The Extended SPaRKy Restaurant Corpus: designing a corpus with variable information density

David M. Howcroft, Dietrich Klakow, Vera Demberg

Department of Language Science and Technology
Saarland Informatics Campus, Saarland University, Germany
{howcroft,vera}@coli.uni-saarland.de   dietrich.klakow@lsv.uni-saarland.de

Abstract

Natural language generation (NLG) systems rely on corpora for both hand-crafted approaches in a traditional NLG architecture and for statistical end-to-end (learned) generation systems. Limitations in existing resources, however, make it difficult to develop systems which can vary the linguistic properties of an utterance as needed. For example, when users' attention is split between a linguistic and a secondary task such as driving, a generation system may need to reduce the information density of an utterance to compensate for the reduction in user attention.

We introduce a new corpus in the restaurant recommendation and comparison domain, collected in a paraphrasing paradigm, where subjects wrote texts targeting either a general audience or an elderly family member. This design resulted in a corpus of more than 5000 texts which exhibit a variety of lexical and syntactic choices and differ with respect to average word & sentence length and surprisal. The corpus includes two levels of meaning representation: flat 'semantic stacks' for propositional content and Rhetorical Structure Theory (RST) relations between these propositions.

Index Terms: natural language generation, corpora, surprisal, variation

1. Introduction

Natural language generation (NLG) has a long history of corpus-based design, whether studying human behavior to develop hand-crafted components within the traditional NLG architecture or training 'end-to-end' statistical systems. However, the meaning representations (MR) that existing corpora tend to use are fairly impoverished: large recent resources used for training neural networks contain sets of "dialog acts" which lack hierarchical structure [1, 2, 3]. And even those corpora that are accompanied by a meaning representation which includes hierarchical discourse structure still suffer from a relatively limited range of possible lexicalizations.

In order to further improve current NLG systems and make them suitable for a wider range of applications, users, and usage situations, we need to develop NLG and dialogue systems which exhibit a decent range of linguistic variation. In addition to simply providing for more natural interactions with dialogue systems by reducing repetition, variation allows the system to adapt to different requirements, styles, and user preferences. In order to be able to learn such variability using data-driven systems, we therefore need training data which exhibits at least as much variation as we'd like to see in our target applications.

Imagine, for example, usage scenarios of a given app with a distracted user paying only partial attention to the dialogue system interaction, or usage of the app by elderly users—in each of these situations, we humans would adapt our communication to the different circumstances [4, 5, 6], and so should NLG components of dialog systems, if they are to communicate successfully and be more human-like.

Prior work has looked at adapting NLG systems to exhibit different personality traits and accommodate different user preferences at the levels of content selection and sentence structure [7, 8, 9, 10], but adapting the linguistic difficulty of system utterances has been less explored. While work on automatic text simplification has addressed the issue of the linguistic difficulty of texts, it only does so in the context of text-to-text generation, where the system must rewrite an existing text to achieve, e.g., a lower reading level. For concept-to-text generation, on the other hand, there are currently no systems that can dynamically change the level of linguistic complexity in their output.

Choosing an appropriate measure of linguistic complexity is difficult, but there is evidence from psycholinguistic research suggesting that a text's *information density* (i.e. its Shannon information or *surprisal*) is one relevant factor. In written text, for example, we know that information density significantly influences processing difficulty [11] in reading. With respect to production, speakers are more likely to reduce phonemes and omit optional syntactic markers if they are less informative in context [12, 13, 14]. That is to say, humans are sensitive to information density, in both comprehension and production, so we can expect users' behavior to differ when interacting with a system which adapts its information density intelligently. This motivates the development of a corpus for data-to-text generation which includes the requisite variation with respect to linguistic difficulty in general and information density in particular.

This paper presents a corpus designed to exhibit variation in information density as well as word choice, enabling the development of adaptive NLG systems which can vary the information density of their utterances based on user needs.

We begin by detailing existing resources for training NLG systems in Section 2. Section 3 details the design and development of the corpus, while Section 4 provides qualitative and statistical descriptions of this corpus. Directions for future work and our conclusions are presented in Section 5.

2. Existing resources for data-driven NLG

Existing resources either (1) provide a fairly flat meaning representation paired with crowdsourced texts or (2) provide hierarchical meaning representations paired with the outputs of a traditional, hand-crafted NLG system. Both types of corpora are fairly limited in terms of lexical variety, although recent work has improved on earlier approaches to crowdsourcing corpora by using images to elicit text [3]. The only publicly-available corpus which we are aware of that contains the richer meaning representation of the second kind is the SPaRKy Restaurant Corpus. Our work here therefore builds on this corpus to extend it in terms of variability of lexicalizations and target audience.

2.1. Crowdsourced corpora

Mairesse et al. [1] collected the BAGEL corpus in order to train a dynamic Bayesian network to generate natural language utterances. The corpus consists of 404 utterances which were manually aligned to 202 ‘dialogue acts’ at the phrase level. These dialogue acts follow the Cambridge University Engineering Department’s standard dialogue act guidelines [15]. The basic format of these acts is: `act_type(slot1=value1;slot2=value2)`, where `act_type` denotes the type of dialogue act (e.g. `inform`ing the user of a fact or the system of a preference) and each of the `slot`s and `value`s represents a property to be expressed (e.g. a venue `name` or the type of `food` served at a restaurant).

For the BAGEL corpus, these dialogue acts are further broken down into a sequence of acts expressing individual properties (i.e. `act_type(slot1=value1)`, `act_type(slot2=value2)`). These individual properties are then aligned to (sequences of) words in the corpus texts in order to make the problem of data-driven NLG tractable using dynamic Bayesian networks on such a small dataset.

The SFX-hotel and SFX-restaurants corpora [2] follow the BAGEL corpus in using a flat semantic structure, but address the size limitations by providing 5k utterances paired with dialogue acts in each of two domains. These dialogue acts are only aligned at the level of whole utterances, no longer decomposing a single large dialogue act into a fixed sequence of semantic stacks, but suffer from a similar lack of lexical variation.

Novikova et al. [3] improved on this lack of variation by using a different elicitation paradigm. Where both the BAGEL and SFX corpora relied on textual representations of the dialogue acts when asking crowdsourced workers to write a new text, Novikova et al. demonstrated that using images to elicit descriptions resulted in greater lexical variation. The portion of their corpus which has already been released contains 1243 utterances paired with dialogue acts.

In order to train a generation system regarding word usage, the system needs to see multiple instances of each word during training. When we consider only words which occur at least five times in the corpus after removing business names, the SFX-hotel dataset contains only 338 words, the SFX-restaurants corpus only 353 words, and the Novikova et al. corpus only 238.

2.2. The SPaRKy Restaurant Corpus

In contrast with these crowdsourced corpora, the SPaRKy Restaurant Corpus (SRC; a.k.a. the MATCH corpus, [9]) includes more complex meaning representations based on Rhetorical Structure Theory (RST, [16]). The SRC consists of 1760 texts produced by the MATCH system using a sentence planner with rhetorical knowledge (SPaRKy) and the meaning representations given as input to this system, as well as the MRs produced at intermediate stages of generation.

A separate content selection module decided which facts to express to users and identified the RST relations between these facts. The output of this process was the *unordered text plans* which served as input to the rest of the NLG system. The system then fixed the relative order of these propositions in a tree-structure before lexicalizing these text plans (i.e. choosing which words to use to express their content) and realizing the actual surface form of the utterance.

> We are adding variety to an existing dialogue system and we need your help!
>
> In this task, you will be given a text about one or more restaurants written by our existing system.
>
> **Your job is to express the same facts, describing the restaurant(s) as you would describe them to your...**
>
> DEFAULT: **friends or family.**
> ELDERLY: **85-year-old grandmother.**

Figure 1: *Instructions to participants on Prolific Academic, with the difference between the* DEFAULT *and* ELDERLY *conditions indicated on a separate line.*

While the crowdsourced corpora tended to focus only on shorter dialogue system responses, the SRC offers a wider variety of texts, in part due to the inclusion of RST relations. The SRC texts include recommendations for individual restaurants as well as comparisons of multiple restaurants. The texts range in length from 1 to 25 sentences, although the majority of texts are no longer than 6 sentences.

The vocabulary of the SRC, however, is severely restricted, with only 99 unique words appearing more than 5 times. Therefore, a new corpus is needed if we are to apply data-driven methods to generate texts with rich meaning representations based on RST with good lexical variety.

3. Corpus design and development

We used a paraphrasing paradigm to collect new utterances based on the text plans of the SRC. As the only available corpus with higher level meaning representations, the SRC provides a natural basis for this work.

Similar to the image-based elicitations of Novikova et al. [3], a paraphrasing task elicits greater lexical variation by reducing the tendency of subjects to copy exact phrases, which occurs when dialogue acts and similar meaning representations are used.[1] In order to elicit variation with respect to text difficulty in general and information density in particular, we manipulated the instructions as shown in Figure 1 and discussed in the next section.

After collecting the paraphrases, we generated dialogue acts and sets of RSTs based on the SRC for all of them. We then manually corrected the annotations for 1344 of these.

3.1. Corpus collection

We deployed the paraphrasing task on the Prolific Academic crowdsourcing platform[2] using the LingoTurk [17] framework. Participants were paid 1 GBP for an average of 7 minutes work. We required participants to be native speakers of English living in English-speaking countries.[3]

Figure 1 shows the instructions to participants. Our instructions targeted variation by emphasizing that participants should re-write texts in their own words. We relied on the phenomenon of ‘elderspeak’ to elicit variation with respect to information density, asking subjects to imagine that they were ad-

[1] In a pilot, we evaluated a tabular format for information presentation and found that subjects tended to copy slot values exactly and use more formulaic syntactic structures than when paraphrasing a text.

[2] `https://www.prolific.ac`

[3] The detailed specifications, along with the LingoTurk experiment we created for paraphrase collection, will be released with the corpus.

```
1. inform(ref=Lemongrass_Grill, price=22)
2. inform(ref=Lemongrass_Grill,
          cuisine=Thai)
3. inform(ref=Monsoon, price=26)
4. inform(ref=Monsoon, cuisine=Vietnamese)

contrast(infer(1,2),infer(3,4))
```

Figure 2: *The propositions or* REFDA*s for the text in Example 1 (numbered) and the text plan for the same.*

dressing their "85-year-old grandmother" in the ELDERLY condition rather than simply addressing their familiars as in the DEFAULT condition.

We chose to use an elderly relative as the audience for this manipulation rather than, e.g., a child, because children are not expected to require the same information as an adult speaker in order to express a preference for where to eat. With an elderly relative, however, our participants could be expected to produce relatively standard adult-directed language while compensating for stereotyped cognitive decline.

Each participant saw only one condition and paraphrased two recommendations of a single restaurant and two comparisons of two restaurants. Each of the 672 SRC texts used as paraphrasing prompts was paraphrased at least 4 times in each condition, yielding more than 2600 texts in each condition.

Given the importance of good text quality for training inputs to an NLG system, we processed these texts to normalize spelling and restaurant name mentions (e.g. correcting 'restuarant', ensuring the restaurant 'Il Mulino' was not abbreviated to 'Mulino'). We also ensured that every sentence begins with a capital letter and ends with a punctuation mark. For the 1344 texts we manually annotated, we further corrected uses of non-standard sentence-final punctuation (e.g. 'run-on' sentences using commas in place of full stops).[4]

3.2. Annotations

We manually annotated 1344 ($\sim$25%) texts with their propositional content and the discourse tree structure over this content. This allowed us to evaluate the accuracy of the basic annotations derived from the SRC, which include their propositional content and merely the *set* of RST relations between them.

We split the annotation task into two stages. First, we annotated the propositions in the text using a variation of the dialogue act representation described in Section 2.1. Then, we connected these propositions in a single-rooted tree structure with RST relations as non-terminal nodes.

For example, consider the following paraphrase from the DEFAULT condition:

(1) Lemongrass Grill is cheaper at 22 dollars, and it serves Thai food, whilst Monsoon is slightly more expensive at 26 dollars and serves Vietnamese food.

In Figure 2, the first dialogue act indicates that the average price at Lemongrass Grill is 22 dollars. We represent each proposition as an individual dialogue act with a `ref` slot to mark which referent is associated with it. These are referred to as REFDAs. Marking referents is necessary because our texts can include multiple referents and each property (i.e. slot-value pair) needs to be associated with the correct one. Marking reference in this way can also enable a system to learn when to apply different strategies for referring expressions (e.g. when to use the full name of a restaurant versus a pronoun).[5]

```
1. inform(ref=Amy's_Bread, quality=best)
2. inform(ref=Amy's_Bread, decor=mediocre)
3. inform(ref=Amy's_Bread, service=decent)
4. inform(ref=Amy's_Bread, price=12)
5. inform(ref=Amy's_Bread,
          food_quality=excellent)

justify-ns(1,contrast(infer(2,3),
                      infer(4,5)))
```

> From the selected restaurants Amy's Bread has the best overall quality since while it has mediocre decor and decent service, its price is only 12 dollars and their food quality is excellent.

Figure 3: REFDA*s, text plan, and text exemplifying the* `justify` *relation.*

After annotating the sequence of REFDAs for the text, we annotate the tree structure connecting them. In Figure 2, this means we first group together properties (1) and (2) under an `infer` relation, and likewise properties (3) and (4). This relation serves as an underspecified discourse relation, in many cases serving as a simple marker of conjunction. More interesting is the `contrast` relation from RST, which indicates that the discourse is contrasting these two sets of propositions. There is also a marker for *implicit* contrast which we used in cases where there is no explicit contrastive connective (as it is in Example 1 with the word 'whilst') but the underlying text plan marks the contrast. This is annotated so that researchers have the flexibility of treating these `icontrasts` as `infer` relations, `contrast` relations, or as their own relation to be learned, depending on their needs.

Following the SRC annotations, we distinguish between nucleus-satellite (`ns`) and satellite-nucleus (`sn`) justifications. The `justify-ns` relation indicates that its first argument is supported by the assertions of its second argument, and vice versa the `justify-sn` relation. This is exemplified in Figure 3. The `elaboration` relation similarly marks that one (set of) propositions elaborates upon the claim made by another.

In addition to these manual annotations, we provide automatically generated sets of propositions and RST relations for each paraphrase based on its SRC source text.

3.2.1. Results of annotations

The original texts for these 1344 paraphrases we manually annotated contained about 6k propositions. An interesting observation we made during annotation was that subjects sometimes altered the content while paraphrasing. Specifically, they changed 580 of the 6000 propositions, e.g., describing a restaurant's decor as 'excellent' instead of merely 'good'. Participants also dropped $\approx$ 320 propositions. The majority of the paraphrases, however, were not affected by these alterations, with 830 of the 1344 we annotated preserving all of the original

[4] The uncorrected paraphrases are also provided with the corpus.

[5] As a convenience, the corpus release includes scripts for converting sequences of REFDAs into (sequences of) CUED-standard dialogue acts as used by [2].

Table 1: *Lexical variation across corpora in the restaurant domain. 'Vocab' is the number of words occurring at least 5 times in the corpus. 'Manual ESRC' is the subset of the data we manually annotated. 'Extended SRC' is our corpus of paraphrases.*

corpus	# texts	Vocab
BAGEL	404	74
SFX-restaurant	5192	353
NLR	1243	238
SRC	1760	99
Manual ESRC	1344	309
Extended SRC	5356	577

contentb We believe that this observation is a relevant potential source of noise also for other types of data-collection for NLG involving paraphrasing or picturing.

4. Corpus statistics

Our corpus consists of more than 5300 texts along with meaning representations based on the paraphrased original SRC source and manually quality-checked and corrected meaning representations for 1344 of these. The corpus exhibits a wide range of lexical variation and variation with respect to information density, word & sentence length, and proposition density.

4.1. Lexical variation

Our corpus contains more than 1500 unique words, more than 500 of which occur 5 or more times. This is a marked improvement of the $\approx$ 65 words occurring in the portion of the original SRC which we used to as prompts in our paraphrasing task. Consider, for example, the relatively stilted way in which price is communicated in the SRC: the average cost of a meal at a restaurant is always described using a genitive determiner phrase modifying the word 'price' and a simple copula. For this one simple property, our corpus includes: 'costs', 'is', 'has food for', 'with a price of', 'is priced at', 'for N dollars you can eat at X', 'expect to pay N dollars', etc. Table 1 compares the vocabulary size to existing resources.

4.2. Variation in information density across texts

Our data collection included an explicit manipulation of the intended audience to elicit variation with respect to text difficulty. Asking participants to address 'their 85-year-old' grandmother was effective in getting them to produce texts with significantly lower average information density, as expected and as reflected in Table 2 and Figure 4. The table also shows that our DEFAULT and ELDERLY subcorpora differ significantly with respect to average sentence length & word length, the average number of facts per sentence, and the average number of sentences per text.

4.3. Meaning representation statistics

While every text in the corpus is associated with set of propositions and RST relations from the SRC, in this section we focus only on the 1344 texts for which we have manual annotations. Our corpus includes $\approx$ 5700 REFDA tokens consisting of $\approx$ 360 unique REFDA types. Converting these into the CUED dialogue act style, we have 570 unique dialogue acts with 2284 dialogue act tokens. Considering only the combination of slots and not the values for these dialogue acts, there are 107 unique dialogue acts of which 56 occur at least 5 times in the corpus.

Table 2: *Properties of the texts in the* DEFAULT *and* ELDERLY *conditions, with the significance of the differences between the means based on Welch's* t*-test. n.s. = not significant*

	Average Value		
text property	DEFAULT	ELDERLY	p
info density	9.11	8.90	$< 10^{-8}$
word length	4.27	4.16	$< 10^{-15}$
sentence length	17.8	16.5	$< 10^{-13}$
# words / REFDA	7.49	7.35	0.2 (n.s.)
# REFDAs / sentence	2.52	2.33	< 0.001
# of sentences / text	1.84	1.98	$< 10^{-8}$
# of REFDAs / text	4.23	4.21	0.8 (n.s.)

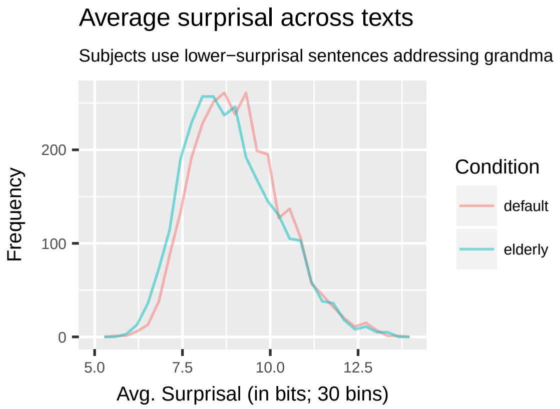

Figure 4: *Relative frequency of different average surprisals across the texts in our corpus.*

At the level of the tree structured text plans, we have 187 unique tree structures, when we ignore the labels of the leaf nodes, which are the individual REFDAs. If we collapse all leaf nodes into their parent, we have 102 unique tree structures, of which 34 occur at least five times in the corpus.

5. Conclusion

We have presented a new corpus of 5356 texts for training generation systems, supplementing the 1760 texts of the SPaRKy Restaurant Corpus with greatly improved lexical variation. Building on the SRC facilitated the inclusion of a more complex meaning representation in addition to flat semantic stacks, enabling work on training generation systems with a more complex discourse structure. A central contribution of our work is the fact that our corpus contains different types of addressees, and as a result varies in complexity. Moreover, the corpus includes information density estimates for all utterances, allowing the development of NLG systems which adapt the information density of their utterances for different users and situations.

6. Acknowledgements

We thank Florian Pusse for extending and helping with LingoTurk, Merel Scholman for feedback on annotation guidelines, and Stalin Varanasi & Ben Peters for testing the guidelines. Thanks also go to Alyssa Price & Mittul Singh for proofreading. This work was supported by the DFG through SFB 1102 'Information Density and Linguistic Encoding'.

7. References

[1] F. Mairesse, M. Gasic, F. Jurcicek, S. Keizer, B. Thomson, K. Yu, and S. Young, “Phrase-based statistical language generation using graphical models and active learning,” Uppsala, Sweden, pp. 1552–1561, July 2010. [Online]. Available: http://www.aclweb.org/anthology/P10-1157

[2] T.-H. Wen, M. Gasic, N. Mrkšić, P.-H. Su, D. Vandyke, and S. Young, “Semantically conditioned lstm-based natural language generation for spoken dialogue systems,” Lisbon, Portugal, pp. 1711–1721, September 2015. [Online]. Available: http://aclweb.org/anthology/D15-1199

[3] J. Novikova, O. Lemon, and V. Rieser, *Proc. of the 9th International Natural Language Generation conference*. Association for Computational Linguistics, 2016, ch. Crowd-sourcing NLG Data: Pictures Elicit Better Data., pp. 265–273. [Online]. Available: http://aclweb.org/anthology/W16-6644

[4] F. A. Drews, M. Pasupathi, and D. L. Strayer, “Passenger and cell phone conversations in simulated driving,” *Journal of Experimental Psychology: Applied*, vol. 14, no. 4, pp. 392–400, 2008.

[5] J. G. Gaspar, W. N. Street, M. B. Windsor, R. Carbonari, H. Kaczmarski, A. F. Kramer, and K. E. Mathewson, “Providing Views of the Driving Scene to Drivers’ Conversation Partners Mitigates Cell-Phone-Related Distraction,” *Psychological Science*, vol. 25, no. 12, pp. 2136–2146, 2014.

[6] E. Becic, G. S. Dell, K. Bock, S. M. Garnsey, T. Kubose, and A. F. Kramer, “Driving impairs talking,” *Psychonomic Bulletin & Review*, vol. 17, no. 1, pp. 15–21, 2010.

[7] F. Mairesse and M. A. Walker, “PERSONAGE: Personality Generation for Dialogue,” pp. 496–503, 2007. [Online]. Available: http://acl.ldc.upenn.edu/P/P07/P07-1063.pdf

[8] M. A. Walker, S. Whittaker, and A. Stent, “Generation and evaluation of user tailored responses in dialogue,” *Cognitive Science*, vol. 28, no. 5, pp. 811–840, 2004.

[9] M. A. Walker, A. Stent, F. Mairesse, and R. Prasad, “Individual and domain adaptation in sentence planning for dialogue,” *Journal of Artificial Intelligence Research*, vol. 30, pp. 413–456, 2007.

[10] V. Demberg, A. Winterboer, and J. D. Moore, “A strategy for information presentation in spoken dialog systems,” *Computational Linguistics*, vol. 37, no. 3, pp. 489–539, 2011.

[11] V. Demberg and F. Keller, “Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity,” *Cognition*, vol. 109, no. 2, pp. 193–210, 2008.

[12] D. Jurafsky, A. Bell, M. Gregory, and W. D. Raymond, *Probabilistic Relations between Words: Evidence from Reduction in Lexical Production*. John Benjamins, 2001, vol. 45, pp. 229–254.

[13] W. D. Raymond, R. Dautricourt, and E. Hume, “Word-internal /t,d/-deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors,” *Language Variation and Change*, vol. 18, no. 01, pp. 55–97, 2006.

[14] “Speakers optimize information density through syntactic reduction,” *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[15] S. Young, “Cued standard dialogue acts,” Cambridge University Dialogue Systems Group, Tech. Rep., 2009.

[16] W. C. Mann and S. A. Thompson, “Rhetorical structure theory: Toward a functional theory of text organization,” *Text: Interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 243–281, 1988.

[17] F. Pusse, A. Sayeed, and V. Demberg, “LingoTurk: managing crowdsourced tasks for psycholinguistics,” San Diego, California, pp. 57–61, June 2016. [Online]. Available: http://www.aclweb.org/anthology/N16-3012