Enhancing the Expression of Contrast in the SPaRKy Restaurant Corpus

David M. Howcroft and **Crystal Nakatsu** and **Michael White**
Department of Linguistics
The Ohio State University
Columbus, OH 43210, USA
{howcroft,cnakatsu,mwhite}@ling.osu.edu

Abstract

We show that Nakatsu & White's (2010) proposed enhancements to the SPaRKy Restaurant Corpus (SRC; Walker et al., 2007) for better expressing contrast do indeed make it possible to generate better texts, including ones that make effective and varied use of contrastive connectives and discourse adverbials. After first presenting a validation experiment for naturalness ratings of SRC texts gathered using Amazon's Mechanical Turk, we present an initial experiment suggesting that such ratings can be used to train a realization ranker that enables higher-rated texts to be selected when the ranker is trained on a sample of generated restaurant recommendations with the contrast enhancements than without them. We conclude with a discussion of possible ways of improving the ranker in future work.

1 Introduction

To lessen the need for handcrafting in developing generation systems, Walker et al. (2007) extended the overgenerate-and-rank methodology (Langkilde and Knight, 1998; Mellish et al., 1998; Walker et al., 2002; Nakatsu and White, 2006) to complex information presentation tasks involving variation in rhetorical structure. They illustrated their approach by developing SPaRKy (Sentence Planning with Rhetorical Knowledge), a sentence planner for generating restaurant recommendations and comparisons in the context of the MATCH (Multimodal Access To City Help) system (Walker et al., 2004), and showed that SPaRKY can produce texts comparable to those of MATCH's template-based generator.

Despite the evident importance of expressing contrast clearly in making comparisons among restaurants, Nakatsu (2008) surprisingly found that most of the examples involving contrastive connectives in the SPaRKy Restaurant Corpus (SRC) received low ratings by the human judges. Even though the low ratings were not necessarily directly attributable to the use of a contrastive connective in many cases, Nakatsu conjectured that the large proportion of low-rated examples containing contrastive connectives would make it difficult to train a ranker to learn to use contrastive connectives effectively without augmenting the corpus with better examples of contrast. Subsequently, Nakatsu and White (2010) proposed a set of enhancements to the SRC intended to better express contrast—including ones employing multiple connectives in the same clause that are problematic for RST (Mann and Thompson, 1988)—and showed how they could be generated with Discourse Combinatory Categorial Grammar (DCCG), an extension of CCG (Steedman, 2000) designed to enable multi-sentence grammar-based generation. However, Nakatsu and White did not evaluate empirically whether these contrast enhancements were successful.

In this paper, we show that Nakatsu & White's (2010) proposed SRC contrast enhancements do indeed make it possible to generate better texts: in particular, we present an initial experiment that shows that the oracle best restaurant recommendations including the contrast enhancements have significantly higher human ratings for naturalness than comparable texts without these enhancements, and which suggests that even a basic n-gram ranker trained on the enhanced recommendations can select texts with higher ratings. The paper is structured as follows. In Section 2, we review Nakatsu & White's proposed enhancements to the SRC for better expressing contrast—including the use of *structural connectives* together with *discourse adverbials*—and how they can be generated with DCCG. In Sec-

Proceedings of the 14th European Workshop on Natural Language Generation, pages 30–39,
Sofia, Bulgaria, August 8-9 2013. ©2013 Association for Computational Linguistics

tion 3, we first present a validation experiment showing that naturalness ratings gathered on Amazon's Mechanical Turk (AMT) are comparable to those for the same texts in the original SRC; then, we present our method of generating and selecting a sample of new restaurant recommendation texts with and without the contrast enhancements for rating on AMT. In Section 4, we describe how we trained discriminative n-gram rankers using cross validation on the gathered ratings. In Section 5, we present the oracle and cross validation results in terms of mean scores of the top-ranked text. In Section 6, we analyze how the individual contrast enhancements affected the naturalness ratings and discuss issues that may be still hampering naturalness. Finally, in Section 7, we conclude with a summary and a discussion of possible ways of creating improved rankers in future work.

2 Enhancing Contrast with Discourse Combinatory Categorial Grammar

Figure 1 (Nakatsu, 2008) shows examples from the SRC where some of the SPaRKy realizations are clearly more natural than others. In Nakatsu's experiments, she found that the use of contrastive connectives was negatively correlated with human ratings, and that an n-gram ranker learned to disprefer texts containing these connectives. In analyzing these unexpected results, Nakatsu noted two factors that appeared to hamper the naturalness of the contrastive connective usage. First, consistent with Grote et al.'s (1995) observation that *however* and *on the other hand* (unlike *but* and *while*) signal that the clause they attach to is the more important one, we might expect realizations to be preferred when these connectives appear with the more desirable of the contrasted qualities. Such preferences do indeed appear to be present in the SRC: for example, in Figure 1, alts 8 & 13—where the better property is ordered second—are rated highly, while alts 7 & 11—where the better property is ordered first—are rated poorly. Nakatsu further observed that in human-authored comparisons, when the second clause expresses the lesser property, it is often qualified by *only* or *just*; consistent with this observation, alts 7 & 11 do seem to improve with the inclusion of these modifiers.

The second factor noted by Nakatsu that may contribute to the awkwardness of *however* and *on the other hand* is that both of these connectives seem to be rather "grand" for the rather simple contrasts in Figure 1, and may sound more natural when used with heavier arguments.

Based on these observations, Nakatsu and White (2010) proposed a set of enhancements to the SRC, all of which are exemplified in Figure 2.[1] The enhancements include (i) optional summary statements that give an overall assessment of each restaurant based on the average of their property values, thereby allowing contrasts to be expressed over larger text spans; (ii) adverbial modifiers *only*, *just* and *merely* to express a lesser value of a given property than one mentioned earlier;[2] (iii) the modifers *also* and *too* to signal the repetition of the same value for a given property (Striegnitz, 2004); and (iv) contrastive connectives for different properties of the same restaurant, exemplified here by the contrast between decent decor and mediocre food quality for Bienvenue.

In the text plan in Figure 2, <1>–<4> correspond to the propositions in the original SRC text plan and (1')–(2') are the new summary-level propositions. Following Webber et al. (2003), Nakatsu and White (2010) take *only*, *merely*, *just*, *also*, and *too* to be **discourse adverbials**, whose discourse relations are allowed to cut across the primary tree structure established by the other relations in the figure. Note that in addition to going beyond RST's limitation to tree-structured discourses, the example also contains clauses employing multiple discourse connectives, where one is a **structural connective** (such as *however* or *while*) and the other is a discourse adverbial.

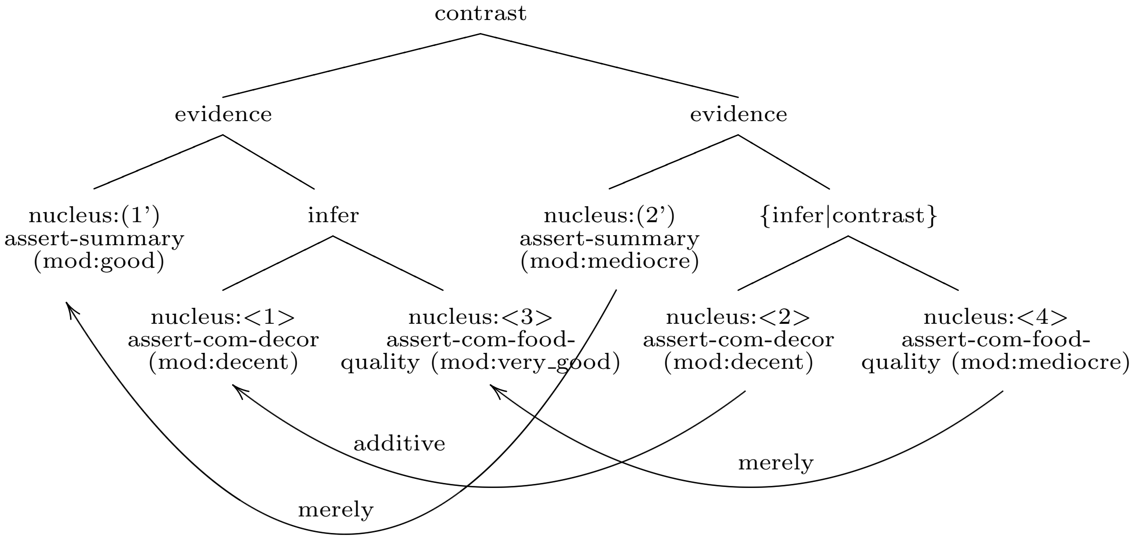
To realize such texts, Nakatsu & White introduce Discourse Combinatory Categorial Grammar (DCCG), an extension of CCG (Steedman, 2000) to the discourse level. DCCG follows Discourse Lexicalized Tree Adjoining Grammar (Webber, 2004) in providing a lexicalized treatment of structural connectives and discourse adverbials, but differs in doing so in a single CCG, rather than separate sentence-level and discourse-level grammars whose interaction is not straightforward. As such, DCCG requires no changes to the OpenCCG realizer (White, 2006b; White, 2006a; White and Ra-

[1] In the text, words intended to help indicate similarities and contrasts are italicized. Note that we have added *overall* and *on the whole* to the summary statements to better indicate their summarizing role.

[2] The second value must be a less extreme one on the same side of the scale; in principle, it could be *merely poor* rather than *horrible*, but such low attribute values did not occur in the corpus.

Strategy	Alt #	Rating	Rank	Realization
C2	3	3	7	Sonia Rose has very good decor but Bienvenue has decent decor.
	7	1	16	Sonia Rose has very good decor. On the other hand, Bienvenue has decent decor.
	8	4.5	13	Bienvenue has decent decor. Sonia Rose, on the other hand, has very good decor.
	10	4.5	5	Bienvenue has decent decor but Sonia Rose has very good decor.
	11	1	12	Sonia Rose has very good decor. However, Bienvenue has decent decor.
	13	5	14	Bienvenue has decent decor. However, Sonia Rose has very good decor.
	14	5	3	Sonia Rose has very good decor while Bienvenue has decent decor.
	15	4	4	Bienvenue has decent decor while Sonia Rose has very good decor.
	17	1	15	Bienvenue's price is 35 dollars. Sonia Rose's price, however, is 51 dollars. Bienvenue has decent decor. However, Sonia Rose has very good decor.

Figure 1: Some alternative (Alt) realizations of SPaRKy sentence plans from a COMPARE2 (C2) plan, with averaged human ratings (Rating; 5 = highest rating) and ranks (Rank; 1 = top ranked) assigned by an n-gram ranker (Nakatsu, 2008)



(1'): Sonia Rose is a good restaurant overall.

<1>: It has decent decor and

<3>: very good food quality.

(2'): *However*, Bienvenue is *just* a mediocre restaurant on the whole.

<2>: *While* it *also* has decent decor,

<4>: it *only* has mediocre food quality.

Figure 2: Modified SPaRKy text plan for text with new relations and summary statements intended to enhance contrast (Nakatsu and White, 2010)

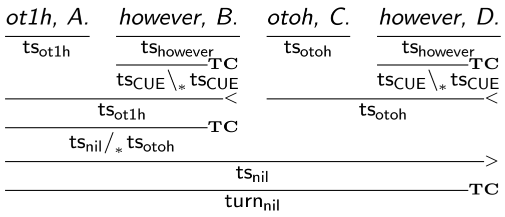


Figure 3: DCCG derivation of nested contrast relations (Nakatsu and White, 2010)

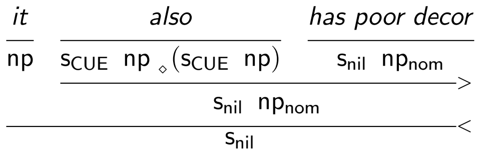


Figure 4: DCCG derivation of a clause with the discourse adverbial *also* (Nakatsu and White, 2010)

jkumar, 2009) in order to generate texts that vary in size from single sentences to entire paragraphs.

In DCCG, the technique of **cue threading** is used to allow structural connectives—including paired ones such as *on the one hand … on the other hand*—to project beyond the sentence level, while allowing no more than one to be active at a time. In this way, structural connectives can be nested, as sketched in Figure 3, but cannot cross. In the figure, the value of the cue feature for each text segment (ts) is shown (where otlh and otoh abbreviate *on the one hand* and *on the other hand*); these cue values can be propagated through a derivation, allowing the discourse relations to project, but must be discharged (to nil) in a complete derivation, thereby ensuring that the intended discourse relations are actually realized. By contrast, discourse adverbials introduce their relations anaphorically and are transparent to cue threading, as sketched in Figure 4, making use of typical adverb categories syntactically. See Nakatsu and White (2010) for further details.

3 Crowd Sourcing Ratings

To collect human judgements from a diverse group of speakers of US English, we used Amazon's Mechanical Turk service (AMT) to run two experiments. In the first experiment, subjects rated the naturalness of 174 passages used in Walker et al.'s (2007) study. As detailed in Section 5, this validation experiment confirmed that the judgements collected on AMT correlate with those of the raters in Walker et al.'s (2007) study. Our second experiment collected ratings on 300 passages realized with modifications for better contrast expression (WITHMODS) and 300 passages without these modifications (NOMODS), both realized using OpenCCG. While this does not admit a direct comparison to the realizations produced by Walker et al. (2007), this controls for differences between the generators other than the variable of interest: the contrastive enhancements. In addition to these materials, five passages from the SRC were seen by all subjects to control for anomalous subject behavior.

3.1 Survey Format

Each survey used demographic questions to determine the native speaker status of the subject. Instructions for completing comprehension questions and rating realizations followed the demographic questions.[3] Each subject saw fifteen stimuli, each consisting of a sample user query and the target passage as in Figure 5. After reading the stimulus, the subject answered a yes-or-no comprehension question (see §3.2). Finally the subject rated the naturalness of the passage on a seven-point Likert scale ranging from *very unnatural* to *very natural*. At the survey's conclusion, the subject could offer free-form feedback, explain their responses, or ask questions of the researchers. The average completion time across all experiments was about ten minutes.

Passage selection is detailed in §3.3 and §3.4.

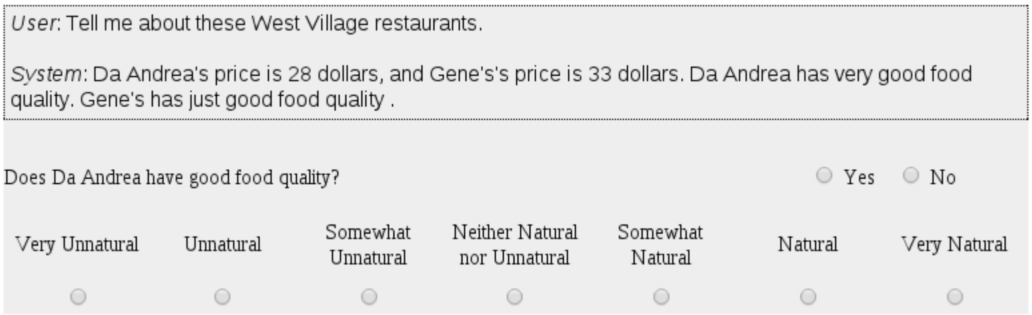
3.2 Quality Control

We used three strategies to filter out low-quality responses from AMT subjects.

Comprehension Questions A template-based yes-or-no question (exemplified in Figure 5) followed each passage. Subjects who answered less than 75% of these questions correctly were rejected and not paid, in accordance with the protocol approved by our human subjects review board. Responses from three subjects were excluded from analysis on this basis.

Uniform Ratings When a subject gave the same rating for all passages in a given survey (and in disagreement with other subjects), we took this to mean that the subject was paying attention only

[3] These materials, along with the generated passages and their ratings are available at `http://www.ling.ohio-state.edu/~mwhite/data/enlg13/`.



User: Tell me about these West Village restaurants.

System: Da Andrea's price is 28 dollars, and Gene's's price is 33 dollars. Da Andrea has very good food quality. Gene's has just good food quality .

Does Da Andrea have good food quality? ○ Yes ○ No

Very Unnatural	Unnatural	Somewhat Unnatural	Neither Natural nor Unnatural	Somewhat Natural	Natural	Very Natural
○	○	○	○	○	○	○

Figure 5: Sample survey stimulus and comprehension question

Method	# subjects excluded
Comprehension Questions	3
Uniform Answers	1
SAME5	0
Native Speaker Status	2

Table 1: Number of subjects excluded based on quality control measures or native language.

to the comprehension questions that ensured payment. Only one subject was excluded on this basis, though they were still paid for answering the comprehension questions correctly.

SAME5 Passages Five passages were chosen from the original SRC realizations for which the original ratings (from Walker et al. 2007) were identical for both judges. The passages were selected such that the first and third authors of this paper agreed with the general valence and relative rankings of the passages. That is, we took two unambiguously bad realizations, two unambiguously good realizations, and one realization near the middle of the spectrum to represent a gold standard for rating to compare subjects against. If any subject's ratings on these five passages were clear outliers, we could remove that subject's data for anomalous behavior, but this measure proved unnecessary for the subjects in the present study.

3.3 Validating AMT

Data Selection In this experiment, we sampled 174 of the 1757 realizations from the SRC rated by subjects A and B in Walker et al.'s (2007) experiment.

The SRC realizations were divided randomly into two groups. Within one group, realizations were labelled by subject A's rating for that realization. Subject B's rating was used for the other group. Taking the poles of the rating scale and its midpoint, the realizations were further partitioned into six sets: realizations rated 1, 3, and 5 by subject A and realizations rated 1, 3, and 5 by subject B. This division of the data ensured that the realizations used would cover the full spectrum of ratings while being representative of the SRC ratings with respect to, e.g., inter-annotator ratings correlations.

From each of these six sets, we chose 10 COMPARE2, 10 COMPARE3, and 10 RECOMMEND realizations,[4] each of these groups representing a different realization task in the SRC. The COMPARE2 and COMPARE3 tasks involved the comparison of two restaurants or three or more restaurants, respectively. In the RECOMMEND context, the sytem had to generate a recommendation for a single restaurant.

Subject Demographics Thirty-six subjects responded to this survey initially, but one was rejected based on a failure to answer the comprehension questions and data from another had to be excluded for non-native speaker status. Two additional subjects were recruited to replace their data. This resulted in a subject pool with a mean age (std. dev.) of 34.67 (9.35) years. Twenty-four subjects identified as female and twelve identified as male. Each subject received $2.50 for the survey, estimated to take approximately 20 minutes.

3.4 Rating OpenCCG Realizations

Data Selection We selected 15 content plans (CPs) from the SRC where the use of the contrastive modifiers was licensed: five COMPARE2, five COMPARE3, and five RECOMMEND CPs. Each of the 112 textplans (TPs) that produced

[4] Except that subject A used the rating '5' less than subject B. To compensate, we used as many 5-point ratings as were available from subject A and then filled in the remainder of the 10 slots with realizations rated '4'. We mirrored these selections in the data from subject B for consistency.

the SRC realizations for these CPs was then preprocessed for realization in OpenCCG both with contrast enhancements (WITHMODS) and without them (NOMODS).

Both structural choices and ordering choices are encoded in these TPs.[5] Structural choices include decisions about how to group the restaurant properties to be expressed, such as deciding whether to describe one restaurant in its entirety and then the other (i.e. a *serial* structure) or alternating between one restaurant and the other, directly contrasting particular attributes (i.e. a *back-and-forth* structure). Ordering choices fixed the order of presentation of restaurant attributes in serial plans and the order of presentation of attribute contrasts in back-and-forth plans. As discussed in §6, there turn out to be interesting interactions between these aggregation choices and the contrast enhancements, interactions which we did not explore directly in this experiment.

Processing each TP produced a different LF for each possible combination of aggregation choices and contrastive modifications, resulting in approximately 41k logical forms (LFs) for the TPs WITHMODS and 88k LFs for the TPs with NOMODS.[6]

Each realization received two language model (LM) scores, one based on the semantic classes used during realization (LM_{SC}) and one based on the Gigaword corpus (LM_{GW}). LM_{SC} used a trigram model over modified texts based on the SRC where specific entities (e.g. restaurant names like *Caffe Buon Gusto*) were replaced with their semantic class (e.g. *RESTAURANT*). The LM scores were normalized by CP, such that the scores for a given CP summed to 1 in each LM. These were then linearly combined with weights slightly preferring the LM_{SC} score to produce a combined LM score for each realization.

Sampling then proceeded without replacement, weighted by the combined LM score for each realization. For the NOMODS sample, 20 realizations were chosen this way, but, in the WITHMODS sample, a series of regular expression filters were used to ensure adequate representation of the modifications in the surveys. These filters selected (without replacement) 10 realizations such that every contrastive modification licensed by a particular CP was represented, leaving 10 realizations to be selected by weighted sampling without replacement.

This process resulted in 300 passages in each of the two conditions (WITHMODS, NOMODS): 20 realizations for each of the 15 CPs. Each survey included 5 realizations WITHMODS paired by CP with 5 realizations with NOMODS as well as the SAME5 realizations. As noted earlier, pairing realizations in this way helps to control for differences in the variety of aggregation choices and surface realizations used in the SRC as opposed to our SRC-inspired grammar for OpenCCG.

Subject Demographics Sixty-eight subjects responded to these 180 surveys initially. Subjects were allowed to complete up to six distinct surveys. One subject's data was excluded for non-native status and another's was excluded on the basis of uniform ratings (as detailed in §3.2). To compensate for the eight surveys completed by these subjects and ten surveys mistakenly administered in draft format, we recollected data for 18 of the 180 surveys. This resulted in a final pool of 80 subjects with an average (std. dev.) age 37.15 (13.5) years. Forty identified as female, thirty-nine identified as male, and one identified as non-gendered.

Because subjects in the validation study completed the survey in about 10 minutes on average with a standard deviation of about 5 minutes, we scaled the pay to $2.00 per survey in this experiment. Since subjects could participate in this experiment multiple times, they could receive up to $12.00 for their contribution.

4 Training a Text Ranker

To perform the ranking, we trained a basic n-gram ranker using SVM$^{\text{light}}$ in preference ranking mode.[7] We used the average ratings obtained in §3 as target value.

The feature set was composed of 2 types of features. The first feature type are the two language model scores from §3.4, LM_{SC} and LM_{GW}. The second feature type consisted of n-gram counts. We indexed the unigrams and bigrams in each corpus and used each as a feature whose value was the number of times it appeared in a given realization.

We trained the ranker on, and extracted n-gram

[5] This differs from Walker et al. (2007), wherein reorderings were allowed in mapping from tp-trees to sp-trees and d-trees.

[6] In future work we will explore a probabilistic rather than exhaustive mapping algorithm to produce only LFs that are more likely to result in more fluent realizations—not unlike the weighted aggregation done by Walker et al.'s (2007) sentence plan generator.

[7] SVM$^{\text{light}}$ is an implementation of support vector machines by (Joachims, 2002).

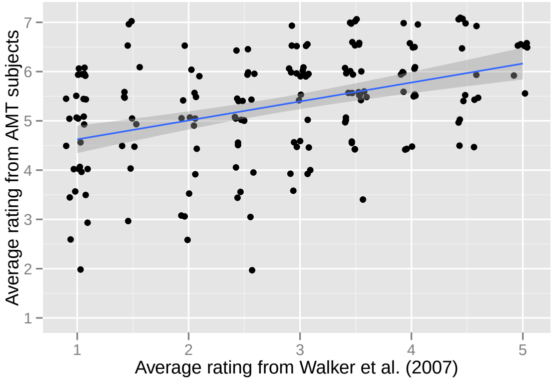


Figure 6: Average ratings from our experiment and Walker et al. (2007), accompanied by a line of best fit. Jitter (0.1) applied to each point minimizes overlap.

features from, 3 different corpora drawn from the data selection in §3.4. The first corpus contains 299 selections WITHMODS (1 selection was discarded for only being rated once), the second corpus contains 300 selections with NOMODS, and the third corpus contains BOTH of the first two corpora combined.

To train and test the ranker, we performed 15-fold cross-validation on each corpus. Within each training fold, we had 14 training examples, corresponding to 14 CPs. Each training example consisted of all of a given CP's realizations and their ratings. After training, the realizations for the remaining CP were ranked.

In order to evaluate the ranker, we used the TopRank metric (Walker et al., 2007). For each of the ranked CP realization sets, we extracted the target values (i.e. the average rating given by subjects) of the highest ranked realization. We then averaged the target scores of all of the top-ranked realizations across the 15 training folds to produce the Top Rank metric. The oracle best score is the score of the highest rated realization, as determined by the average score assigned to that realization by the subjects.

5 Results

Validation Figure 6 shows the correlation between the average ratings of our subjects on AMT and the average ratings assigned by subjects A and B in Walker et al. (2007). This correlation was 0.31 ($p < 0.01$, Kendall's tau), while the correlation between subjects A and B was only 0.28 ($p < 0.01$, Kendall's tau). On this basis we conclude that using AMT workers as subjects to rate sentences for their naturalness is at least as reasonable as having two expert annotators labelling realizations for their overall quality.

	BOTH	WITHMODS	NOMODS
human	6.61 (0.28)	6.46 (0.43)	6.49 (0.26)
bigram	6.00 (0.58)	5.62 (0.83)	5.51 (1.02)

Table 2: TopRank scores and standard deviations for the oracle (human) & bigram (bigram) ranks.

SAME5 Comparison There was no significant difference ($p = 0.16$, using Welch's t-test) between the scores given to the SAME5 stimuli in the two experiments,[8] indicating that subjects used the rating scale similarly in both experiments. The mean ratings for the rest of the validation realizations was 5.31 (1.43) and the mean for the OpenCCG-based realizations in the ranking experiment was 4.96 (1.51), which is significantly lower according to Welch's t-test ($p < 0.01$). This highlights the underlying differences between the two generation systems, validating our choice to use OpenCCG for both the WITHMODS and NOMODS realizations to better examine the impact of the contrast enhancements.

Ranking Table 2 reports the oracle results, along with our ranker's results, using the TopRank metric. Most indicative of the benefit of the contrastive enhancements is the performance of the oracle score for the BOTH (6.61) condition compared to the NOMODS condition (6.49), which is significantly higher according to a paired t-test ($p = 0.01$).

We also found that the bigram ranker with the averaged raw ratings was better at predicting the top rank of the combined (BOTH) corpus (6.00 vs. oracle-best of 6.61) than either of the other two, and better on the WITHMODS condition (5.62) than on the NOMODS condition (5.51). However, a two-tailed t-test revealed that the difference was not quite signficant between BOTH and NOMODS at the conventional level ($p = 0.06$), though the p-value did meet the 0.1 threshold sometimes employed in small-scale experiments. The performance of the different rankers, as compared to the oracle scores, can be seen in Figure 7.

These preliminary results with a simple ranker

[8] Validation experiment mean (std. dev.) 4.89 (1.79) versus 5.10 (1.75) in the ranking experiment.

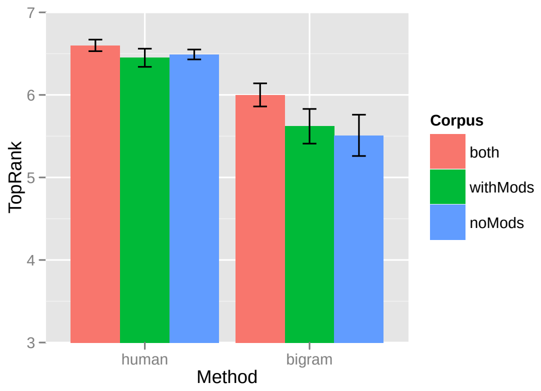


Figure 7: TopRank scores for each of the rankers with standard error bars.

are promising, motivating future work on improving the ranker in addition to enlarging the dataset.

6 Discussion

To assess the impact of the enhancement options, we performed a linear regression between the contrast-related patterns we used for data selection and the normalized ratings, with scikit-learn's implementation of the Bayesian Ridge method of regularizing weights.[9] In looking at examples, we found that the number of discourse adverbials appeared to be a factor, so we then added these counts as features. The coefficients and corpus counts appear in Table 3. The results show that the discourse adverbials were effective some of the time, especially when used sparingly and in conjunction with *while*. The "heavier" contrastive connectives *however* and *on the one/other hand* were dispreferred, perhaps in part because they ended up appearing too often with small, single-restaurant contrasts, as there were relatively few examples of summary statements, most of which were somewhat disfluent due to a medial choice for *overall / on the whole*.

pattern	coeff	count
$\mid$ disc advb $\mid = 1$	0.23	102
while	0.19	38
also has	0.13	47
has . . . too	0.12	39
has only	0.09	43
while . . . disc advb	0.09	16
contrastive . . . *overall*	0.07	8
has just	0.04	46
however . . . disc advb	0.03	4
but	-0.03	20
, however ,	-0.05	10
only has	-0.06	30
has merely	-0.11	46
on the whole	-0.14	33
just has	-0.16	29
merely has	-0.16	8
$\mid$ disc advb $\mid = 2$	-0.18	32
. however ,	-0.21	64
on the other hand	-0.21	40
$\mid$ disc advb $\mid >= 3$	-0.27	50
overall	-0.29	34
on the one hand	-0.36	22

Table 3: Coefficients of linear regression between contrast-related patterns and normalized ratings, along with pattern counts, where disc adv is one of *just, only, merely, also, too* and contrastive is one of *while, however, on the one/other hand*

Table 4 shows examples that illustrate both successes and remaining issues. At the top, two pairs of examples are given where the normalized average ratings are higher with the inclusion of *just* and *only*, and where the rating drops off greatly when *however* is used with a lesser value and no adverbial of this kind, as expected. At the bottom, the first example shows one instance where the use of multiple adverbials is dispreferred. A possible factor here may be that in addition to there being several similar adverbials in a row, they all involve long-distance antecedents, which may be difficult to process. Finally, the last example shows a realization that receives a relatively high rating despite the use of two adverbials; note, however, that since this passage uses a back-and-forth text plan, the antecedents of the adverbials are all very local.[10]

Turning to the survey feedback, many subjects provided insightful comments regarding the task. The most frequent comment pointed out that our comprehension questions sometimes precipitated a false implicature: when asked if a restaurant had decent decor, subjects commented that they felt that answering "no" meant implying that it had terrible decor. Similar problems occurred when a restaurant had, e.g., *very good* decor and the subjects were asked if it had *good* decor. Despite occasional deviations from our intended exact-match interpretation of these questions, no subjects were excluded for scoring too low as a result of this.

[9] http://scikit-learn.org/stable/modules/linear_model.html

[10] As one reviewer points out, there's also an interaction between how attributes are aggregated and the ability to express contrast. For example, contrasting the attributes for which a restaurant scores highly with those for which it scores poorly requires the aggregation of attributes with like valence, as in "This restaurant has superb decor and very good service but only mediocre food quality." Our future work on aggregation will explore this interaction as well.

Strategy	Mods?	Rating	Realization
C2	Y	1.13	Da Andrea's price is 28 dollars. Gene's's price is 33 dollars. Da Andrea has very good food quality while Gene's has *just* good food quality.
C2	N	0.73	Da Andrea's price is 28 dollars. Gene's's price is 33 dollars. Da Andrea has very good food quality while Gene's has good food quality.
C2	Y	1.04	Da Andrea's price is 28 dollars. Gene's's price is 33 dollars. Da Andrea has very good food quality. However, Gene's has *only* good food quality.
C2	N	-0.63	Da Andrea's price is 28 dollars. Gene's's price is 33 dollars. Da Andrea has very good food quality. However, Gene's has good food quality.
C3	Y	-1.85	Daniel and Jo Jo offer exceptional value among the selected restaurants. Daniel, *on the whole*, is a superb restaurant. Daniel's price is 82 dollars. Daniel has superb decor. It has superb service and superb food quality. Jo Jo, *overall*, is an excellent restaurant. Jo Jo's price is 59 dollars. Jo Jo *just* has very good decor. It *just* has excellent service. It has *merely* excellent food quality.
C2	Y	1.12	Japonica's price is 37 dollars while Dojo's price is 14 dollars. Japonica has excellent food quality while Dojo has *merely* decent food quality. Japonica has decent decor. Dojo has *only* mediocre decor.

Table 4: Examples illustrating successful and problematic contrast enhancements

In order to elicit rankings at a variety of points on the naturalness scale, our selection included a number of realizations with lower quality overall, which subjects picked up on. For example, one subject commented that, "Repeatedly using the name of each restaurant over and over in simple sentences make[s] almost all of these excerpts sound horrifyingly awkward," while another observed, "The constant [use] of more sentences, instead of using conjunction words . . . makes it seem as if the system is rambling and lost in though[t] process."

Several subjects also pointed out that it would be more natural to discuss the cost of an average meal at a restaurant than to state that a restaurant's price is some particular number of dollars. Though these domain-specific lexical preferences are tangential to the focus of this paper, they suggest that exploring options to expand the range of realizations for more naturally expressing these properties might be a fruitful direction for future work.

In addition to expressing an explicit preference for serial rather than back-and-forth text-plans, subjects also commented that higher level contrastive adverbials like *however* work better when they are used sparingly at a high level, reinforcing the findings in our regressions. We also received suggestions for future work improving the expression of contrast: some subjects suggested that using *better* and *worse* to make explicit comparisons between restaurants would improve the naturalness, and one subject suggested explicitly stating which restaurant is (say) the *cheapest* as in White et al. (2010).

7 Conclusions and Future Work

In this paper, we have shown using ratings gathered on AMT that Nakatsu & White's (2010) proposed enhancements to the SPaRKy Restaurant Corpus (Walker et al., 2007) for better expressing contrast do indeed make it possible to generate better texts, and an initial experiment suggested that even a basic n-gram ranker can do so automatically. A regression analysis further revealed that while using a few discourse adverbials sparingly was effective, using too many discourse adverbials had a negative impact, with antecedent distance potentially an important factor. In future work, we plan to improve upon this basic n-gram ranker to take these observations into account and validate these initial findings on a larger dataset. In the process we will explore the interaction between contrast expression and aggregation and seek to better model the felicity conditions for "weighty" top level adverbials such as *however*.

Acknowledgments

This work was supported in part by NSF grant IIS-1143635. Special thanks to the anonymous reviewers, the Clippers computational linguistics discussion group at Ohio State, and to Mark Dras, Francois Lareau, and Yasaman Motazedi at Macquarie University.

References

Brigitte Grote, Nils Lenke, and Manfred Stede. 1995. Ma(r)king concessions in English and German. In *Proc. of the Fifth European Workshop on Natural Language Generation*.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proc. KDD*.

Irene Langkilde and Kevin Knight. 1998. The practical value of n-grams in generation. In *Proc. INLG-98*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization. *TEXT*, 8(3):243–281.

Chris Mellish, Alistair Knott, Jon Oberlander, and Mick O'Donnell. 1998. Experiments using stochastic search for text planning. In *Proc. INLG-98*.

Crystal Nakatsu and Michael White. 2006. Learning to say it well: Reranking realizations by predicted synthesis quality. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1113–1120, Sydney, Australia, July. Association for Computational Linguistics.

Crystal Nakatsu and Michael White. 2010. Generating with discourse combinatory categorial grammar. *Linguistic Issues in Language Technology*, 4(1):1–62.

Crystal Nakatsu. 2008. Learning contrastive connectives in sentence realization ranking. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 76–79, Columbus, Ohio, June. Association for Computational Linguistics.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press.

Kristina Striegnitz. 2004. *Generating Anaphoric Expressions — Contextual Inference in Sentence Planning*. Ph.D. thesis, University of Saalandes & Universit de Nancy.

Marilyn A. Walker, Owen C. Rambow, and Monica Rogati. 2002. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language*, 16:409–433.

M. A. Walker, S. J. Whittaker, A. Stent, P. Maloor, J. D. Moore, M. Johnston, and G Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840.

M. Walker, A. Stent, F. Mairesse, and Rashmi Prasad. 2007. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research (JAIR)*, 30:413–456.

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4).

Bonnie Webber. 2004. D-LTAG: Extending lexicalized TAG to discourse. *Cognitive Science*, 28(5):751–779.

Michael White and Rajakrishnan Rajkumar. 2009. Perceptron reranking for CCG realization. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Singapore, August. Association for Computational Linguistics.

Michael White, Robert A. J. Clark, and Johanna D. Moore. 2010. Generating tailored, comparative descriptions with contextually appropriate intonation. *Computational Linguistics*, 36(2):159–201.

Michael White. 2006a. CCG chart realization from disjunctive logical forms. In *Proc. INLG-06*.

Michael White. 2006b. Efficient Realization of Coordinate Structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75, June.