An Empirical Study of the Communicative Goals Impacting Nominal Expressions

Pamela W. Jordan
Intelligent Systems Program
University of Pittsburgh
jordan@isp.pitt.edu

Contents

1 Introduction

When expressing our thoughts about things in language, we must choose linguistic forms referring to the objects. The problem of choosing such forms is the problem of *generating referring expressions.* Computational work on this topic ([App85, Kro86, Dal92, HH95, Loc95, Rei90, Pas96] has concentrated on how to produce a minimally complex expression that singles out the referent from a contextually determined set of alternative referents. According to these approaches, a descriptor containing information that is not needed to single out the referent would not be minimally complex.

This general approach can be seen as an implementation of the two parts of Grice's *maxim of Quantity*, according to which an utterance should both say as much as is required, and no more than is required [Gri75]. Although such models of Grice's Quantity maxim seem plausible from a theoretical standpoint, recent work on naturally occurring speech has produced compelling evidence that they are overly simplistic [Wal93]. With the goal-directed view of sentence generation, we assume that speakers attempt to satisfy multiple goals with each utterance [App85]. This could be accomplished in many ways. The one we will consider here is a speaker focusing on one goal but opportunistically considering others when faced with choices about how to satisfy that goal. This approach overloads intentions so that one action can contribute to many goals [Pol91]. Overloaded intentions has been used in computational linguistics to show how choices at one level can address multiple goals and allow one to leave some important information unexpressed at another (e.g. clausal connectives at the pragmatic level [DW96] and word choice at the syntax and semantics levels [SW98]).

In this paper, we describe our attempts to empirically identify the degree to which intention overloading may affect content selection for object descriptions and what communicative goals might overload with an identification goal. We limit our study to cases in which references are made to a discourse entity that has already been previously mentioned. We assume that the primary communicative goal of such descriptions is to identify the intended referent. It seems plausible then that any information that is redundant for identification purposes could be present either because it is a cognitive processing artifact [DR95, Pas96, Lev89] or because it helps fulfill some other communicative goal besides identification.

2 The COCONUT Corpus

Our empirical investigations are based on the COCONUT corpus [DJTM99].[1] This corpus contains 24 computer-mediated design dialogues in which two people collaborate on a simple design task, buying furniture for two rooms of a house. The information needed to complete the design task is divided between the two designers in such a way that a good design cannot be achieved without collaboration. With this task, the designers typically describe the furniture items that they believe are relevant to the current subtask and design constraints. It should be noted that, in general, design tasks often require the designers to adjust their problem solving constraints in order to arrive at an agreeable solution [LS97, Lyo95].

The COCONUT task is related to those described in [Wal93, WGR93] but differs in the emphasis and complexity of the task.[2] Each of the two participants in the task is given a separate budget and inventory of furniture that lists the quantities, colors, and prices for each

[1] See http://www.isp.pitt.edu/~intgen/coconut.html.

[2] Walker's similar task is performed by two artificial agents whereas our task and that in Whittaker et.al. is performed by two humans. Whittaker et.al.'s dialogues are spoken whereas ours are written.

item in that inventory.[3] Neither participant knows what is in the other's inventory or the money that the other has. The participants have the same types of knowledge but different instantiations of it. By sharing information, the participants can combine their budgets and can select furniture from each other's inventories. Purchasing decisions are joint: they must be mutually known and approved. The participants are equals in that there is no master-slave or expert-client relationship. Both participants have been briefed on the task goals, incentives and the tools and have had no prior contact.

The participants' main goal is to negotiate the purchases; the items of highest priority are a sofa for the living room and a table and four chairs for the dining room. The participants also have specific secondary goals which further complicate the problem solving task. Participants are instructed to try to meet as many of these goals as possible, and are motivated to do so by associating points with satisfied goals.[4] The secondary goals are: 1) Match colors within a room, 2) Buy as much furniture as you can, 3) Spend all your money.

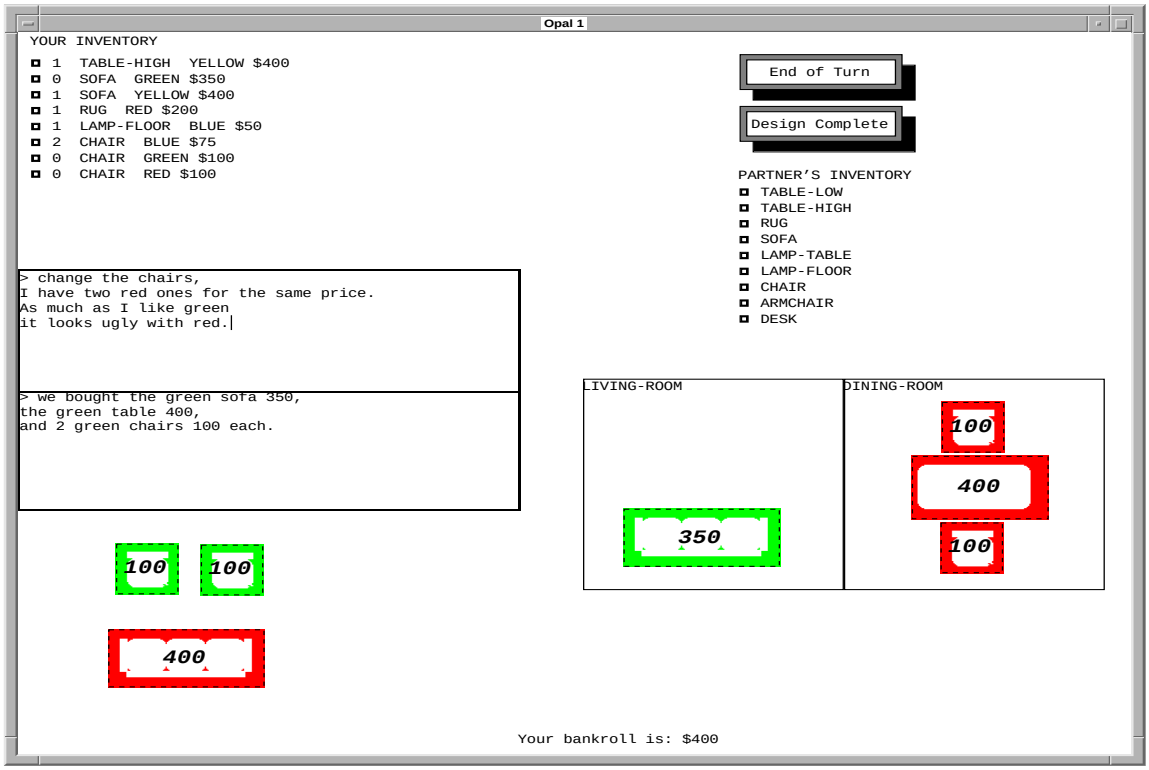

Figure 1: A View of the COCONUT Interface

The participants are in separate rooms and can communicate via the computer interface only. They are asked to maintain private graphical representations of their discussions and incremental agreements. The participants share dialogue windows but the inventories, budgets and updated floor plans are private and appear only on the owner's color display. Figure 1 shows the interface as it looks in the middle of a design session.

The buttons in the upper right corner of Figure 1, "End of Turn" and "Design Complete", enforce turn-taking and initiate the incremental recording of the conversation and the graphics updates. No interruption of the partner's turn is allowed. Also note that only the participants'

[3] In Walker's task this information is committed to memory but in our task the participants have this information in written form.

[4] In Whittaker et.al.'s task the incentives and goals are simpler.

current turns are available, i.e., the turn being currently held in the top dialogue box and the partner's previous turn in the bottom one.

During an incremental recording, the most recently transmitted message is recorded as well as the state of the sender's graphics display. The graphics display record is a description of the furniture icons in the two rooms as well as those that have been created but not assigned to any room. The participants incrementally update the floor plan by placing the furniture icons in meaningful locations. Whenever possible we have used this private information in our corpus analysis as partial evidence of what the speaker's utterance meant and what the hearer understood. However, the primary purpose of the graphics display is as a memory aid for the participants and is only intended secondarily to help clarify possible sources of misunderstanding during analysis.

Note that since a participant does not know what furniture his partner has available, there is a menu (see the mid-right section of the display in Figure 1) that allows a participant to define furniture icons that represent what he understands his partner to have as his partner shares this information with him. There is nothing to prevent the participant from creating an icon for a piece of furniture the partner does not actually have since the menu is general. An icon for a non-existent item could result from either a misunderstanding of his partner's item description or an error in selecting feature values for the item. At minimum the participant must know the type of the furniture item (e.g. chair, table). If the participant does not know or is uncertain about any of the other feature values of the furniture item, he can leave that feature unspecified (i.e. color and purchase price).

The participants first worked through a trial problem to familiarize themselves with the task and the communications setting. During this time they could ask for guidance on using the interface and clarification of the goals and incentives. The participants then solved 1-3 scenarios where the inventories and budgets vary. The problem scenarios ranged from ones where items are inexpensive and the budget is relatively large to ones where the items are expensive and the budget relatively small.

Nothing intrinsic to this design task should result in unusual object descriptions. It is reasonable to assume that design tasks, and the COCONUT task in particular, should not affect the number of coreferences. While we will claim that this specific task does lead to the inclusion of identificationally unneeded properties in descriptions, we expect that this should hold for a wide range of tasks where many object properties are relevant to the problem solving task and where the definition of success is also negotiable.

Because of the non-interruptible setting of the dialogues, attentional limits may also cause redescriptions to be longer than they would otherwise be. Since there will be no interactive feedback in this setting, we expect more instances of overlooked or forgotten objects [JT96].

3 The Empirical Study

During our preliminary investigations of the COCONUT corpus, we noticed that there seemed to be a large number of subsequent references to furniture items using highly redundant descriptions. This observation includes any explicit information in an utterance that helps define a discourse entity. For example, if a discourse entity is mutually known to have the color *red*, then including *red* in the utterance, as with "My chair is red.", makes the description of the discourse entity redundant.

To confirm our initial impressions, we must first determine for each description, what other mutually known items may be salient for the dialogue participants. Following the terminology of [Dal92], we will call these salient, mutually known items, the distractors. Since it is not clear

what definition of the distractor set is correct, we take it as a matter of empirical investigation to try several plausible definitions that relate to current theories in computational linguistics and psycholinguistics. Using several definitions of the distractor set, we can see how much redundancy exists with each definition. If there is a high degree of redundancy no matter what the distractor set definition is, then it is plausible given our current theories that the coreference descriptions do indeed have a high degree of identificational redundancy.

Grosz and Sidner's theory of the attentional and intentional structure of discourse [GS86] provides an account of saliency that is widely accepted. In their theory, the content of a focus space and a stack of focus spaces is defined by the task structure. A change in task or topic indicates the start of a new discourse segment. All of the discourse entities described in a discourse segment are classified as salient for the dialogue participants while that focus space is on the focus stack. The relative saliencies between focus spaces are left unspecified and are still to be determined.

Passonneau [Pas96] uses Grosz and Sidner's theory to define a distractor set. Her distractor set is the union of all the discourse entities in the current discourse segment and all the entities in the last segment that contained the entity to be described. To be conservative, she assumes that if the most recent segment containing the target entity is not the same as the current segment then it is a resumption and the intervening focus spaces should not be included in the distractor set. The descriptive content that is needed to avoid ambiguity and the size of the distractor set are positively correlated. So Passoneau's model, which minimizes the distractors, will also provide a conservative measure of the number of redundant redescriptions in a corpus.

Using this definition for the distractor set, Passonneau found that only 6% of the non-pronominal noun phrases (NPs) in the Pears corpus contained redundant identificational information. We will call this distractor set definition **SEG** and use it to confirm our initial impressions of a high degree of identificational redundancy in the COCONUT corpus. Following Passonneau, we first identified all the NPs that were potentially overspecified by selecting subsequent descriptions that are longer than their previous descriptions. We then filtered this subset of NPs using the **SEG** distractor set definition. With the first step, we found that 49% (96 of 196) of the full NPs in the COCONUT corpus were potentially overspecified. And after filtering these with **SEG**, we found that 31% (60 of 196) of all the non-pronominal NPs were identificationally redundant. This seems to confirm our initial impressions. However, since it is possible that **SEG** is not the best distractor set definition for all genres, we will also try other definitions in the rest of our empirical studies of the corpus.

Several hypotheses based on cognitive processing limitations have been suggested by psychological and computational linguistics research to explain redundancy in object descriptions. Studies by [Deu76, Man86, Son82, Son84] show that it is easier for listeners to identify an overspecified referent than a minimally specified one. Levelt explains these findings by supposing that listeners tend to create a *gestalt* search template for the target object and that an overspecified template makes it easier to search for the referent [Lev89]. For example, searching a group of visible items for "big black bird" is assumed to be easier than "black one" or "black bird" even if the shorter expression uniquely identifies the referent. This could also explain why there is a preference to include a noun related to the type of the target object in the description even when it is redundant. Levelt also claims that the distractor objects should be irrelevant for whether the gestalt template is used. In our empirical testing, we will compare the performance of a selection algorithm, **gestalt**, that is based on this hypothesis, against what the humans did in the COCONUT corpus.

Lexical focus [Pas96] provides another possible explanation for some cases of observed redundancy. This explanation is motivated by the observation that speakers have a tendency

to repeat the last description for the target object in a redescription. While Passonneau doesn't cite any experimental studies to support this observation, Clark and Wilkes-Gibbs' findings in experiments with a Tangram identification task [CWG86] may provide some evidence. They found that between trials, the length of the descriptions get shorter as the participants establish common ground and come to settle on a particular description for each figure. In our empirical testing, we will compare the performance of a selection algorithm, **lex**, that is based on this hypothesis, against what happens in our corpus.

Finally, the unnecessary or redundant information could merely be a side-effect of processing limitations associated with the communicative goal of identification. In making this suggestion, [DR95] point out that it is computationally intractable and therefore psychologically implausible to attempt to find minimal descriptions. This implies that people may use some combination of heuristic approaches that are along the lines of what we just discussed above. Dale & Reiter [DR95] propose a computationally tractable algorithm, **IDAS**, that attempts to capture some of the hypotheses suggested above. This means that it should serve as a good representative approach for identificational content selection. We will compare the output of this algorithm to the COCONUT corpus to see how well it explains the identificational redundancy. If it is not a good match, then this will lend some support to the idea that goal overloading may be affecting the content selection of object descriptions.

To illustrate a possibility for goal overloading in the COCONUT domain, assume that there is an initial constraint setting to match colors. Also assume that the person about to speak has just discussed using a red table and prior to that had introduced tables of a variety of colors and four $100 red chairs, and four $75 green chairs. Finally, assume that she has just decided that she wants to drop the color match constraint so that they can use the cheaper $75 chairs. When she communicates her suggestion that they use these chairs, we hypothesize that she will be less likely to say *the $75 chairs* although it adequately and economically identifies the target chairs and the price of the chairs are highly salient for her. Instead we expect her to say *the green chairs* or *the $75 green chairs.* By choosing "green," she still manages to adequately identify the target chairs while also enabling the hearer to more easily infer that she intends to drop the color match constraint. She has saved the expense of having to explicitly say *Let's forget about matching colors.* [Wal93] and has reduced the risk of the hearer not making the inference [Car92] and thinking there is a mistake in his understanding of the color of the $75 chairs. Alternatively, one might argue that the color property is by default highly salient and would generally be included in the description. However, we found that color is not always included in a full NP redescription even when it rules out some distractors under the **SEG** definition of distractors (43 of 196).

During our initial investigations of the COCONUT corpus, we observed that as the human designers adjusted problem solving constraints, as with the color match constraint in the above example, they rarely talked directly about these adjustments. Given the results of [Wal93], it seems plausible that the redundant property information may be making it easier to infer changes to the problem definition so that it is not necessary to directly communicate the changes. We hypothesize that the inclusion of some redundant property information is related to these problem solving adjustments.

We can also ask how changes in the problem state are related to property saliency. So far, computational models have only used static property saliency hierarchies to guide the choice of which properties to include in an expression and have assumed that dynamic saliency hierarchies that adjust to the context are necessary for deciding which properties to prefer in cases where there are several descriptions of similar length that would uniquely identify the target object.

4 Analysis of the COCONUT corpus

There are two purposes for which we use the COCONUT corpus in our empirical studies. First we wish to check for possible correlations between implicit constraint changes and the properties used to describe the objects in the utterances that cause such inferences. Second we wish to check the performance of algorithms that are based on the explanations for redundancy that we described earlier, as well as the performance when goals to communicate constraint changes are overloaded.

With these purposes in mind, we will need two sorts of corpus annotations: (1) utterance level annotations that capture problem solving and discourse features, and (2) discourse entity level annotations that capture the definitions and updates for discourse entities as a dialogue progresses and the properties selected to describe the discourse entities in subsequent references.

Features of type (1) are needed to indicate what constraint changes were communicated, and whether these changes were communicated explicitly or implicitly. We assumed a set of initial constraint settings that would maximize the number of points earned. In general, these initial settings held true for all of our participants since the task instructions that explained the scoring for solutions was the only common ground that the participants had at the start of the problem solving trials. Annotators were instructed to pick an appropriate constraint description from a given list whenever there was a change to that constraint from its previous setting.

In annotating the task, we also needed to identify the task structure and the discourse segments. We assumed that there were at least three and at most five component actions to be discussed: selecting four chairs for the dining room, selecting a table for the dining room, selecting a sofa for the living room, and selecting a set of optional items for the living room and dining room. We instructed the annotators to indicate what action was being addressed in each utterance by considering whether any furniture items or furniture templates being discussed in the utterance could unambiguously be related to one of these actions. Annotators were also asked to distinguish between when an action was first addressed and when the utterance continued the discussion. If the relation of the furniture item or template to actions was ambiguous, the annotators were instructed to indicate the highest level action that was unambiguous (e.g. select items for the dining room). Contiguous utterances that discussed a particular action were taken to define a discourse segment. Utterances that introduced or restarted action discussions while also continuing active discussions of other actions, were interpreted as starts of embedded discourse segments.

In addition to these problem solving features, information about the furniture entities introduced in the dialogue were also annotated. The main objective was to identify the entity being communicated and how the information about that entity was communicated. Both initial and subsequent references were annotated so that we could capture how the description of a single discourse entity developed during the course of the dialogue. By tracking the discourse entities in the dialogue we could tell when a subsequent reference to an entity might also add new information about the entity or correct erroneous information. For example in, "I have a $200 table. It is green.", entity_1 from the first utterance is ((type table)(owner A)(price 200)). The pronoun "it" in the next utterance corefers to entity_1 but also adds to it new information about the color of the object being referred to. The entity description then gets updated to ((type table)(color green)(owner A)(price 200)). These entity descriptions will serve as input to the content selection algorithms. The algorithms cannot choose to use properties that are new to an entity (i.e. not mutually known) to corefer. In matching an algorithm's selections with those made by the human, choices about whether to describe a new property are not

counted.

For the furniture entities, the annotators were asked to indicate the attribute-value pair information for each discourse entity in an utterance, and the sources for this information (e.g. from the utterance, the NP or locally inferred). Annotators were also asked to indicate whether the discourse entity was new or a coreference to a previous discourse entity and what other discourse entities the current entity might be related to. Here, some of the relevant relations include set, part-of, and class relations. Finally, we also asked the annotators to indicate which action the discourse entity was related to.

To develop and validate the annotation scheme, we conducted intercoder reliability studies using a balanced subset of the corpus. 30% of the corpus was annotated by two annotators.[5] We use the Kappa coefficient of agreement [Kri80, Car96] to assess intercoder reliability; this measure factors out chance agreement between coders. The discourse processing community uses Krippendorf's scale [Kri80] to interpret and apply the Kappa coefficient, which varies between 0 and 1. Krippendorf's scale discounts any variable with $K < .67$, allows tentative conclusions when $.67 < K < .8$, and definite conclusions when $K \geq .8$. Table 1, which shows the intercoder reliability results after 2 development iterations and one partial reconciliation meeting, suggests that all of the features are defined clearly enough so that they can be reliably annotated and used in studies.

Actions & Constraints	Introduce Actions	Continue Actions	Change Constraints	
	.897	.857	.881	
Discourse Entities	Reference Coreference	Discourse Relations	Properties	Entities to Actions
	.863	.819	.861	.857

Table 1: Kappa values for the Annotation Scheme

	Property Used	Property not Used
Implicit change	15	0
Explicit change	21	11
No changes	597	1140

Table 2: Relating Properties to Constraint Changes for Coreference

Table 2 shows that when we look at the descriptions that are coreferences to mutually known discourse entities, properties that are related to constraint changes are more likely to be used in the description than not ($\chi^2 = 41$, $p<0.001$, df=2). If the change is made explicit, related properties are still more likely to be included although not as strongly as with implicit changes. When there are no constraint changes, related properties are less likely to be included. This correlation offers some positive support for the hypothesis that a goal to communicate constraint changes could be overloading with an identification goal.

[5] One annotator's area of expertise is linguistics; the other is the author of this paper.

5 Content Selection Algorithms

We are interested in testing two things when checking the output of content selection algorithms against what happens in the corpus. First, we want to see how well the algorithms match human performance. We don't expect to achieve a perfect match since humans have speech lapses and since there may be contextual effects and effects related to general world knowledge that the algorithms don't model. Second, we want to see whether the overloading of goals to communicate constraint changes affects how well an algorithm matches human performance.

In implementing the various algorithms, there are a number of unknowns to contend with. As we mentioned earlier, we do not know the correct definition of the distractor set, although we do have several contending hypotheses about saliency. We have experimented with several definitions. Another unknown is how perceptual salience is determined for **IDAS** and **gestalt**. Here we have also experimented with a few alternatives.

5.1 The Selection Algorithms

We will test the performance of the three algorithms that we motivated earlier: **IDAS**, **lex**, and **gestalt**. In addition to these three we will also try a variant of **IDAS** that we will call **differences**. It allows us to experiment with a simple way of having the context of the dialogue influence perceptual saliency. The motivation for **differences** will be described below. Finally we will also test three extreme algorithms that will be described below, in order to fix an upper bound on poor performance.

IDAS checks an ordered list of properties and chooses to include the property if it rules out any of the distractors in the focus set. The list of properties is ordered by perceptual saliency; we expect this saliency to be task specific. While it is believed that the perceptual saliency may change as the task progresses, it is generally taken as a simplifying assumption in most implementations that this saliency remains static for the entire dialogue. By examining the frequencies with which these properties are used in the corpus, we determined the perceptual saliency hierarchy for the COCONUT domain to be; type, color, owner, price and quantity.[6]

Differences is similar to **IDAS** in that it uses an ordered list of properties but the selection criterion is different. **Differences** selects the property if the target object's property value is different from the most salient value for this property in the distractor set. The first few selections will be just like IDAS but overall it is more likely to include more properties than IDAS since prominence for each property is based upon the entire distractor set.

The **differences** selection criterion is meant to capture the idea of perceptual prominence observed in experiments by [CSB83]. The experimenters showed one of two photographs to students and asked them to describe the color of the flowers in the photograph. In the two photographs there were four sets of flowers and each set was of a different color. In one photograph the daffodils were made more prominent and in the other nothing was prominent. With the first photograph, the students more frequently described the color of the daffodils and in the other they more frequently asked which flowers the experimenter meant. In the first photograph the daffodils stood out or were somehow different from the others and so they didn't need a unique identifier. For the **differences** algorithm we translated this into finding out what property values are perceptually prominent in the distractor set and then contrasting the target object with those prominent values. If any of the target object's properties are

[6]We assume that the discourse entities are collectives or plurals with homogeneous properties. When non homogeneous property values are indicated by set discourse entity annotations, we generalize the property values (e.g. "ours" for owner and "range" for non homogeneous colors and prices) to create the new discourse entity. However, these generalized property values are never used for coreference in the COCONUT dialogues.

different from the prominent values then those properties are included. IDAS selections, on the other hand, are influenced only by what is currently in the distractor set. A property that contrasts with a prominent one may not get selected because all the distractors with that prominent property have already been eliminated.

The first extreme algorithm that we will examine is **random**. **Random**, loops for the maximum number of property choices that can be made and randomly selects a property value to include regardless of whether the value for that property is known. This tends to skew the number of property values selected to a smaller number and is more like the distributions found in the corpus. The other two extreme selection algorithms that we will test are, **all** and **none**. **All** selects all the mutually known properties for every entity to be described and **none** is an extreme version of **gestalt** that always chooses to express just the object type.

To implement goal overloading for communicating constraint changes, we will first include the related property any time an implicit constraint change is made and then we will address the identification goal using the four selection algorithms described above.

5.2 Distractor Set Definitions

For the distractor set definition we have two extreme approaches, two that are related to Grosz & Sidner's theory of the attentional and intentional structure of discourse, and one that is based simply on recency. The first extreme, which we call **ALL**, includes every furniture discourse entity that has been mentioned thus far. The second extreme includes only the discourse entities that were discussed in the previous utterance. We call this definition, **1UTT**.

The first distractor set definition, based on Grosz & Sidner's theory is, **SEG**, which we described earlier. It duplicates Passonneau's distractor set definition. The second such algorithm, which we call, **SEG+SOLN**, includes all the furniture discourse entities in the current segment plus the discourse entities that are thought to be in the solution set for the problem at that point in the dialogue. The solution set is heuristically determined by assuming that if one of the pair stops talking about an action then it has probably been solved. The last definition, which is based solely on recency, uses all of the entities discussed in the last five utterances. We call this definition, **5UTT**.

5.3 Measuring Performance

algorithms	constraints considered	constraints not considered
IDAS	.68	.67
differences	.68	.67
gestalt	.68	.68
lex	.63	.62
random	NA	.51
all	NA	.45
none	NA	.47

Table 3: Algorithm Performances

To compare the performance of the algorithms to that of humans, we use a measure of the degree of match between the human's and the algorithm's selection of properties for the same discourse entity in the same dialogue context. Inclusion and exclusion of a property both count in the degree of match. We only consider four of the five properties associated with a

furniture entity in the COCONUT domain. These properties are; *color*, *price*, *owner*, *quantity*. We exclude the property *type* from the measure since it is generally assumed to be represented in the expression and since we are not studying the question of when to use pronouns and zero anaphors. So a perfect match means that the algorithm chose to include or exclude the same properties as the human did for a particular entity.

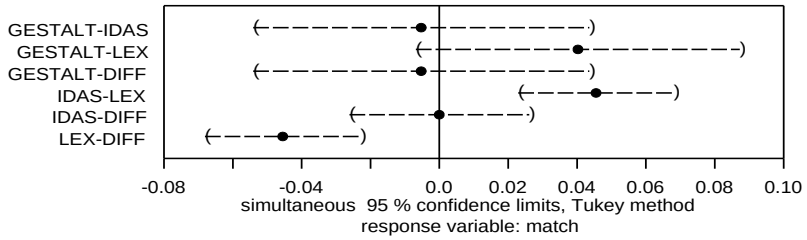

Figure 2: Performance Differences for Algorithms

Doing an analysis of variance (ANOVA) on the performance of the four identification algorithms, **IDAS**, **differences**, **gestalt** and **lex**, and the extremes of **random**, **all** and **none**, we find that their performances are significantly different (F=66, p=0). In Table 3, the means listed for when constraints are not considered, show the performance ranges. Using multiple comparison techniques, we find that the identification algorithms all perform better than the extremes. When we do an ANOVA on just the four identification algorithms, we find that there are significant differences in their performances as well (F=14, p=0). Using multiple comparison techniques, we find that **IDAS** and **differences** are significantly better than **lex** but not significantly different from one another or from **gestalt** (see Figure 2). There is a trend for **gestalt** to perform better than **lex**. Context insensitive algorithms like **gestalt** and **lex** do not perform as well as the minimally sensitive ones.

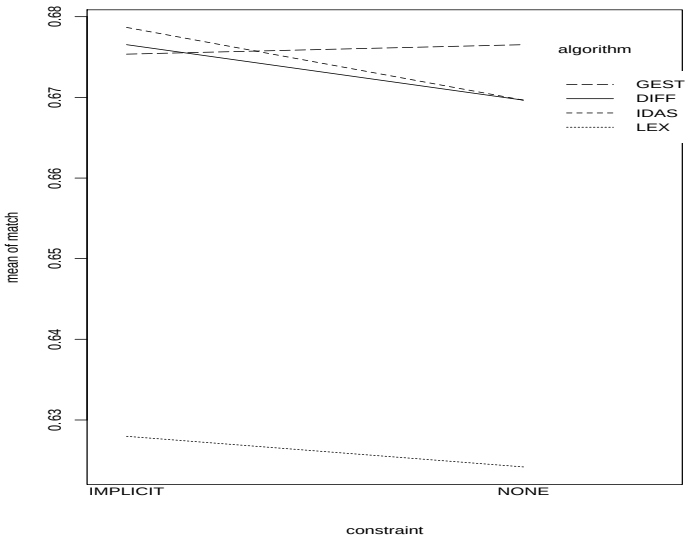

Figure 3: Performance for Identificational Algorithms with and without Constraints

When we overload the goal of communicating constraint changes we see via multiple comparison techniques that it significantly improves the performance when we look at that as the only factor. Comparing the means listed in Table 3 and plotted in Figure 3, gives an idea of the

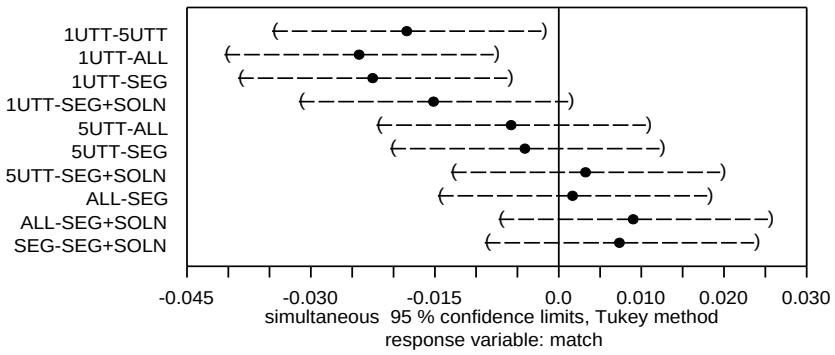

Figure 4: Performance Differences for Distractor Sets

improvements. However, when we consider the same data in terms of levels of algorithms, there is no significant difference in performance. However, we cannot prove there are no differences in performance because the sample size is not large enough to measure the effect of this factor. Here, we use an estimate of power from the Pearson-Hartley charts [Kep91]. According to this estimate we need a sample size of more than 600 coreferences to see whether the factor has no effect. In any case, we can conclude that any performance improvement is very small.

If we look at distractor set definitions, we see that there is no significant difference in performance between **SEG**, **SEG+SOLN**, **ALL** and **5UTT** using multiple comparison techniques (see Figure 4). Of these four distractor set definitions, all but **SEG+SOLN** perform significantly better than **1UTT**, but there is a trend for **SEG+SOLN** to perform better. Of the four that are not significantly different, there is a trend for **ALL** to be the best performer.

6 Conclusions and Future Work

We can conclude from the poor match to human performance that simple, domain independent algorithms are not good models. While there is strong evidence that we need approaches that are more context sensitive, goal overloading will need more study. We have isolated just one possible type of goal overloading and that by itself is not enough to significantly improve current approaches. There could be a different goal or a large set of such goals that would make an impact.

Another type of communicative goal that might overload with an identification goal is expressing reasons or justifications for particular solution suggestions. In our earlier example of how goal to communicate constraint changes could overload with an identification goal, we suggested that the price of the four $75 green chairs was highly salient. If that is the case, then price may have been included to help the hearer infer the reason for the selection; these chairs are cheaper than the alternatives. This reason also helps partially justify dropping the color match constraint. To test for overloading of communicative goals other than ones to communicate constraint changes, we would have to extend our analysis of the corpus by adding other annotation features. Identifying the range of communicative goals associated with this and other types of tasks is a matter of further empirical investigation.

It is possible that a combination of the identificational heuristics would better match human performance. In future work we will apply machine learning techniques to the annotated corpus to see if it will help us find a more sophisticated, context sensitive algorithm. It is also possible that our distractor set definitions are flawed or that distractor set definitions vary with the

the dialogue participants. It was surprising that the **ALL** distractor set definition had a trend towards being the best performer. This could be due to the communications setting for the COCONUT task. Some participants create and maintain graphics icons for all the items that have been discussed. These graphics icons could define the distractor set if both participants assume that the other also keeps a record of all the items discussed.

Finally we saw that some algorithms worked better for some speaker pairs than others, whereas for some pairs, none of the algorithms were good matches. In future work, it would be interesting to examine whether there are factors about these particular interactions that might make one algorithm a better match than another. Otherwise the choice of algorithm may just be due to the preferences of the dialogue pair.

While there is no substitute for a simulation approach when evaluating performance, it only addresses part of the evaluation question. Sometimes we want to find models that explain human performance and at others we are focused on finding computationally feasible models that will suffice. We want to know if the extra effort to find a better model is worthwhile from a system building perspective. To make such a determination, we should measure the intelligibility of dialogue agents who use simple approaches for selecting the content of nominals. As a first step, we could substitute nominals in the human data with ones generated by an algorithm and then readers could rate the dialogues for intelligibility.

References

[App85] Douglas E. Appelt. Some pragmatic issues in the planning of definite and indefinite noun phrases. In *Proceedings of 23rd ACL*, 1985.

[Car92] Jean C. Carletta. *Risk Taking and Recovery in Task-Oriented Dialogue.* PhD thesis, Edinburgh University, 1992.

[Car96] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Lingustics*, 22(2):249–254, 1996.

[CSB83] H.H. Clark, R. Schreuder, and S. Buttrick. Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22:245–258, 1983.

[CWG86] Herbert H. Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22:1–39, 1986.

[Dal92] Robert Dale. *Generating Referring Expressions.* ACL-MIT Series in Natural Language Processing. The MIT Press, 1992.

[Deu76] W. Deutsch. *Sprachliche Redundanz und Objektidentifikation.* PhD thesis, University of Marburg, 1976.

[DJTM99] Barbara Di Eugenio, Pamela W. Jordan, Richmond H. Thomason, and Johanna D. Moore. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *To Appear in International Journal of Human-Computer Studies*, 1999.

[DR95] Robert Dale and Ehud Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, Apr–June 1995.

[DW96] Barbara Di Eugenio and Bonnie Webber. Pragmatic overloading in natural language instructions. *International Journal of Expert Systems*, 9(2):53–84, 1996.

[Gri75] H.P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics III - Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[GS86] Barbara J. Grosz and Candace L. Sidner. Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.

[HH95] Peter A. Heeman and Graeme Hirst. Collaborating on referring expressions. *Computational Lingusitics*, 21(3), 1995.

[JT96] Pamela W. Jordan and Richmond H. Thomason. Refining the categories of miscommunication. In *Proceedings of AAAI Workshop on Detecting, Repairing, and Preventing Human-machine Miscommunication*, CA, 1996. AAAI Press.

[Kep91] Geoffrey Keppel. *Design and Analysis: A Researcher's Handbook, 3rd edition.* Prentice Hall, New Jersey, 1991.

[Kri80] Klaus Krippendorff. *Content Analysis: an Introduction to its Methodology.* Beverly Hills: Sage Publications, 1980.

[Kro86] Amichai Kronfeld. Donnellan's distinction and a computational model of reference. In *Proceedings of 24th ACL*, 1986.

[Lev89] W. J. M. Levelt. *Speaking: From Intention to Articulation.* MIT Press, 1989.

[Loc95] Karen Lochbaum. The use of knowledge preconditions in language processing. In *IJCAI95*, 1995.

[LS97] Claudio Lottaz and Ian Smith. Collaborative design using constraint solving. From Swiss Workshop on Collaborative and Distributed Systems, Lausanne Switzerland. See http://liawww.epfl.ch/ lottaz/ICCS/Design_and_CSP/design_and_CSP.html and http://liawww.epfl.ch/ lottaz/ICCS/Collaboration/index.html, May 1997.

[Lyo95] Kevin W. Lyons. Collaborative design for assembly of complex electro-mechanical products. Presentation abstract for NCMS Manufacturing Technical Conference. See http://elib.cme.nist.gov/made/presentations/ncms.html, May 1995.

[Man86] R. Mangold. *Sensorische Faktoren beim Verstehen uberspezifizierter ObjektBennennungen.* Peter Lang: Frankfurt, 1986.

[Pas96] Rebecca J. Passonneau. Using centering to relax gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech*, 39(2-3):229–264, 1996.

[Pol91] Martha E. Pollack. Overloading intentions for efficient practical reasoning. *Noûs*, 25:513 – 536, 1991.

[Rei90] Ehud Reiter. Generating appropriate natural language object descriptions. Technical Report TR-10-90, Department of Computer Science, Harvard University, 1990. Dissertation.

[Son82] S. Sonnenschein. The effects of redundant communications on listeners: When more is less. *Child Development*, 53:717–729, 1982.

[Son84] S. Sonnenschein. The effect of redundant communications on listeners: Why different types may have different effects. *Journal of Psycholingistic Research*, 13:147–166, 1984.

[SW98] Matthew Stone and Bonnie Webber. Textual economy through close coupling of syntax and semantics. In *workshop7*, Niagra-on-the-Lake, Canada, 1998.

[Wal93] Marilyn A. Walker. *Informational Redundancy and Resource Bounds in Dialogue*. PhD thesis, University of Pennsylvania, 1993.

[WGR93] Steve Whittaker, Erik Geelhoed, and Elizabeth Robinson. Shared workspaces: How do they work and when are they useful? *IJMMS*, 39:813–842, 1993.