CMCPsは英語のノンネイティブスピーカーに及ぼす影響

Xinyu Deng　　中村順一

京都大学

概要

CMCPs (Complex Multiple Cue Phrases) とは、三つ以上の命題を連結した二つ以上の接続詞のことである。本研究では、三つの命題を二つの接続詞で連結したものを対象とした。今回は、CMCPsは英語のノンネイティブスピーカーに及ぼす影響について、二つの経験的な実験をした。一つの実験では、コーパスを利用して、CMCPsに対して、量的な分析を行った。もう一つの実験では、CMCPsが入っている文章のわかりやすさや、首尾一貫性に関して、アンケート調査をした。その後、調査の結果を詳しく分析した。二つの実験の結果から、CMCPsはノンネイティブスピーカーに対して、積極的な作用があることを分かった。研究の成果は文章生成なの専門分野に応用することができる。

The effect of complex multiple cue phrases on non-native speakers of English

Xinyu Deng　　Jun-ichi Nakamura

Kyoto University

Abstract

In this study, complex multiple cue phrases (CMCPs) refer to two cue phrases which signal two discourse relations in an embedded structure. We did two empirical experiments to investigate the effect of CMCPs on non-native users. One experiment was corpus analysis, in which we explored the effect of CMCPs quantitatively. Another was a questionnaire; we analysed the comprehensibility and coherence of texts containing CMCPs. The experiment results showed that CMCPs can help non-native users understand texts. The heuristics obtained from the study can be applied to the natural language generation systems for non-native users.

1 Introduction

Natural Language Generation (NLG) is a sub-field of artificial intelligence that is concerned with building computer software systems that can produce meaningful texts from some underlying nonlinguistic representation of information (Reiter and Dale, 2000). Until now, in the field of NLG, little research has been done for non-native speakers. In this study, we investigate the effect of complex multiple cue phrases (CMCPs) on non-native users. Generally, non-native speakers are divided into three levels: primary (middle school student level), intermediate (high school student level) and advanced (university student level). The users of this study are assumed to be at intermediate level.

From the viewpoint of NLG, the following three texts have the same abstract structure.

Example 1.1

1. You failed the exam. ***But if*** you study hard, you can master English.
2. You failed the exam. ***But*** you can master English ***if*** you study hard.
3. ***Although*** you failed the exam, you can master English ***if*** you study hard.

In each text of Example 1.1, two cue phrases are used to signal two discourse relations. This kind of cue phrases are called *complex multiple cue phrases* (CMCPs) (Oates, 2001). In this study, an embedded structure in which CMCPs occur is defined to have two cue phrases and three propositions. CMCPs have two classes. *Class 1* represents the embedded structure in which the first cue phrase immediately precedes the second one and both cue phrases are attached to the second proposition, e.g., text 1 of Example 1.1. *Class 2* of CMCPs has two sub-classes: *Class 2-1* and *Class 2-2.* In *Class 2-1*, cue phrases precede the second and the third proposition respectively, e.g., text 2. In *Class 2-2*, cue phrases precede the first and the third proposition, e.g., text 3.

(Williams, 2003) pointed out that for poor readers of native speakers, it is better not use *Class 1* because it does not improve coherence. In this study, we explore the effect of *Class 1* and *Class 2-1* on non-native users by two empirical experiments. We are developing a text generation system which is called SILK (Generation *S*ystem for *I*ntermediate *L*evel non-native spea*K*ers on discourse level). We will apply the heuristics obtained from this study to the SILK system. The rest of the paper is arranged as follows. Section 2 describes related work. Section 3 and 4 introduce corpus analysis and questionnaire respectively. In Section 5, we draw a conclusion.

2 Related work

In the field of NLG, almost all researches on cue phrases have been done on single cue phrase and only for the users of native speakers. Although the problem on CMCPs was mentioned by several researchers (Knott, 1996), (Fraser, 1990), (Delin and Scott and Hartley, 1996), it has not been studied in detail. Until now, the studies on multiple cue phrases are much fewer than others. Though (Oates, 2001) is one of the first ones, it focuses on simple multiple cue phrases and just considers the case of native speakers. To my knowledge, our study is the first one on the usage of CMCPs on non-native users.

3 Empirical experiment 1: corpus analysis

In this section, we explore the effect of CMCPs on non-native users by comparing two corpora.

3.1 Two corpora

In order to investigate the effect of CMCPs, we created two corpora: SUB-BNC (for native speakers) is a sub-corpus of BNC (British National Corpus); CNNSE (Corpus for Intermediate

Level Non-Native Speakers of English) was created by the first author. The two corpora have the same size (200,000 words each). According to the Flesch Reading Ease scale, the readability of SUB-BNC is 47.5 (difficult), the readability of CNNSE is 68.7 (easy). We used the following method to make the two corpora comparable.

- The domain is *natural and pure science.*
- The medium is *book.* For CNNSE, all of the texts are extracted from the books published in China and in Japan. The texts are written or rewritten by native speakers.

In addition, the target audience of SUB-BNC is *adult*, while the target audience of the books used to create CNNSE is *high school student.*

3.2 Investigating CMCPs within two corpora

(Knott, 1996) divided the taxonomy of cue phrases into ten categories. We chose the following five categories which are related to our study: "cause" relation, "result" relation, "temporal" relation, "negative polarity" relation, and "hypothetical" relation. The five categories contain 114 cue phrases, excluding repetition of the same one. For each of the cue phrase, we examined whether they occur either before or after every other cue phrase (Oates, 2001) p.106, i.e., we tested 114 × 114 combinations. First, we used program to extract the text fragments containing either of the two kinds of pattern mentioned above. Then we checked manually whether CMCPs occur in those text fragments. We recorded the cue phrase combination and the number of occurrences if *Class 1* or *Class 2-1* of CMCPs was found.

Table 1 shows the results of the corpus analysis. Within SUB-BNC, the frequency of *Class 2-1* is 116 which is much more than that of *Class 1* (61). This proves the opinion that it is better not use *Class 1* for native speakers (Williams, 2003). However, within CNNSE, the frequency of *Class 1* (98) is higher than that of *Class 2-1* (83). It means that for non-native speakers, *Class 1* is preferred. From the viewpoint of psycholinguistics, it can be explained that embedded structures are not easy for non-native speakers to understand. It is necessary to put two cue phrases together to attract more attention of them. Table 1 shows that there is significant difference in CMCPs usage between native and non-native speakers. Chi Square critical values also support this conclusion.

Table 1: Results of corpus analysis (χ^2 = 14.02, p < 0.005)

	Frequency of Class 1	Frequency of Class 2-1
SUB-BNC	61	116
CNNSE	98	83

4 Empirical experiment 2: questionnaire

In this section, we introduce a questionnaire, by which we investigated the comprehensibility and coherence of the texts containing CMCPs.

4.1 Design

The questionnaire has two goals: finding out the relationship between comprehensibility and CMCPs (*Class 1* or *Class 2-1*); exploring whether the embedded structure containing CMCPs (*Class 1* or *Class 2-1*) affects the coherence from the viewpoint of the non-native users. We therefore made two hypotheses (Table 2), and then we tested these hypotheses through experiment. If they are accepted, the hypotheses can be applied in the NLG systems.

Table 2: Hypotheses of the questionnaire

Hypothesis 1	From the viewpoint of *comprehensibility*, texts containing *Class 2-1* are preferred, i.e., the effect of CMCPs on non-native speakers is the same as that on native speakers.
Hypothesis 2	From the viewpoint of *coherence*, texts containing *Class 2-1* are more coherent, i.e, like native speakers, non-native speakers think that *Class 1* can not improve coherence.

The experiment is between-groups design (Hatch and Lazaraton, 1991). In the experiment, there is one independent variables: *Class*, which has two levels, i.e., *Class 1* and *Class 2-1*. Furthermore, we asked human subjects to score the following two dependent variables:

- *Comprehensibility*: the degree of understanding. It has 6 values (see Table 3).
- *Coherence*: how the ideas, thoughts are consistent or connected logically. It has five values (see Table 3).

Table 3: Values of *comprehensibility* and *coherence*

Values	Dependent Variables	
	Comprehensibility	*Coherence*
6	very easy to understand	
5	easy to understand	coherent
4	can understand	fairly coherent
3	difficult to understand	so-so
2	very difficult to understand	fairly uncohernet
1	can not understand	uncoherent

The questionnaire had 20 texts (e.g., Example 4.1.1) which were obtained from high school students' textbooks published in China and in Japan. 8 texts of them contained *Class 1*, 12 texts contained *Class 2-1*. The subjects were asked to select the value of *comprehensibility* and *coherence* in 30 minutes. We had 23 intermediate level non-native subjects. The questionnaires were given to the subjects to be completed in their spare time.

Example 4.1.1

Generally speaking, everything we eat does some good to our body, <u>but if we eat too much of one kind of food and neglect others, we may have too much of one kind of chemical substance and not enough of others.</u>

Comprehensibility: 6 5 4 3 2 1 *Coherence*: 5 4 3 2 1

4.2 Data analysis

In this section, we discuss the experiment results and the conclusions that can be drawn from them.

Comprehensibility

We chose the Mann Whitney U test to analyse the difference in *comprehensibility* between the texts containing *Class 1* and *Class 2-1*. The results (Table 4) showed that there was significant difference between the two kinds of CMCPs. So we rejected Hypothesis 1. Moreover, the average *comprehensibility* of texts containing *Class 1* and *Class 2-1* is 5.03 and 4.49 respectively, i.e., the texts containing *Class 1* are easier to understand. Figure 1 shows the distribution of comprehensibility assessment. The majority of *Class 1* are "easy to understand" (52.7%) and "very easy to understand" (25.5%). While for *Class 2-1*, the majority are "can understand" (45.3%) and "easy to understand" (43.1%).

Table 4: The output of the Mann Whitney U Test for *comprehensibility*

Inde-Var	Cases	Z value	2-tail Sig.
Class	460	-7.748	.000

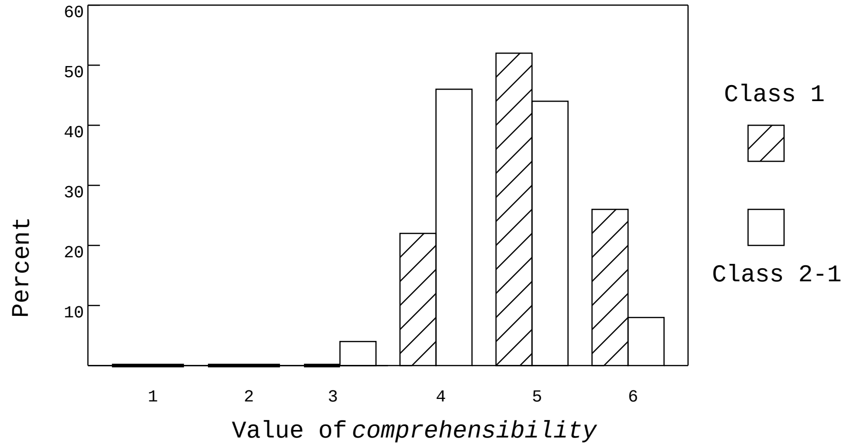

Figure 1: The *comprehensibility* of texts containing CMCPs

Coherence

We chose the Mann Whitney U test to analyse the difference in *coherence* between texts containing *Class 1* and *Class 2-1*. The results (see Table 5) showed that there was no significant difference between the two kinds of CMCPs. So we rejected Hypothesis 2. This means that compared with native speakers, non-native speakers are not sensitive to the *coherence* of texts. Figure 2 shows that the difference in distribution of coherence assessment between *Class 1* (17.4%, 47.3%, 35.3%) and *Class 2-1* (14.1%, 43.5%, 42.4%) is not significant.

Table 5: The output of the Mann Whitney U Test for *coherence*

Inde-Var	Cases	Z value	2-tail Sig.
Class	460	-.889	.374

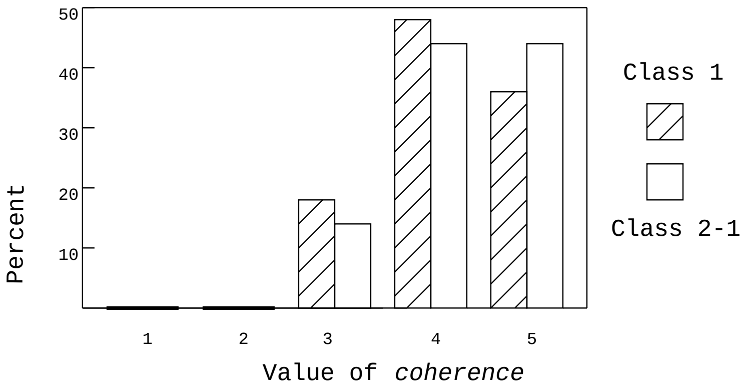

Figure 2: The *coherence* of texts containing CMCPs

5 Conclusion

This study shows that there is difference in CMCPs usage between native and non-native speakers. Compared with *Class 2-1*, *Class 1* is preferred by non-native users. We found that non-native speakers are not sensitive to the *coherence* of texts. Instead, they pay more attention to the position of the cue phrases. This may be the reason why texts containing *Class 1* of CMCPs is easier to understand. We think the heuristics obtained from this study can be applied not only to the SILK system but also those NLG systems for non-native users.

References

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems.* Cambridge University Press.

Sandra Williams. 2003. *Language choice models for microplanning and readability.* Proceedings of the Student Workshop of the Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference (HLT-NAACL Student Workshop), Edmonto.

Evelyn Hatch and Anne Lazaraton. 1991. *The Research Manual: Design and Statistics for Applied Linguistics.* Newbury House Publishers.

Judy Delin and Donia Scott and Anthony Hartley. 1996. *Language Specific Mappings from Semantics to Syntax.* Proceedings of the 16th COLING. pp. 96-101, Helsinki.

Bruce Fraser. 1990. *An Approach to Discourse Markers.* Journal of Pragmatics, 14:383-395.

Alistair Knott. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations.* Ph.D. Thesis, University of Edinburgh.

Sarah Oates. 2001. *Generating Multiple Discourse Markers in Text.* MPhil Thesis, Information Technology Research Institute, University of Brighton.

Gisela Redeker. 1990. *Ideational and Pragmatic Markers of Discourse Structure.* Journal of Pragmatics 14:367-381.