. Hence, the two rows are not aggregated. The algorithm next compares row 1 to row 3. In this case, with $\sigma = 0.0767$, $P(aggr|\sigma) = 0.7$, which is greater than the value for $P(\neg aggr|\sigma)$. Hence the two are aggregated. At this

[5]The similiarity values of property pairs have already been calculated in the `orderBySimilarity` function.

point, rows 1 and 3 have been aggregated while row 2 has not, yielding the following output:

- LF1 = *academic*
- LF2 = *postgraduate* $\vee$ *undergraduate*

These logical forms could be realised as:

- NP1 = *the academic*
- NP2 = *the postgraduate and the undergraduate*

This approach assumes that when a GRE algorithm is called during the NLG process, the output need not necessarily be a single referring expression or NP. Indeed, based on the findings presented in Section 2, we would argue that there are cases where it is better not to produce a single plural referring expression, as this would sacrifice coherence. The approach also raises interesting possibilities for realisation. GRE is usually assumed to belong to the microplanning stage in the NLG architecture. At this point, the system is in the process of fleshing out the contents of a message. Presumably, a referring expression for some set $\mathcal{R}$ will be embedded within a sentence minimally containing a VP. In case more than one referring expression is generated, as in the above example, realisation options include generating multi-part messages related by function words such as *also* and *too*. For example, suppose the sentence being generated were $\{e_1, e_2, e_3\}$ *be in the room*, and the output were the two LFs given above. Realisation options would include sentences such as *The academic was in the room. So were the postgraduate and the undergraduate.*.

4 Remarks on complexity

In this section, we make a few remarks about the complexity of the content-selection algorithm. The function `distinguishReferents` simply traverses $\mathcal{PO}$ and tests, for each property, whether it is useful for some element of $\mathcal{R}$. Thus, letting $n = |\mathcal{R}|$ and $m = |\mathcal{PO}|$, it has a worst case complexity in the order of $n \cdot m$.

Clearly, the main source of complexity lies in the pairwise comparison of property sets of intended referents in the construction of the preference order. For a given pair $\{r_i, r_j\}$, `orderBySimilarity` compares the sets $\mathcal{C}_{r_i}$ and $\mathcal{C}_{r_j}$, a quadratic operation. On the other hand, this process is repeated for all 2-subsets of $\mathcal{R}$, giving it an overall complexity in the order of

$$k^2 \cdot \binom{n}{2} \equiv \frac{n! \times k^2}{2(n-2)!}$$

where k is the number of properties in C_r. The complexity of the function therefore grows the greater the number of intended referents. It is plausible to assume that efficiency will be reasonable in the average case. For instance, in the corpus of definite descriptions from which the sample in Experiment 1 was taken, NPs with more than two conjuncts are seldom attested. On the other hand, that experiment did not consider NPs realised as morphological plurals (e.g. *the boys in the room*), which presumably tend to arise where several entities with shared properties formed a distinct group (for instance, all the entities are in a well-defined spatial region which distinguishes them from all distractors). Our algorithm has no way of identifying such groups of entities, except by comparing their properties in the way outlined. We return to this issue briefly in the Conclusion.

5 Extensions

Our algorithm is still prone to redundancy, which arguably is not always desirable. An extension to the current algorithm has been implemented which slightly alters the construction of the preference order. In this version, the *satellite set* for each property is calculated at the outset. The satellite set, as proposed by [van Deemter and Halldórsson, 2001], contains the set of entities that falls in the extension of the property. During the call to `orderBySimilarity`, once the similarity value has been calculated for a property pair, the number of distractors falling in the extension of each property in the pair is also calculated as the difference between the satellite set of the property and the set of intended referents. We call the cardinality of the resulting set the *distraction value*, DV of a property. The similarity value of the property pair is scaled by dividing by the highest DV in the pair. As a result, the higher the DV of some property in a pair, the lower the value associated with it by the function, modulo the similarity value. This increases the likelihood that highly valued properties have a low DV. In case $|\mathcal{R}| = 1$, the preference order is simply constructed in terms of the DV's of the properties of r, since calculating the similarity between referents is excluded.

This approach is related to the approach of Siddharthan and Copestake [2004], where a preference order is constructed by calculating, for each property of an intended referent, a *Contrastive Quotient*, which indicates the extent to which distractor properties are unrelated to the referent property, and a *Similarity Quotient*, which indicates the opposite trend. These are calculated by looking up distractor properties in the WordNet synonym and antonym sets of the referent property. The *Discriminating Quotient* is the difference between the two. This approach works for the singleton case, which is the focus of their work. In the case of sets, our approach could be interpreted as an attempt to balance two forces, trying to maximise similarity between intended referents in $\mathcal{R}$ and maximise the difference between the referents and their distractors.

6 Conclusions and future work

We have described an approach to generation which focuses on maximising the coherence of a description, by considering the common or similar properties of intended referents. This approach was motivated empirically, on the basis of corpus and experimental data. Our algorithm distinguishes between the selection of properties and their aggregation into logical forms, avoiding the combinatorial problem of testing for the usefulness of disjunctions of various lengths in the search process. On the other hand, the combinatorial problem faced by most GRE algorithms dealing with sets resurfaces to some extent in the dynamic preference ordering. The scope of the approach presented here is limited in two respects. First, we have only discussed distributive reference. This raises the

more general problem of representing such properties, which take sets, not singletons in their extension (but see [Stone, 2000]). Second, as noted in Section 4, our algorithm deals with distributive properties that create salient groups of entities by virtue of being shared, since such properties are assigned a similarity value of 1 by `orderBySimilarity`. However, this approach may incur considerable overhead in the case of salient groups with large cardinalities, since the algorithm undertakes a pairwise comparison.

In further work, we plan to extend our approach in two directions. First, we are furthering the empirical work, looking at perceptual factors involved in grouping during reference tasks performed by human speakers. This is being carried out both in production and *identification* tasks, in order to develop heuristics for ensuring that referring expressions are easy to understand. Second, we are also looking at alternative ways of addressing the problems raised here computationally. In particular, a constraint-satisfaction approach to GRE seems promising, especially when addressing the issue of brevity versus coherence discussed in Section 5. From that perspective, the GRE problem can be defined as one of emphasising the properties that group entities into a conceptual gestalt, and maximising their distinctiveness from other entities in the domain. A constraint-satisfaction approach may also be useful for overcoming the limitations pointed out here, with respect to collective properties [Stone, 2000] and perceptual grouping via shared distributive properties. Promising work on GRE and perceptual grouping has been undertaken by Funakoshi *et al.* [2004], applying the reference resolution algorithm proposed by Thorisson [1994] to generation. In future work, we plan to tackle the problem of perceptual grouping more directly.

7 Appendix. Formalisation of the algorithm

In this section, we give a formalisation of the algorithm presented in Section 2, couched in set-theoretic terms. Numbered steps in each subsection correspond to steps cross-referenced in Sections 3.1 and 3.2 respectively.

7.1 Initialisation of data structures

1: $\mathcal{DOM} \leftarrow \{e_1, e_2, \ldots, e_n\}$
// The domain of entities.
2: $\mathcal{R} \leftarrow \{r_1, r_2, \ldots, r_n \mid r_i \in \mathcal{DOM}\}$
// The set of intended referents.
3: $\mathcal{PROP} \leftarrow \{p_1, p_2, \ldots, p_n\}$
// The set of properties in the domain.
4: $\mathcal{P} \leftarrow \{p_1, p_2, \ldots, p_n \mid p_i \in \mathcal{PROP} : \exists r \in \mathcal{R}[r \in \|p_i\|]\}$
// The set of relevant properties.
5: $\mathcal{D} \leftarrow \mathcal{DOM} - \mathcal{R}$
// The set of distractors.
6: **for all** $r_i \in \mathcal{R}$ **do**
7: $C_{r_i} \leftarrow \{p_1, p_2, \ldots, p_n \mid p_i \in \mathcal{P} : r_i \in \|p_i\|\}$
// Initialise the property set C_{r_i} for each referent r_i.
8: $D_{r_i} \leftarrow \{d_i, d_j, \ldots, d_n \mid d_i \in \mathcal{D} : \exists p_i \in C_{r_i}[d_i \in \|p_i\|]\}$
// Initialise the distractor set for each $r_i \in \mathcal{R}$.
9: $UD_{r_i} \leftarrow \emptyset$
// Initialise the description for r_i.
10: **end for**
11: $\mathcal{PO} \leftarrow \emptyset$
// Initialise preference order $\mathcal{PO}$ to the empty set.
12: $\mathcal{UD} \leftarrow \emptyset$
// Initialise underspecified description $\mathcal{UD}$.

7.2 Function `orderBySimilarity`

Require: $\mathcal{R} \neq \emptyset$
Require: $\mathcal{P} \neq \emptyset$
Require: $C_{r_i} \neq \emptyset$ forall $r_i \in \mathcal{R}$
1: **for all** $\{r_i, r_j\} \in \mathcal{R}^2$ **do**
2: **for** $\{p_i, p_j \mid p_i \in C_{r_i} \wedge p_j \in C_{r_j}\}$ **do**
3: **if** $p_i = p_j$ **then**
4: $\sigma(p_i, p_j) = 1$
5: **else if** $p_i \neq p_j$ **then**
6: $\sigma(p_i, p_j) = wasps(p_i, p_j)$
7: **end if**
8: $\mathcal{PO} \leftarrow \mathcal{PO} \cap \{\langle p_i, p_j \rangle\}$
9: **end for**
10: **end for**
// Calculate similarity between properties and add property pairs to $\mathcal{PO}$.
11: $\mathcal{PO} = \langle A, \leq \rangle$ such that:
$\forall P \in A$:
$P = \langle p_i, p_j \rangle : p_i \in C_{r_i} \wedge p_j \in C_{r_j} \wedge r_i \neq r_j$
$P_i \leq P_j \leftrightarrow \sigma(P_i) \leq \sigma(P_j)$
// $\mathcal{PO}$ is a poset of 2-subsets of $\mathcal{P}$, ordered by similarity.
12: **return** $\mathcal{PO}$

7.3 Function `distinguishReferents`

Require: $\mathcal{R} \neq \emptyset$
Require: $\mathcal{PO} \neq \emptyset$
Require: $\mathcal{UD} = \emptyset$
Require: $\mathcal{PO} = \emptyset$
1: **for all** $\langle p_i, p_j \rangle \in \mathcal{PO}$ **do**
2: **for** $r_i \in \mathcal{R}$ **do**
3: **if** $p_i \in C_{r_i} \wedge \|p_i\| - D_{r_i} \neq \emptyset$ **then**
4: $D_{r_i} \leftarrow D_{r_i} \cap \|p_i\|$
// Remove distractors for r_i.
5: $UD_{r_i} \leftarrow UD_{r_i} \cap \{p_i\}$
// Update description for r_i.
6: **end if**
7: **if** $D_{r_i} = \emptyset$ **then**
8: $\mathcal{UD} \leftarrow \mathcal{UD} \cup \{UD_{r_i}\}$
// Update the underspecified description $\mathcal{UD}$.
9: **end if**
10: **end for**
11: **end for**
12: **return** $\mathcal{UD}$

References

[Arts, 2004] Anja Arts. *Overspecification in Instructive Texts*. PhD thesis, Univiersity of Tilburg, 2004.

[Bard *et al.*, 1996] E. G. Bard, D. Robertson, and A. Sorace. Magnitude estimation of linguistic acceptability. *Language*, 72, 1996.

[Barry *et al.*, 1997] C. Barry, C. M. Morrison, and A. W. Ellis. Naming the snodgrass and vanderwart pictures: Effects of age of acquisition, frequency and name agreement. *Quarterly Journal of Experimental Psychology*, 50A(3):560–585, 1997.

[Dale and Reiter, 1995] Robert Dale and Ehud Reiter. Computational interpretation of the gricean maxims in the generation of referring expressions. *Cognitive Science*, x:8, 1995.

[Damian *et al.*, 2001] M. F. Damian, G. Vigliocco, and W. J. Levelt. Effects of semantic context in the naming of pictures and words. *Cognition*, 81:B77–B86, 2001.

[Damian, 2000] M. F. Damian. Semantic negative priming in picture categorisation and naming. *Cognition*, 76:B45–B55, 2000.

[Dell, 1986] G. S. Dell. A spreading activation theory of retrieval in sentence production. *Psychological Review*, 93:283–321, 1986.

[Ford and Olson, 1975] W. Ford and D. Olson. The elaboration of the noun phrase in children's object descriptions. *Journal of Experimental Child Psychology*, 19:371–382, 1975.

[Funakoshi *et al.*, 2004] K. Funakoshi, S. Watanabe, N. Kuriyama, and T. Tokunaga. Generating referring expressions using perceptual groups. In *Proceedings of the 3rd International Conference on Natural Language Generation, INLG-04*, 2004.

[Gardent, 2002] Clare Gardent. Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-2002*, 2002.

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

[Horacek, 2003] H. Horacek. A best-first search algorithm for generating referring expressions. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, EACL-2003.*, 2003.

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

[Kilgarriff and Tugwell, 2001] A. Kilgarriff and D. Tugwell. Word sketch: Extraction and display of significant collocations for lexicography. In *Proceedings of the Collocations Workshop in Association with ACL-2001.*, 2001.

[Levelt, 1989] W. M. J. Levelt. *Speaking: From Intention to Articulation.* MIT Press, 1989.

[Levelt, 1998] W. J. Levelt. Models of word production. *Trends in Cognitive Science*, 3:223–232, 1998.

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

[Meyer, 1996] A. S. Meyer. Lexical access in phrase and sentence production: Results from picture-word interference experiments. *Journal of Memory and Language*, 35:477–496, 1996.

[Minnen *et al.*, 2001] G. Minnen, J. J. Carroll, and D. Pearce. Applied morphological processing of english. *Natural Language Engineering*, 7(3):207–223, 2001.

[Pechmann, 1989] Thomas Pechmann. Incremental speech production and referential overspecification. *Linguistics*, 27:89–110, 1989.

[Pederson *et al.*, 2004] T. Pederson, S. Patwardhan, and J. Michelizzi. Wordnet::similarity — measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence, AAAI-2004.*, 2004.

[Resnik, 1995] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI-1995.*, 1995.

[Rock, 1983] I. Rock. *The Logic of Perception.* MIT Press, 1983.

[Roelofs, 2000] A. Roelofs. Weaver++ and other computational models of lemma retrieval and word form encoding. In L. Wheeldon, editor, *Aspects of Language Production.* Psychology Press, UK, 2000.

[Schriefers and Pechmann, 1988] H. Schriefers and T. Pechmann. Incremental production of referential noun phrases by human speakers. In M. Zock and G. Sabah, editors, *Advances in Natural Language Generation*, volume 1. London: Pinter, 1988.

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

[Snodgrass and Vanderwart, 1980] J. G. Snodgrass and M. Vanderwart. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):174–215, 1980.

[Sonnenschein, 1982] S. Sonnenschein. The effects of redundant communication: When less is more. *Child Development*, 53:717–729, 1982.

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

[Thorisson, 1994] K. R. Thorisson. Simulated perceptual grouping: An application to human-computer interaction. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society.*, 1994.

[van Deemter and Halldórsson, 2001] K. van Deemter and M. Halldórsson. Logical form equivalence: The case

of referring expressions generation. In *Proceedings of the European Workshop on Natural Language Generation, ENLG-01*, 2001.

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

[Vigliocco *et al.*, 2002] G. Vigliocco, M. Lauer, M.F. Damian, and W. J. Levelt. Semantic and syntactic forces in noun phrase production. *Journal of Experimental Psychology: Human Perception and Performance*, 28:46–58, 2002.

[Wertheimer, 1938] M. Wertheimer. Laws of organization in perceptual forms. In W. Ellis, editor, *A Source Book of Gestalt Psychology.* Routledge & Kegan Paul, 1938.