Reformulating Discourse Connectives for Non-Expert Readers

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

Napoleon Katsos
Research Centre for English and Applied Linguistics
University of Cambridge
nk248@cam.ac.uk

Abstract

In this paper we report a behavioural experiment documenting that different lexico-syntactic formulations of the discourse relation of *causation* are deemed more or less acceptable by different categories of readers. We further report promising results for automatically selecting the formulation that is most appropriate for a given category of reader using supervised learning. This investigation is embedded within a longer term research agenda aimed at summarising scientific writing for lay readers using appropriate paraphrasing.

1 Introduction

There are many reasons why a speaker/writer would want to choose one formulation of a discourse relation over another; for example, maintaining thread of discourse, avoiding shifts in focus and issues of salience and end weight. There are also reasons to use different formulations for different audiences; for example, to account for differences in reading skills and domain knowledge. In this paper, we present a psycholinguistic experiment designed to illuminate the factors that determine the appropriateness of particular realisations of discourse relations for different audiences. The second part of this paper focuses on training a natural language generation system to predict which realisation choices are more felicitous than others for a given audience. Our paraphrases include eight different constructions. Consider 1a.–d. below:

(1) a. Tom ate **because** he was hungry.
b. Tom ate **because of** his hunger.
c. Tom's hunger **caused** him to eat.
d. The **cause** of Tom's eating was his hunger.

These differ in terms of the lexico-syntactic properties of the discourse marker (shown in bold font). Indeed the discourse markers here are conjunctions, prepositions, verbs and nouns. As a consequence the propositional content is expressed either as a clause or a noun phrase ("*he was hungry*" vs "*his hunger*", etc.). Additionally, the order of presentation of propositional content can be varied to give four more lexico-syntactic paraphrases:

(1) e. **Because** Tom was hungry, he ate.
f. **Because of** his hunger, Tom ate.
g. Tom's eating **was caused by** his hunger.
h. Tom's hunger was the **cause** of his eating.

It is clear that some formulations of this propositional content are more felicitous than others; for example, 1a. seems preferable to 1d., but for a different propositional content, other formulations might be more felicitous (for instance, example 4, section 3.1, where the passive seems in fact preferable). While discourse level choices based on information ordering play a role in choosing a formulation, it is of particular interest to us that some decontextualised information orderings within a sentence are deemed unacceptable. Any summarisation task that considers discourse coherence should not introduce sentence-level unacceptability.

We now summarise our main research questions:

1. Are some formulations of a discourse relation more felicitous than others, given the same propositional content?
2. Does the reader's level of domain expertise affect their preferred formulation?
3. What linguistic features determine which formulations are acceptable?
4. How well can a natural language generator be trained to predict the most appropriate formulation for a given category of reader?

In this paper, we focus on causal relations because these are pervasive in science writing and are integral to how humans conceptualise the world. The 8 formulations selected are 2 information orderings

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 1002–1010,
Los Angeles, California, June 2010.

of 4 different syntactic constructs; thus we explore a fairly broad range of constructions.

With regard to genre, we have a particular interest in scientific writing, specifically biomedical texts. Reformulating such texts for lay audiences is a highly relevant task today and many news agencies perform this service; e.g., Reuters Health summarises medical literature for lay audiences and BBC online has a Science/Nature section that reports on science. These services rely either on press releases by scientists and universities or on specialist scientific reporters, thus limiting coverage of a growing volume of scientific literature in a digital economy. Thus, reformulating technical writing for lay audiences is a research area of direct relevance to information retrieval, information access and summarisation systems.

At the same time, while there are numerous studies about the effect of text reformulation on people with different literacy levels or language deficits (see section 2), the issue of expert vs lay audiences has received less attention. Further, most studies focus on narrative texts such as news or history. However, as Linderholm et al. (2000) note, results from studies of causality in narrative texts might not carry over to scientific writing, because inferences are made more spontaneously during the reading of narrative than expository texts. Thus comparing expert vs lay readers on the comprehension of causal relations in scientific writing is a most timely investigation.

In section 2, we relate our research to the existing linguistic, psycholinguistic and computational literature. Then in section 3, we describe our psycholinguistic experiment that addresses our first two research questions and in section 4 we present a computational approach to learning felicitous paraphrases that addresses the final two questions.

2 Background and related work

2.1 Expressing causation

Linguists generally consider five different components of meaning (Wolff et al., 2005) in causal expressions: (a) occurrence of change in patient, (b) specification of endstate, (c) tendency and concordance, (d) directness and (e) mechanism. The expressions we consider in this paper, "because" (**conjunction**), "because of" (**preposition**) and "cause" as noun or verb (**periphrastic causatives**) express (a), (b) and in some instances, (c). This is in contrast to **affect verbs** that only express (a), **link verbs** that express (a–b), **lexical causatives** that express (a–d) and **resultatives** that express (a–e). These distinctions are illustrated by the sentences in example 2:

(2) a. Sara **kicked** the door. (affect verb – end state not specified)

b. The door's **breaking** was linked to Sara. (link verb – end state specified, but unclear that door has a tendency to break)

c. Sara **caused** the door to break. / The door broke **because of** Sara. (periphrastic / preposition – indirect; the door might have a tendency to break)

d. Sara **broke** the door. (lexical causative – directness of action is specified)

e. Sara **broke** the door **open**. (resultative – end state is "open")

There is much literature on how people prefer one type of causative over the other based on these five components of meaning (e.g. see Wolff et al. (2005)). What is less understood is how one selects between various expressions that carry similar semantic content. In this paper we consider four constructs "because of", "because", and "cause" as a verb and a noun. These express the components of meaning (a–c) using different syntactic structures. By considering only these four lexically similar constructs, we can focus on the role of the lexis and of syntax in determining the most felicitous expression of causation for a given propositional content.

2.2 Discourse connectives and comprehension

Previous work has shown that when texts have been manually rewritten to make discourse relations such as *causation* explicit, reading comprehension is significantly improved in middle/high school students (Beck et al., 1991). Further, connectives that permit pre-posed adverbial clauses have been found to be difficult for third to fifth grade readers, even when the order of mention coincides with the causal (and temporal) order; for instance, 3b. is more accessible than 3a. (e.g. from Anderson and Davison (1988)).

(3) a. **Because** Mexico allowed slavery, many Americans and their slaves moved to Mexico during that time.

b. Many Americans and their slaves moved to Mexico during that time, **because** Mexico allowed slavery.

Such studies show that comprehension can be improved by reformulating text; e.g., making causal relations explicit had a facilitatory effect for readers with low reading skills (Linderholm et al., 2000; Beck et al., 1991) and for readers with low levels of domain expertise (Noordman and Vonk, 1992). Further, specific information orderings were found to be facilitatory by Anderson and Davison (1988).

However, it has not been investigated whether readers with different levels of domain expertise are facilitated by any specific lexico-syntactic formulation among the many possible explicit realisations of a relation. This is a novel question in the linguistics literature, and we address it in section 3.

2.3 Connectives and automatic (re)generation

Much of the work regarding (re)generation of text based on discourse connectives aims to simplify text in certain ways, to make it more accessible to particular classes of readers. The PSET project (Carroll et al., 1998) considered simplifying news reports for aphasics. The PSET project focused mainly on lexical simplification (replacing difficult words with easier ones), but more recently, there has been work on syntactic simplification and, in particular, the way syntactic rewrites interact with discourse structure and text cohesion (Siddharthan, 2006). Elsewhere, there has been renewed interest in *paraphrasing*, including the replacement of words (especially verbs) with their dictionary definitions (Kaji et al., 2002) and the replacement of idiomatic or otherwise troublesome expressions with simpler ones. The current research emphasis is on automatically learning paraphrases from comparable or aligned corpora (Barzilay and Lee, 2003; Ibrahim et al., 2003). The text simplification and paraphrasing literature does not address paraphrasing that requires syntactic alterations such as those in example 1 or the question of appropriateness of different formulations of a discourse relation.

Some natural language generation systems incorporate results from psycholinguistic studies to make principled choices between alternative formulations. For example, SkillSum (Williams and Reiter, 2008) and ICONOCLAST (Power et al., 2003) are two contemporary generation systems that allow for specifying aspects of style such as choice of discourse marker, clause order, repetition and sentence and paragraph lengths in the form of constraints that can be optimised. However, to date, these systems do not consider syntactic reformulations of the type we are interested in. Our research is directly relevant to such generation systems as it can help such systems make decisions in a principled manner.

2.4 Corpus studies and treebanking

There are two major corpora that mark up discourse relations – the RST Discourse Treebank based on Rhetorical Structure Theory (Mann and Thompson, 1988), and the Penn Discourse Treebank (Webber et al., 2005). Neither is suitable for studies on the felicity of specific formulations of a discourse relation. As part of this research, we have created a corpus of 144 real text examples, reformulated in 8 ways, giving 1152 sentences in total.

There have been numerous corpus studies of discourse connectives, such as studies on the discourse-role disambiguation of individual cue-phrases in spoken and written corpora (e.g., Hirschberg and Litman (1993)), the substitutability of discourse connectives (e.g., Hutchinson (2005)), and indeed corpus studies as a means of informing the choice of discourse relations to consider in a theory (e.g., Knott and Dale (1994); Knott (1996)). A distinguishing feature of our approach relative to previous ones is an in-depth study of syntactic variations; in contrast, for example, Knott's taxonomy of discourse relations is based on the use of a substitution text that precludes variants of the same relation having different syntax.

3 Linguistic acceptability study

3.1 Dataset creation

We have constructed a dataset that can be used to gain insights into differences between different realisations of discourse relations. In the following, we will illustrate such rewriting situations using an example from a medical article. As mentioned previously, we are particularly interested in complex syntactic reformulations; in example 4 below, a. is from the original text and b.–h. are reformulations. There are two examples each of formulations using "*because*", "*because of*", the verb "*cause*" and the noun "*cause*" with different ordering of propositional content. This provides us with 8 formulations per example sentence; for example:

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

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

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

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

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

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

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

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

Our corpus contains 144 such examples from three genres (see below), giving 1152 sentences in total. These 144 examples contain equal numbers of original sentences (18) of each of the 8 types. The manual reformulation is formulaic, and it is part of our broader research effort to automate the process using transfer rules and a bi-directional grammar. The example above is indicative of the process. To make a clause out of a noun phrase (examples 4c.–f.), we introduce either the copula or the verb “occur”, based on a subjective judgement of whether this is an event or a continuous phenomenon. Conversely, to create a noun phrase from a clause, we use a possessive and a gerund; for example (simplified for illustration):

(5) a. Irwin had triumphed because he was so good a man.

b. The cause of Irwin’s having triumphed was his being so good a man.

Clearly, there are many different possibilities for this reformulation; for example:

(5) b’. The cause of Irwin’s *triumph* was his being so good a man.

b”. The cause of Irwin’s *triumph* was his *exceptional goodness as* a man.

As part of our wider research agenda, we are exploring automatic reformulation using transfer rules and a bi-directional grammar. In this context, given our immediate interest is in the discourse markers, we restrict our reformulation method to only generate sentences such as 5b. This not only makes automation easier, but also standardises data for our experiment by removing an aspect of subjectivity from the manual reformulation.

We used equal numbers of sentences from three different genres[1]:

- **PubMed Abstracts**: Technical writing from the Biomedical domain
- **BNC World**: Article from the British National Corpus tagged as World News
- **BNC Natural Science**: Article from the British National Corpus tagged as Natural Science. This covers popular science writing in the mainstream media

There were 48 example sentences chosen randomly from each genre, such that there were 6 examples of each of the 8 types of formulation)

3.2 Experimental setup

Human judgements for acceptability for each of the 1152 sentences in our corpus were obtained using the WebExp package (Keller et al., 2008 to appear).[2] We investigated acceptability because it is a measure which reflects both ease of comprehension and surface well-formedness.

The propositional content of 144 sentences was presented in 8 formulations. 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. Thus, while each participant read an equal number of sentences in each formulation type, they never read more than one formulation of the same propositional content. Each group saw 18 original and 126 reformulated sentences in total, 48 from each genre. This experimental design allows all statistical comparisons between the eight types of causal formulations to be within-participants.

Acceptability judgements were elicited on the sentences without presenting the preceding context

[1] PubMed URL: http://www.ncbi.nlm.nih.gov/pubmed/
The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk

[2] Note that the reformulations are, strictly speaking, grammatical according to the authors’ judgement. We are testing violations of acceptability, rather than grammaticality per se.

from the original text. The participants were University of Cambridge students (all native English speakers with different academic backgrounds). Post experimentally we divided participants in two groups based on having a Science or a non-Science background[3]. Rather than giving participants a fixed scale (e.g. 1–7), 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 Bard et al. (1996); Cowart (1997); Keller (2000)).

3.3 Magnitude estimation

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, using any positive number, even decimals, so that higher scores were assigned to more acceptable sentences. The advantage of Magnitude estimation is that the researcher does not make any assumptions about the number of linguistic distinctions allowed. Each subject makes as many distinctions as they feel comfortable. 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.

3.4 Results

42 out of 48 participants (19 science students and 23 non-science students) completed the experiment, giving us 3–6 ratings for each of the 1152 sentences. Figure 1 shows the average z-scores with standard error bars for Science and non-Science students for each of the three genres. The first six columns show the scores for only the 144 Original Sentences. Note that science students find PubMed sentences most acceptable (significantly more than BNC Natural Science; t-test, $p < .005$), while among non-science students there is a numerical tendency to find the world news sentences most acceptable. Both categories of participants disprefer sentences from the popular science genre. Columns 7–12 show the average z-scores for the 1008 reformulated sentences. Let us note that these are significantly lower than for the originals (t-test, $p < .001$).

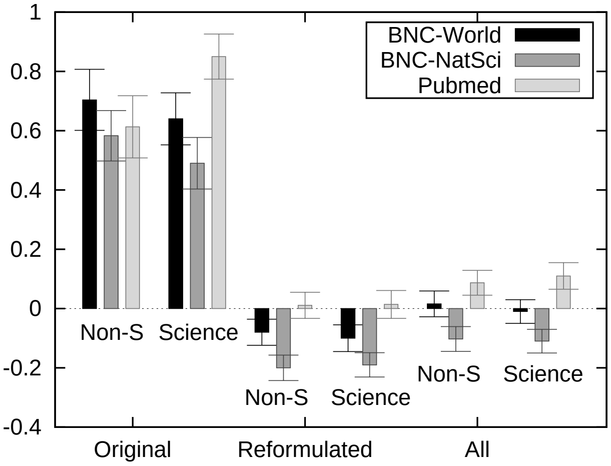

Figure 1: Preferences by Field of Study – **Science** or **Non-S**cience.

Some of these results are as expected. With regard to genre preferences, scientists might find the style of technical writing acceptable because of familiarity with that style of writing. Second, with regard to the average score for original and reformulated sentences, some reformulations just don't work for a given propositional content. This pulls the average for reformulated sentences down. However, on average 2 out of 7 reformulations score quite high.

It is interesting that the popular science genre is least preferred by both groups. This suggests that reformulating technical writing for lay readers is not a trivial endeavour, even for journalists.

Now consider Figure 2, which shows the average z-scores for only PubMed sentences for science and non-science students as a function of sentence type. For non-science students reading PubMed sentences, three formulations are strongly dispreferred – "a is caused by b", "because b, a" and "b is the

[3] Participants provided subject of study prior to participation in the experiment. Our classification of Science consists of Life Sciences(Genetics/Biology/etc), Chemistry, Environmental Science, Engineering, Geology, Physics, Medicine, Pharmacology, Veterinary Science and Zoology. Non-Science consists of Archaeology, Business, Classics, Education, Literature&Languages, International Relations, Linguistics, Maths, Music, Politics and Theology.

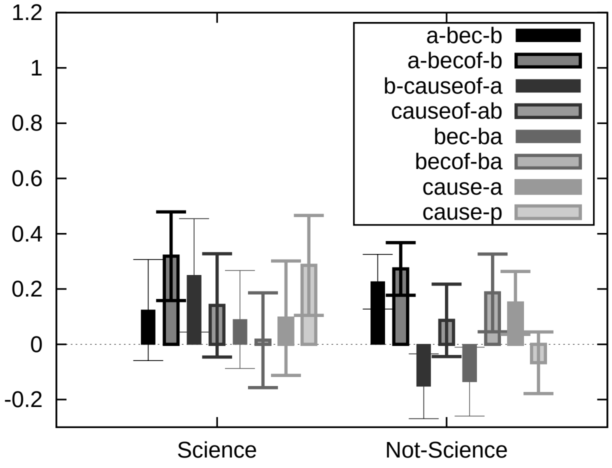

Figure 2: PubMed type preferences

cause of a". The last two are significantly lower than "a because b", "a because of b" and "because of b, a" (t-test, $.005 < p < .01$). On the contrary, there are no strong preferences among the science students and all the error bars overlap. Let us now look at some specific differences between science and non-science students:

1. Science students prefer sentences in the passive voice, while these are strongly dispreferred by non-science students. While active voice is the canonical form in English, much of science is written in the passive by convention. This difference can thus be explained by different levels of exposure.
2. Non-science students disprefer the use of "cause" as a noun while science students don't (columns 3–4 and 11–12).
3. Non-science students prefer "because of b, a" to "because b, a" while science students show the opposite preference.

The lack of strongly dispreferred formulations in the Science students is most likely due to two factors: (a) the group's familiarity with this genre and (b) their expert knowledge compensates for acceptability even for relatively odd formulations. In the absence of exposure and background knowledge, the non-Science students display clear preferences.[4]

Note that these preferences are not surprising. The preference for canonical constructs such as active voice and conjunction in infix position are well documented. Our claim however, is that blindly rewriting all instances of globally dispreferred constructs with globally preferred constructs is counterproductive because not all formulations are acceptable for any given propositional content. This claim is easily verified. Table 1 shows the average z-scores of non-science students when one formulation of each of the PubMed sentences is selected based only on the global preferences in Figure 2. Such rewriting invariably makes matters worse. In the next section we present a more intelligent approach.

Selection Method	Av. z
Always select original sentence	.61
Replace cause-p, b-causeof-a and causeof-ab with cause-a & bec-ba with a-bec-b	.48
Replace cause-p with cause-a, b-causeof-a with causeof-ab & bec-ba with a-bec-b	.47
Always select most preferred type (a-becof-b)	.27

Table 1: Selecting a formulation of PubMed sentences for non-science students using their global preferences.

4 Machine learning experiment

The first question we address is: for a given propositional content, which formulations are acceptable and which are not? This is a useful question for multiple reasons. In this paper, our interest stems from our desire to selectively rewrite causation based on the properties of the sentence as well as global preferences of categories of users. More generally, this information is important for summarisation tasks, where sentences might appear in different contexts and different information orderings might be desirable for reasons of coherence. Knowing which formulations are acceptable in isolation for a given propositional content is thus important.

Since Magnitude estimation scores are freescale, we first need to determine how high a score needs to be for that formulation to be considered acceptable. Our solution is to (a) treat the original formulation as acceptable and (b) treat any reformulations with a higher average z-score than the original as also acceptable. We find that roughly 3 formulations (the original and another two) out of 8 are acceptable on average. Our data is summarised below:

- 1152 Sentences in total (144 originals, 1008 reformulations)
    - 361 labelled as acceptable (31%; 144 originals, 217 reformulations)

[4] While we only show the averages for all sentences, the distributions for original and reformulated sentences look remarkably similar.

- 791 labelled as unacceptable (69%; 791 reformulations)

4.1 Features

We use shallow features derived from the sentence, as well as the textual genre. Sentences were parsed using the RASP parser (Briscoe and Carroll, 2002). The features we extract are as follows:

1. Type (8 values: cause_a, cause_p, a_bec_b, bec_ba, a_becof_b, becof_ba, a_causeof_b, causeof_ba)
2. Genre (3 values: pubmed, bnc-world, bnc-natsci)
3. Complexity: As an indication of the complexity of the propositional content, we use the following:
   (a) Length Features
      - length (in words) of the sentence and each clause
      - length (as proportion of total length) of each clause

   (b) Whether the causative is embedded in a relative clause or other construct
   (c) The presence or absence of copula in each clause (e.g., "because there is...")
   (d) Whether the causation is quantified (e.g., "a major cause of...")

The only feature that varies between the eight formulations of the same sentence is the "type" feature; the "genre" and "complexity" features are constant across reformulations. The reason for using 3(c–d) as features is that expressions such as "because there is" might be better formulated as "because of" and that it is hard to find an exact reformulation when quantifiers are present (e.g., "a major cause of" is not equivalent to "often because of").

Machine performance on this task is not very good (First Run, Table 2). The problem is that some propositional content is harder to formulate than others. Therefore good formulations of some propositional content might have much lower scores than even mediocre formulations of other propositional content. This makes it hard to learn a function that distinguishes good from bad formulations for any particular propositional content. To overcome this, we run the classifier twice. Given 8 formulations of 144 sentences $S_{i=1..144,j=1..8}$, the first run gives us 1152 probabilities $p_{weka1}(S_{ij})$ for the acceptability of each sentence, independent of propositional content (these are test-set probabilities using 10-fold cross-validation). We then run the machine learner again, with this new feature $relative$:

Classifier	Accuracy	Kappa
Baseline	.69	0
First Run	.72	.23
Second Run	.85	.65
Only PubMed	.89	.73

Table 2: Accuracy and Agreement of classifier relative to human judgement.

Genre	Class	P	R	F
All Genres	Good	.72	.78	.75
	Bad	.91	.89	.90
Only PubMed	Good	.89	.89	.89
	Bad	.89	.97	.92

Table 3: Precision, Recall and F-measure of classifier (second run) relative to human judgement.

- The ratio of the test-set probability (from the first run) to the highest of the 8 test-set probabilities for the different formulations of that sentence:

$$relative_{i=a,j=b} = \frac{p_{weka1}(S_{i=a,j=b})}{max_{i=a,j=1..8}(p_{weka1}(S_{i=a,j}))}$$

Thus probabilities for acceptability are normalised such that the best score for a given propositional content is 1 and the other 7 formulations score less than or equal to 1. The second classifier uses these relative probabilities as an extra feature.

4.2 Results

Our results are summarised in Table 2 (accuracy and agreement) and Table 3 (f-measure). We experimented with the Weka toolkit (Witten and Frank, 2000) and report results using "weka.classifiers.trees.J48 -C 0.3 -M 3" and 10-fold cross-validation for both runs.[5]

Table 2 shows that the first run performs at around baseline levels, but the second run performs significantly better (using z-test, p=0.01 on % Accuracy), with acceptable agreement of $\kappa = 0.65$[6]. This increases to 89% ($\kappa = .73$) when we only consider technical writing (PubMed genre). Table 3 shows that precision, recall and f-measure are also around .90 for PubMed sentences.

[5] J48 outperformed other Weka classifiers for this task.

[6] Following Carletta (1996), we measure agreement in κ, which follows the formula $K = \frac{P(A)-P(E)}{1-P(E)}$ where P(A) is observed, and P(E) expected agreement. κ ranges between -1 and 1. κ=0 means agreement is only as expected by chance. Generally, κ of 0.8 are considered stable, and κ of 0.69 as marginally stable, according to the strictest scheme applied in the field.

Left out feature	First Run		Second Run	
	Acc	κ	Acc	κ
Length	.71	-.01	.78	.33
Quantified	.71	.20	.75	.36
Embedded	.69	.15	.78	.37
Copula Present	.72	.20	.79	.44

Table 4: Accuracy and Kappa of classifier when complexity features are left out.

All our context features proved useful for the classification task, with the length features being the most useful. Table 4 shows the performance of the classifier when we leave out individual features.

It thus appears that we can determine the acceptable formulations of a sentence with high accuracy. The next question is how this information might be used to benefit a text regeneration system. To evaluate this, we combined our predictions with the user preferences visible in figure 2 as follows:

- We calculate a prior $prior_j$ for each formulation of type j using the z-score distribution for non-science students in Figure 2.
- We calculate $prior_{j=b}.p_{weka2}(S_{i=a,j=b})$ for each formulation $S_{i=a,j=b}$ of sentence a and type b, where $p_{weka2}(S_{i=a,j=b})$ is the probability returned by the classifier (second run) for formulation b of sentence a.
- **Selectively Reformulate:** We reformulate only the four dispreferred constructs (cause_p, bec_ba, causeof_ab, b_causeof_a) using the formulation for which the prior times the classifier probability is the maximum; i.e, for sentence a, we select $max_{i=a,j=1..8}(prior_j.p_{weka2}(S_{i=a,j}))$.

Table 5 shows the impact this reformulation has on the acceptability of the sentences. Our algorithm selects one formulation of each PubMed sentence based on our prior knowledge of the preferences of non-science students, and the Weka-probabilities for acceptability of each formulation of a sentence. Our selective reformulation increases the average z-score from .613 to .713. This is now comparable with the acceptability ratings of non-scientists for sentences from the world news genre. Note that reformulation only using priors resulted in worse results (Table 1).

However there remains scope for improvement. If we had an oracle that selected the best formulation of each sentence (as scored by non-scientists), this would result in an average score of 1.04.

Genre	Version	z-score
PubMed	Randomly Selected	–.17
PubMed	Original Sentences	.61
PubMed	Selectively Reformulate	.71
PubMed	Selected by Oracle	1.04
BNC World	Original Sentences	.70

Table 5: Average z-scores for non-science students. Selective reformulation increases the acceptability scores of sentences drawn from technical writing to levels comparable to acceptability scores of sentences drawn from news reports on world news (their most preferred genre).

5 Conclusions and future work

In this investigation we report that science and non-science university students have different global preferences regarding which formulations of causation are acceptable. Using surface features that reflect propositional complexity, a machine classifier can learn which of 8 formulations of a discourse relation are acceptable (with Accuracy $= .89$ and Kappa $= .73$ for sentences from the PubMed genre). Using the global preferences of non-science students as priors, and combining these with machine classifier predictions of acceptability, we have demonstrated that it is possible to selectively rewrite sentences from PubMed in a manner that is personalised for non-science students. This boosts the average z-score for acceptability from .613 to .713 on PubMed sentences, a level similar to scores of non-scientists for sentences from their most preferred World News genre. We have thus shown that there is potential for reformulating technical writing for a lay audience – differences in preferences for expressing a discourse relation do exist between lay and expert audiences, and these can be learnt.

While in this paper we focus on the discourse relation of causation, other discourse relations commonly used in scientific writing can also be realised using markers with different syntactic properties; for instance, *contrast* can be expressed using markers such as “while”, “unlike”, “but”, “compared to”, “in contrast to” or “the difference between”. As part of our wider goals, we are in the process of extending the number of discourse relations considered. We are also in the process of developing a framework within which we can use transfer rules and a bi-directional grammar to automate such complex syntactic reformulation.

Acknowledgements

This work was supported by the Economic and Social Research Council (Grant Number RES-000-22-3272). We would also like to thank Donia Scott, Simone Teufel and Ann Copestake for many discussions that influenced the scope of this work, and John Williams and Theodora Alexopoulou for their suggestions on experimental design.

References

R.C. Anderson and A. Davison. 1988. Conceptual and empirical bases of readibility formulas. In Alice Davison and G. M. Green, editors, *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*. Lawrence Erlbaum Associates, Hillsdale, NJ.

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

R. Barzilay and L. Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *HLT-NAACL 2003*, pp 16–23.

I.L. Beck, M.G. McKeown, G.M. Sinatra, and J.A. Loxterman. 1991. Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly*, pp 251–276.

E.J. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proc. of the 3rd International Conference on Language Resources and Evaluation*, pp 1499–1504, Gran Canaria.

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proc. of AAAI98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pp 7–10, Madison, WI.

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

J. Hirschberg and D. Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

B. Hutchinson. 2005. Modelling the substitutability of discourse connectives. In *ACL '05: Proc. of the 43rd Annual Meeting on Association for Computational Linguistics*, pp 149–156, Morristown, NJ, USA. Association for Computational Linguistics.

A. Ibrahim, B. Katz, and J. Lin. 2003. Extracting paraphrases from aligned corpora. In *Proc. of The Second International Workshop on Paraphrasing*.

N. Kaji, D. Kawahara, S. Kurohash, and S. Sato. 2002. Verb paraphrase based on case frame alignment. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pp 215–222, Philadelphia, USA.

F. Keller, S. Gunasekharan, N. Mayo, and M. Corley. 2008, to appear. Timing accuracy of web experiments: A case study using the webexp software package. *Behavior Research Methods*.

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

A. Knott and R. Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

A. Knott. 1996. *A Data-Driven Methodology for Motivating a Set of Discourse Relations*. Ph.D. thesis, Ph. D. thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK.

T. Linderholm, M.G. Everson, P. van den Broek, M. Mischinski, A. Crittenden, and J. Samuels. 2000. Effects of Causal Text Revisions on More-and Less-Skilled Readers' Comprehension of Easy and Difficult Texts. *Cognition and Instruction*, 18(4):525–556.

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

L. G. M. Noordman and W. Vonk. 1992. Reader's knowledge and the control of inferences in reading. *Language and Cognitive Processes*, 7:373–391.

R. Power, D. Scott, and N. Bouayad-Agha. 2003. Generating texts with style. *Proc. of the 4 thInternational Conference on Intelligent Texts Processing and Computational Linguistics*.

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

B. Webber, A. Joshi, E. Miltsakaki, R. Prasad, N. Dinesh, A. Lee, and K. Forbes. 2005. A Short Introduction to the Penn Discourse TreeBank. *Treebanking for discourse and speech: proceedings of the NODALIDA 2005 special session on Treebanks for spoken language and discourse*.

S. Williams and E. Reiter. 2008. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, 14(04):495–525.

I. Witten and E. Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

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*, pp 29–48.