Offline Sentence Processing Measures for testing Readability with Users

Advaith Siddharthan
Department of Computing Science
University of Aberdeen
advaith@abdn.ac.uk

Napoleon Katsos
Department of Theoretical and Applied Linguistics
University of Cambridge
nk248@cam.ac.uk

Abstract

While there has been much work on computational models to predict readability based on the lexical, syntactic and discourse properties of a text, there are also interesting open questions about how computer generated text should be evaluated with target populations. In this paper, we compare two offline methods for evaluating sentence quality, magnitude estimation of acceptability judgements and sentence recall. These methods differ in the extent to which they can differentiate between surface level fluency and deeper comprehension issues. We find, most importantly, that the two correlate. Magnitude estimation can be run on the web without supervision, and the results can be analysed automatically. The sentence recall methodology is more resource intensive, but allows us to tease apart the fluency and comprehension issues that arise.

1 Introduction

In Natural Language Generation, recent approaches to evaluation tend to consider either "naturalness" or "usefulness". Following evaluation methodologies commonly used for machine translation and summarisation, there have been attempts to measure naturalness in NLG by comparison to human generated gold standards. This has particularly been the case in evaluating referring expressions, where the generated expression can be treated as a set of attributes and compared with human generated expressions (Gatt et al., 2009; Viethen and Dale, 2006), but there have also been attempts at evaluating sentences this way. For instance, Langkilde-Geary (2002) generate sentences from a parsed analysis of an existing sentence, and evaluate by comparison to the original. However, this approach has been criticised at many levels (see for example, Gatt et al. (2009) or Sripada et al. (2003)); for instance, because there are many good ways to realise a sentence, because typical NLG tasks do not come with reference sentences, and because fluency judgements in the monolingual case are more subtle than for machine translation.

Readability metrics, by comparison, do not rely on reference texts, and try to model the linguistic quality of a text based on features derived from the text. This body of work ranges from the Flesch Metric (Flesch, 1951), which is based on average word and sentence length, to more systematic evaluations of various lexical, syntactic and discourse characteristics of a text (cf. Pitler et al. (2010), who assess readability of textual summaries). Some researchers have also suggested measuring edit distance by using a human to revise a system generated text and quantifying the revisions made (Sripada et al., 2003). This does away with the need for reference texts and is quite suited to expert domains such as medicine or weather forecasting, where a domain expert can easily correct system output. Analysis of these corrections can provide feedback on problematic content and style. We have previously evaluated text reformulation applications by asking readers which version they prefer (Siddharthan et al., 2011), or through the use of Likert scales (Likert, 1932) for measuring meaning preservation and grammaticality (Siddharthan, 2006). However, none of these approaches tell us very much about the comprehensibility of a text for an end reader.

To address this, there has been recent interest in task based evaluations. Task based evaluations directly evaluate generated utterances for their utility

NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations (PITR 2012)., pages 17–24,
Montréal, Canada, June 7, 2012.

to the hearer. However, while for some generation areas like reference (Gatt et al., 2009), the real world evaluation task is obvious, it is less so for other generation tasks such as surface realisation or text-to-text regeneration or paraphrase. We are thus keen to investigate psycholinguistic methods for investigating sentence processing as an alternative to task based evaluations.

In the psycholinguistics literature, various offline and online techniques have been used to investigate sentence processing by readers. Online techniques (eye-tracking (Duchowski, 2007), neurophysiological (Friederici, 1995), etc.) offer many advantages in studying how readers process a sentence. But as these are difficult to set up and also resource intensive, we would prefer to evaluate NLG using offline techniques. Some offline techniques, such as Cloze tests (Taylor, 1953) or question answering, require careful preparation of material (choice of texts and questions, and for Cloze, the words to leave out). Other methods, such as magnitude estimation and sentence recall (cf. Sec 3 for details), are more straightforward to implement. In this paper, we investigate magnitude estimation of acceptability judgements and delayed sentence recall in the context of an experiment investigating generation choices when realising causal relations. Our goal is to study how useful these methods are for evaluating surface level fluency and deeper comprehensibility. We are interested in whether they can distinguish between similar sentences, and whether they can be used to test hypotheses regarding the effect of common generation decisions such as information order and choice of discourse marker. We briefly discuss the data in Section 2, before describing our experiments (Sections 3.1 and 3.2). We finish with a discussion of their suitability for more general evaluation of NLG with target readers.

2 Data

We use a dataset created to explore generation choices in the context of expressing causal relations; specifically, the choice of periphrastic causative (Wolff et al., 2005) and information order. The dataset considers four periphrastic causatives (henceforth referred to as discourse markers): "*because*", "*because of*", the verb "*cause*" and the noun "*cause*" with different lexico-syntactic properties. We present an example from this dataset below (cf. Siddharthan and Katsos (2010) for details):

(1) a. Fructose-induced hypertension **is caused by** increased salt absorption by the intestine and kidney. **[b_caused-by_a]**

b. Increased salt absorption by the intestine and kidney **causes** fructose-induced hypertension. **[a_caused_b]**

c. Fructose-induced hypertension occurs **because of** increased salt absorption by the intestine and kidney. **[b_because-of_a]**

d. **Because of** increased salt absorption by the intestine and kidney, fructose-induced hypertension occurs. **[because-of_ab]**

e. Fructose-induced hypertension occurs **because** there is increased salt absorption by the intestine and kidney. **[b_because_a]**

f. **Because** there is increased salt absorption by the intestine and kidney, fructose-induced hypertension occurs. **[because_ab]**

g. Increased salt absorption by the intestine and kidney is the **cause of** fructose-induced hypertension. **[a_cause-of_b]**

h. The **cause of** fructose-induced hypertension is increased salt absorption by the intestine and kidney. **[cause-of_ba]**

In this notation, "a" represents the cause,"b" represents the effect and the remaining string indicates the discourse marker; their ordering reflects the information order in the sentence, for example, "a_cause-of_b" indicates a cause-effect information order using "cause of" as the discourse marker. The dataset consists of 144 sentences extracted from corpora (18 sentences in each condition (discourse marker + information order), reformulated manually to generated the other seven conditions, resulting in 1152 sentences in total.

Clearly, different formulations have different levels of fluency. In this paper we explore what two offline sentence processing measures can tell us about their acceptability and ease of comprehension.

3 Method

3.1 Magnitude estimation of acceptability

Human judgements for acceptability for each of the 1152 sentences in the dataset were obtained using the WebExp package (Keller et al., 2009). Note that

the reformulations are, strictly speaking, grammatical according to the authors' judgement. We are testing violations of acceptability, rather than grammaticality per se. This mirrors the case of NLG, where a grammar is often used for surface realisation, ensuring grammaticality.

Acceptability is a measure which reflects both ease of comprehension and surface well-formedness. We later compare this experiment with a more qualitative comprehension experiment based on sentence recall (cf. Section 3.2). Rather than giving participants a fixed scale, we used the magnitude estimation paradigm, which is more suitable to capture robust or subtle differences between the relative strength of acceptability or grammaticality violations (see, for example, Bard et al. (1996); Cowart (1997); Keller (2000)). One advantage of magnitude estimation is that the researcher does not make any assumptions about the number of linguistic distinctions allowed. Each participant makes as many distinctions as they feel comfortable. Participants were given the following instructions (omitting those that relate to the web interface):

1. Judge acceptability of construction, not of meaning;
2. There is no limit to the set of numbers you can use, but they must all be positive - the lowest score you can assign is 0. In other words, make as many distinctions as you feel comfortable;
3. Always score the new sentence relative to the score you gave the modulus sentence, which you will see on the top of the screen;
4. Acceptability is a continuum, do not just make yes/no judgements on grammaticality;
5. Try not to use a fixed scale, such as 1–5, which you might have used for other linguistic tasks previously.

Design: The propositional content of 144 sentences was presented in eight conditions. Eight participant groups (A–H) consisting of 6 people each were presented with exactly one of the eight formulations of each of 144 different sentences, as per a Latin square design. This experimental design allows all statistical comparisons between the eight types of causal formulations and the three genres to be within-participant. The participants were University of Cambridge students (all native English speakers). Participants were asked to score how acceptable a modulus sentence was, using any positive number. They were then asked to score other sentences relative to this modulus, so that higher scores were assigned to more acceptable sentences. Scores were normalised to allow comparison across participants, following standard practice in the literature, by using the z-score: For each participant, each sentence score was normalised so that the mean score is 0 and the standard deviation is 1 ($z_{ih} = \frac{x_{ih} - \mu_h}{\sigma_h}$), where z_{ih} is participant h's z-score for the sentence i when participant h gave a magnitude estimation score of x_{ih} to that sentence. μ_h is the mean and σ_h the standard deviation of the set of magnitude estimation scores for user h.

3.2 Sentence Recall

Acceptability ratings are regarded as a useful measure because they combine surface judgements of grammaticality with deeper judgements about how easy a sentence is to understand. However, one might want to know whether an inappropriate formulation can cause a breakdown in comprehension of the content of a sentence, which would go beyond the (perhaps) non-detrimental effect of a form that is dispreferred at the surface level. To try and learn more about this, we conducted a second behavioural experiment using a sentence recall methodology. As these experiments are harder to conduct and have to be supervised in a lab (to ensure that participants have similar conditions of attention and motivation, and to prevent "cheating" using cut-and-paste or note taking techniques), we selected a subset of 32 pairs of items from the previous experiment. Each pair consisted of two formulations of the same sentence. The pairs were selected in a manner that exhibited a variation in the within-pair difference of acceptability. In other words, we wanted to explore whether two formulations of a sentences with similar acceptability ratings were recalled equally well and whether two formulations of a sentence with different acceptability ratings were recalled differently.

Design: 32 students at the University of Cambridge were recruited (these are different partici-

pants from those in the acceptability experiment in Section 3.1, but were also all native speakers). We created four groups A–D, each with eight participants. Each Group saw 16 sentences in exactly one of the two formulation types, such that groups A–B formed one Latin square and C–D formed another Latin square. These 16 sentences were interleaved with 9 filler sentences that did not express causality. For each item, a participant was first shown a sentence on the screen at the rate of 0.5 seconds per word. Then, the sentence was removed from the screen, and the participant was asked to do two arithmetic tasks (addition and subtraction of numbers between 10 and 99). The purpose of these tasks was to add a load between target sentence and recall so that the recall of the target sentence could not rely on internal rehearsal of the sentence. Instead, research suggests that in such conditions recall is heavily dependent on whether the content and form was actually comprehended (Lombardi and Potter, 1992; Potter and Lombardi, 1990). Participants then typed what they recalled of the sentence into a box on the screen.

We manually coded the recalled sentences for six error types (1–6) or perfect recall (0) as shown in Table 1. Further, we scored the sentences based on our judgements of how bad each error-type was. The table also shows the weight for each error type. For any recalled sentences, only one of (0,1,5,6) is coded, i.e., these codes are mutually exclusive, but if none of the positive scores (0,1,5,6) have been coded, any combination of error types (2,3,4) can be coded for the same sentence.

4 Results

4.1 Correlation between the two methods

The correlation between the differences in acceptability (using average z-scores for each formulation from the magnitude estimation experiment) and recall scores (scored as described above) for the 32 pairs of sentences was found to be significant (Spearman's rho=.43; p=.01). A manual inspection of the data showed up one major issue regarding the methodologies: our participants appear to penalise perceived ungrammaticalities in short sentences quite harshly when rating acceptability, but they have no trouble recalling such sentences accurately. For example, sentence a. in Example 2 below had an average acceptability score of 1.41, while sentence b. only scored .13, but both sentences were recalled perfectly by all participants in the recall study:

(2) a. It is hard to imagine that it was the cause of much sadness.

b. It is hard to imagine that because of it there was much sadness.

Indeed the sentence recall test failed to discriminate at all for sentences under 14 words in length. When we removed pairs with sentences under 14 words (there were eight such pairs), the correlation between the differences in magest and recall scores for the 24 remaining pairs of sentences was even stronger (Spearman's rho=.64; $p<.001$).

Summary: The two methodologies give very different results for short sentences. This is because comprehension is rarely an issue for short sentences, while surface level disfluencies are more jarring to participants in such short sentences. For longer sentences, the two methods correlate strongly; for such sentences, magnitude estimations of acceptability better reflect ease of comprehension. In retrospect, this suggests that the design of an appropriate load (we used two arithmetic sums) is an important consideration that can affect the usefulness of recall measures. One could argue that acceptability is a more useful metric for evaluating NLG as it combines surface level fluency judgements with ease of comprehension issues. In Siddharthan and Katsos (2010), we described how this data could be used to train an NLG component to select the most acceptable formulation of a sentence expressing a causal relation. We now enumerate other characteristics of magnitude estimation of acceptability that make them useful for evaluating sentences. Then, in Section 4.3, we discuss what further information can be gleaned from sentence recall studies.

4.2 Results of magnitude estimation study

Distinguishing between sentences: We found that magnitude estimation judgements are very good at distinguishing sentences expressing the same content. Consider Table 2, which shows the average acceptability for the n-best formulation of each of the

Weight	Error Code	Error Description
+0.5	0	Recalled accurately (clauses A and B can be valid paraphrases, but the discourse connective (TYPE) is the same)
+0.4	1	Clauses A and B are recalled accurately but the relation is reformulated using a different *but valid* discourse marker
-0.25	2	The discourse marker has been changed in a manner that modifies the original causal relation
-0.5	3	Clause B (effect) recall error (clause is garbled)
-0.5	4	Clause A (cause) recall error (clause is garbled)
+0.25	5	Causal relation and A and B are recalled well, but some external modifying clause is not recalled properly
+0.25	6	Causality is quantified (e.g., "major cause") and this modifier is lost or changed in recall (valid paraphrases are not counted here)

Table 1: Weighting function for error types.

144 sentences (n=1–8). We see that the best formulation averages .89, the second best .57 and the worst formulation -.90. Note that it is not always the same formulation types that are deemed acceptable – if we always select the most preferred type (a_caused_b) for each of the 144 sentences, the average acceptability is only .12.

n =	1	2	3	4	5	6	7	8
Av. Z =	.89	.57	.33	.13	-.12	-.33	-.58	-.90

Table 2: Average acceptability for the n^{th} best formulation of each of the 144 sentences.

Testing hypotheses: In addition to distinguishing between different formulations of a sentence, varying generation choices systematically allows us to test any hypotheses we might have about their effect on acceptability. Indeed, hypothesis testing was an important consideration in the design of this experiment. For instance, various studies (Clark and Clark, 1968; Katz and Brent, 1968; Irwin, 1980) suggest that for older school children, college students and adults, comprehension is better for the cause-effect presentation, both when the relation is implicit (no discourse marker) and explicit (with a discourse marker). We can then test specific predictions about which formulations are likely to be more acceptable.

H1 We expect the cause-effect information order to be deemed more acceptable than the corresponding effect-cause information order.

H2 As all four discourse markers are commonly used in language, we do not expect any particular marker to be globally preferred to the others.

We ran a 4 (discourse marker) x 2 (information order) repeated measures ANOVA. We found a main effect of information order ($F1(1, 49) = 5.19$, $p = .017$) and discourse marker ($F1(3, 147) = 3.48$, $p = .027$). Further, we found a strong interaction between information order and formulation type, $F1(3, 147) = 19.17$, $p<.001$. We now discuss what these results mean.

Understanding generation decisions: The main effect of discourse marker was not predicted (Hypothesis H2). We could try and explain this empirically. For instance, in the BNC "because" as a conjunction occurs 741 times per million words, while "cause" as a verb occurs 180 times per million words, "because of" 140 per million words and "cause" as a noun 86 per million words. We might expect the more common markers to be judged more acceptable. However, there was no significant correlation between participants' preference for discourse marker and the BNC corpus frequencies of the markers (Spearman's rho=0.4, $p>0.75$). This suggests that corpus frequencies need not be a reliable indicator of reader preferences, at least for discourse connectives. The mean z-scores for the four discourse markers are presented in Table 3

To explore the interaction between discourse marker and information order, a post-ANOVA Tukey HSD analysis was performed. The significant

Discourse Marker	Average Z-score
Cause (verb)	0.036
Because of	0.028
Because	-0.011
Cause (noun)	-0.028

Table 3: Av. z-scores for the four discourse markers

effects are listed in Table 4. There is a significant preference for using "because" and "because of" in the effect-cause order (infix) over the cause-effect order (prefix) and for using "cause" as a verb in the cause-effect order (active voice) over the effect-cause order (passive voice). Thus, hypothesis H1 is not valid for "because" and "because of", where the canonical infix order is preferred, and though there are numerical preferences for the cause-effect order for "cause" as a noun we found support for hypothesis H1 to be significant only for "cause" as a verb. Table 4 also tells us that if the formulation is in cause-effect order, there is a preference for "cause" as a verb over "because" and "because of". On the other hand, if the formulation is in the reverse effect-cause order, there is a preference for "because" or "because of" over "cause" as a verb or as a noun.

Summary: This evaluation provides us with some insights into how generation decisions interact, which can be used prescriptively to, for example, select a discourse marker, given a required information order.

4.3 Results of sentence recall study

While magnitude estimation assessments of acceptability can be used to test some hypotheses about the effect of generation decisions, it cannot really tease apart cases where there are surface level disfluencies from those that result in a breakdown in comprehension. To test such hypotheses, we use the sentence recall study.

Testing hypotheses: Previous research (e.g., Engelkamp and Rummer (2002)) suggests that recall for the second clause is worse when clauses are combined through coordination (such as "therefore" or "and") than through subordination such as "because". The explanation is that subordination better unifies the two clauses in immediate memory. We would expect this unification to be even greater when the cause and effect are arguments to a verb. Thus, compared to "because", we would expect recall of the second clause to be higher for "cause" as a verb or a noun, due to the tighter syntactic binding to the discourse marker (object of a verb). Likewise, compared to "cause", we would expect to see more recall errors for the second clause when using "because" as a conjunction. Our hypotheses are listed below:

H3 For "cause" as a verb or a noun, there will be fewer recall errors in "a" and "b" compared to "because" or "because of", because of the tighter syntactic binding.

H4 For "because" as a conjunction, there will be more recall errors in the second clause than in the first clause; i.e., for "b_because_a", clause "a" will have more recall errors than "b" and for "because_ab", clause "a" will have fewer recall errors than "b".

Table 5 shows the average incidence of each error type per sentences in that formulation (cf. Table 1). Note that the totals per row might add up to slightly more than 1 because multiple errors can be coded for the same sentence.

Table 5 shows that "because" and "because of" constructs result in more type 3 and 4 recall errors in clauses "a" and/or "b" compared with "cause" as either a noun or a verb. This difference is significant (z-test; $p<.001$), thus supporting hypothesis H3.

Further, for "because", the recall errors for the first clause are significantly fewer than for the second clause (z-test; $p<.01$), thus supporting hypothesis H4. In contrast, for the cases with "cause" as a verb or noun, both A and B are arguments to a verb (either "cause" or a copula), and the tighter syntactic binding helps unify them in immediate memory, resulting in fewer recall errors that are also distributed more evenly between the first and the second argument to the verb.

We make one further observation: passive voice sentences appear to be reformulated at substantial levels (19%), but in a valid manner (type 1 errors). This suggests that the dispreference for passives in the acceptability study is about surface level form rather than deeper comprehension. This would be a

(a) Ordering Effects

Marker	Preference	p-value
because	**effect-cause (.12) is preferred over cause-effect (-.14)**	p<.001
because of	**effect-cause (.13) is preferred over cause-effect (-.11)**	p<.001
cause (verb)	**cause-effect (.12) is preferred over effect-cause (-.05)**	p=.0145
cause (noun)	cause-effect (.01) is preferred over effect-cause (-.11)	p=.302

(b) Discourse Marker Effects

Order	Preference	p-value
effect-cause	**'because' (.12) is preferred over 'cause (noun)' (-.11)**	p<.001
effect-cause	**'because-of' (.13) is preferred over 'cause (noun)' (-.11)**	p<.001
effect-cause	**'because-of' (.13) is preferred over 'cause (verb)' (-.05)**	p=.001
effect-cause	**'because' (.12) is preferred over 'cause (verb)' (-.05)**	p=.002
effect-cause	'cause (verb)' (-.05) is preferred over 'cause (noun)' (-.11)	p=.839
effect-cause	'because' (.12) is preferred over 'because-of' (.13)	p=.999
cause-effect	**'cause (verb)' (.13) is preferred over 'because' (-.14)**	p<.001
cause-effect	**' cause (verb)' (.13) is preferred over 'because-of' (-.06)**	p=.006
cause-effect	'cause (verb)' (.13) is preferred over 'cause (noun)' (.01)	p=.165
cause-effect	'cause (noun)' (.01) is preferred over 'because' (-.14)	p=.237
cause-effect	'because-of' (-.06) is preferred over 'because' (-.14)	p=.883
cause-effect	'cause (noun)' (.01) is preferred over 'because-of' (-.06)	p=.961

Table 4: Interaction effects between information order and discourse marker (mean z-scores in parentheses; significant effects in bold face).

reasonable conclusion, given that all our participants are university students.

Summary: Overall we conclude that sentence recall studies provide insights into the nature of the comprehension problems encountered, and they corroborate acceptability ratings in general, and particularly so for longer sentences.

5 Conclusions

In this paper, we have tried to separate out surface form aspects of acceptability from breakdowns in comprehension, using two offline psycholinguistic methods.

We believe that sentence recall methodologies can substitute for task based evaluations and highlight breakdowns in comprehension at the sentence level. However, like most task based evaluations, recall experiments are time consuming as they need to be conducted in a supervised setting. Additionally, they require manual annotation of error types, though perhaps this could be automated.

Acceptability ratings on the other hand are easy to acquire. Based on our experiments, we believe that acceptability ratings are reliable indicators of comprehension for longer sentences and, particularly for shorter sentences, combine surface form judgements with ease of comprehension in a manner that is very relevant for evaluating sentence generation or regeneration, including simplification.

Both methods are considerably easier to set up and interpret than online methods such as self paced reading, eye tracking or neurophysiological methods.

Acknowledgements

This work was supported by the Economic and Social Research Council (Grant Number RES-000-22-3272).

References

E.G. Bard, D. Robertson, and A. Sorace. 1996. Magnitude estimation for linguistic acceptability. *Language*, 72(1):32–68.

H.H. Clark and E.V. Clark. 1968. Semantic distinctions and memory for complex sentences. *The Quarterly Journal of Experimental Psychology*, 20(2):129–138.

type	err0	err1	err2	err3	err4	err5	err6
b_because_a	0.62	0.18	0.02	0.10	0.18	0.00	0.05
because_ab	0.80	0.03	0.03	0.20	0.10	0.00	0.00
b_because-of_a	0.78	0.11	0.02	0.06	0.07	0.00	0.06
because-of_ab	0.73	0.00	0.00	0.17	0.17	0.10	0.00
a_cause-of_b	0.89	0.04	0.06	0.00	0.00	0.00	0.07
cause-of_ba	0.75	0.06	0.04	0.06	0.08	0.00	0.06
a_caused_b	0.83	0.05	0.02	0.03	0.03	0.07	0.00
b_caused-by_a	0.77	0.19	0.00	0.00	0.04	0.00	0.02

Table 5: Table of recall errors per type.

W. Cowart. 1997. *Experimental Syntax: applying objective methods to sentence judgement*. Thousand Oaks, CA: Sage Publications.

A.T. Duchowski. 2007. *Eye tracking methodology: Theory and practice*. Springer-Verlag New York Inc.

J. Engelkamp and R. Rummer. 2002. Subordinating conjunctions as devices for unifying sentences in memory. *European Journal of Cognitive Psychology*, 14(3):353–369.

Rudolf Flesch. 1951. *How to test readability*. Harper and Brothers, New York.

A.D. Friederici. 1995. The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and language*, 50(3):259–281.

A. Gatt, A. Belz, and E. Kow. 2009. The tuna-reg challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European workshop on natural language generation*, pages 174–182. Association for Computational Linguistics.

J.W. Irwin. 1980. The effects of explicitness and clause order on the comprehension of reversible causal relationships. *Reading Research Quarterly*, 15(4):477–488.

E.W. Katz and S.B. Brent. 1968. Understanding connectives. *Journal of Verbal Learning & Verbal Behavior*.

F. Keller, S. Gunasekharan, N. Mayo, and M. Corley. 2009. Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, 41(1):1.

Frank Keller. 2000. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, University of Edinburgh.

I. Langkilde-Geary. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 12th International Natural Language Generation Workshop*, pages 17–24. Citeseer.

R. Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

L. Lombardi and M.C. Potter. 1992. The regeneration of syntax in short term memory* 1. *Journal of Memory and Language*, 31(6):713–733.

E. Pitler, A. Louis, and A. Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554. Association for Computational Linguistics.

M.C. Potter and L. Lombardi. 1990. Regeneration in the short-term recall of sentences* 1. *Journal of Memory and Language*, 29(6):633–654.

Advaith Siddharthan and Napoleon Katsos. 2010. Reformulating discourse connectives for non-expert readers. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010)*, Los Angeles, CA.

A. Siddharthan, A. Nenkova, and K. McKeown. 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, 37(4):811–842.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

S. Sripada, E. Reiter, and I. Davy. 2003. SumTime-Mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10.

W.L. Taylor. 1953. " cloze procedure": a new tool for measuring readability. *Journalism Quarterly; Journalism Quarterly*.

Jette Viethen and Robert Dale. 2006. Algorithms for generating referring expressions: do they do what people do? In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 63–70.

P. Wolff, B. Klettke, T. Ventura, and G. Song. 2005. Expressing causation in English and other languages. *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin*, pages 29–48.