6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

REPAIR PATTERNS IN SPONTANEOUS CHINESE DIALOGS: MORPHEMES, WORDS, AND PHRASES

Shu-Chuan Tseng
Institute of Linguistics, Academia Sinica, Taiwan
E-mail: tsengsc@gate.sinica.edu.tw

ABSTRACT

Applying experimental and empirical results of linguistic analyses on Chinese speech repairs, this paper presents a new line of research on the relationship of constituent boundaries and speech repairs in spoken Mandarin Chinese. A judgement experiment was carried out to look for particularity of Chinese repairs, so that preliminary results can be obtained. Based on the results, regular repair patterns were determined in relation to three constituent levels: morpheme, word, and phrase. Furthermore, a detailed corpus analysis on Chinese spontaneous dialogs shows that these three constituent units in spoken Chinese influence the production of speech repairs in different ways. The interrelationship of the constituent boundaries and the repair patterns is empirically illustrated. This leads to the conclusion that research on speech repairs not only helps language understanding systems in the way that they can cope with spontaneous speech phenomena. This kind of research also provides empirical cues to the construction of linguistic theories.

1. INTRODUCTION

In the context of the Chinese linguistics, the interaction of morphemes, words, and phrases is a fascinating issue [2], [10]. This is because the orthographic and the semantic systems in Chinese are closely related to each other. And each morpheme is potentially a word, or even a phrase. However, how to acquire an appropriate methodology to empirically explore and describe this relationship has proved to be greatly difficult. This paper introduces a new possibility to do so, that is, to investigate this issue by determining repair patterns in spoken Chinese dialogs.

Recently, studies on speech repairs have provided practical and theoretical cues in various research fields: the natural language processing (NLP) [1], [6], [12] as well as psycholinguistic [9], [11] and computational linguistic fields [7], [14]. In addition, cross-linguistic comparative studies on speech repairs have also been done recently [5], [13]. Sequential template-based descriptions of speech repairs are often adopted, for instance the following notation: 1) the reparandum, 2) the editing term, and 3) the alteration. The reparandum represents the words, which are to be corrected or repeated, whereas the alteration is the actual correction or repetition. The editing term occurs between the reparandum and the alteration, usually in the form of discourse particles or pauses.

In addition to the template-based description of speech repairs, further analyses are necessary to establish a theory of speech repairs with practical and applicable linguistic contributions. Thus, this paper intends to determine the linguistic features of speech repairs with regard to the morphological, semantic, and syntactic perspectives by investigating authentic spontaneous speech data. For this purpose, spontaneous corpus data have been collected, annotated, and analyzed.

2. DETERMINING REPAIR PATTERNS

Before the corpus analysis is started, a pre-experiment has been designed to obtain possible Chinese speech repair patterns.

2.1 A Pre-Test

32 hypothetical Chinese repairs of the following structures were manipulated by the author, 1) determiner adjective noun and 2) adjective noun. The placement of repair sequences varies by correcting or repeating all potential syntactic and semantic segments in turn. 20 Chinese native speakers, aged from 26 to 37, were shown these 32 potential Chinese repairs and asked to give their judgment as to which repairs are "appropriate" according to their linguistic competence.

2.2 Preliminary Results

The results have shown that the phrasal beginning is the most likely position for native speakers to repair or repeat their speech stretch. Besides, morphemes containing more semantic content are also more likely to be corrected or repeated than morphemes such as particles or classifiers. To be more specific, among the 16 sequences, which were identified as appropriate repair sequences by more than one third of the overall subjects, only two repairs were operated within phrases. In these two sequences, the morphemes where the repair process is executed carry the main semantic content of the words involved. The following examples should clarify this point: "這個(this) 藍(blue) 藍色(blue-color) 的(structural particle) 螺絲(screw)"[1] was identified as a possible repair by 13 subjects, while the sequence "這個(this) 藍色(blue-color) 色(color) 的

[1] The Pinyin transcription: Zhe ge lan lan se de luo si.

(structural particle) 螺絲(screw)"[2] was only recognized by two subjects as a possible repair pattern. In these two patterns, "藍" (blue) plays the major semantic role in the adjective "blue" and "色", which means "color", is merely part of a morphological construction. To hesitate at the morpheme "藍" seems to be more appropriate and effective than to hesitate and repair at the morpheme "色". In fact, the latter one is a morphologically relevant hesitation. The other 14 repair sequences are initiated at the phrasal beginning position, i.e. either determiners or adjectives are repeated or repaired. For instance, the sequence "那個(that) 那個(that) 黑(black) 色(color) 的(structural particle) 螺絲(screw)"[3] was accepted by more than the half overall subjects.

These results partially confirm the conjecture proposed by Chui [3] that the completion of Chinese repairs is more directly related to the lexical complexity of the problem words than their syntactic characteristics. With regard to the linguistic strata involved in Chinese repairs, the boundaries (morphological, word, and phrasal) play an essential role. Our preliminary results on Chinese repair patterns can be summarized as follows:

> The most likely position for the Chinese speech repairs to be initiated is the phrasal boundary, where the second most likely position is the morpheme carrying the semantic content of the problem word involved.

3. REPAIRS IN CHINESE SPOKEN DIALOGS

Monologs usually contain less frequent speech repairs, because there is no communicative necessity to repair inappropriate speech. In order to investigate repair patterns in Chinese spoken dialogs, spontaneous speech data were analyzed in the corpus analysis.

3.1 Spontaneous Speech Data

Taiwan Putonghua Corpus (TWPTH), where Putonghua refers to Mandarin Chinese, was recorded in Taiwan. The speakers were all born in Taiwan and their first language is Taiwanese (Southern Min). The speakers were given the instructions in advance to speak in usual conversation style and they could speak on any topic they wanted to, or even on no topic at all. Thus, the spontaneous and conversation-oriented speech data were obtained. A total of 40 speakers were recorded including five dialogs and 30 monologs. Three dialogs were analyzed for the study in this paper and each is about 20 minutes long. In total, 373 immediate speech repetitions and repairs were identified in these three dialogs and they were annotated according to the POS system developed for the Sinica Corpus [4].

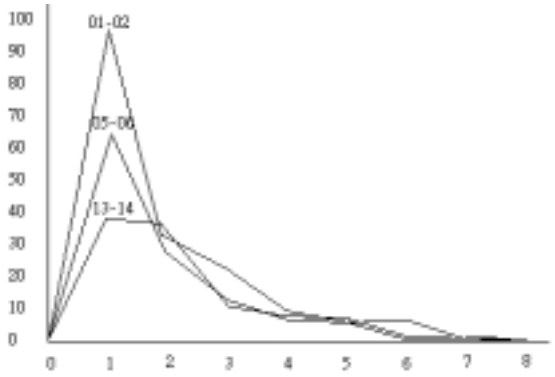
[2] The Pinyin trancription: Zhe ge lan se se de luo si.

[3] The Pinyin trancription: Na ge na ge hei se de luo si.

3.2 Retracing in Repairs: A Quantitative Analysis

In the framework of speech repairs, one of the most essential issues is the retracing phenomenon. Three dialogs out of the TWPTH corpus have been investigated; they are annotated as 01-02, 05-06, and 13-14, respectively. Figure 1 illustrates graphically how similar the patterns across these three dialogs are, especially in respect of the number of retraced morphemes. This fact supports the notion that these three pairs of different subjects have identifiable parallel repairing behavior.

Figure1: Repair Pattern w.r.t. Morphemes Retraced



In order to get a clearer picture of retracing behavior in Chinese speech repairs, we look at Table 1. Clearly, we see that the most frequent cases of retracing in Chinese speech repairs are one-morpheme-retracing. Speakers tend to repair their speech directly after they made the inappropriate production. 51.7% of speech repairs found in this study are related to one retraced morpheme. This is independent of whether the retraced morphemes are word-initial or phrase-initial. The quantitative distance between the reparandum and the alteration counts essentially. 24.4% of speech repairs were produced by retracing two morphemes to the reparandum, and 11% three morphemes. This means, 87% of speech repairs found in the corpus were retraced at most three morphemes to the reparandum beginning, in other words, three Chinese characters, irrespective of the syntactic and semantic position of the morphemes.

In Chinese, each morpheme is written in the form of a character and each character has its own original meaning. However, there are more and more compound words in modern Chinese than in the classical Chinese, where the majority of words are usually monosyllabic. Our results on the retracing phenomenon illustrate that Chinese speakers prefer repairing their speech back to the nearest meaning-carrying morphemes to repairing or repeating their errors in complete words or phrases. This conclusion empirically supports the n-gram method in statistical natural language processing approaches for Chinese speech repairs, for instance in the repair sequence "那(then) 一

(one) 一般(usually) 的(structural particle) 話(cases), 你(you) 在(in) 中心(center) 訓練(training) 的(structural particle) 時候(time)"[4], only one morpheme is directly involved and the window looking for problem sequences in speech needs only to search one morpheme backwards.

Table 1: Retraced Morphemes in Total

# of Retraced Morphemes	Occurrence	Percentage
0	2	0.5%
1	193	51.7%
2	91	24.4%
3	41	11.0%
4	19	5.1%
5	13	3.5%
6	9	2.4%
7	3	0.8%
8	2	0.5%
Total	373	100%

3.3 Constituent Boundary vs. Speech Repairs: A Qualitative Analysis

The initiation of the reparandum and the alteration in the majority of the Chinese repairs investigated in this study were located at the morphological, word, and phrasal boundaries at the same time. It is significantly less frequent that repairs were started at morphological boundaries where no word and phrasal boundaries are involved. In Table 2, a clear picture of Chinese repair patterns can be drawn. 85.3% of speech repairs were started at the morphological, word, and phrasal levels. 8.3% of speech repairs were initiated at the morphological and word levels, but not directly related to the syntactic level. For instance, in the sequence "從(from) 早(early) 早上(morning) 就(immediately) 帶(bring) 帶(bring) 他(him) 去(go)"[5] the prepositional phrase "from morning" was not repaired at the phrasal beginning, but at the word-initial position.

It is possible that a Chinese speech stretch has constituent boundaries, which are only word boundaries or both word and phrase boundaries. However, it is impossible that phrasal boundaries are not at the word boundaries. This fact is empirically observable. In Table 2, no speech repairs were retraced to phrasal boundaries, which are not word boundaries. If speech repairs are found within words, neither word boundaries nor phrase boundaries are relevant. These cases make up about 6.4% of the total speech repairs investigated in this study.

[4] The Pinyin trancription: Na yi yi ban de hua, ni zai zhong xin xun lian de shi hou.

[5] The Pinyin trancription: Cong zao zao shang jiu dai dai ta qu.

Table 2: Repair Patterns w.r.t Constituent Boundaries

Morpheme	Word	Phrase	Occurrence
+	+	+	318
+	+	-	31
+	-	+	0
+	-	-	24
Total			373

4. REPAIR PATTERNS

Based on the previous results, a descriptive system of Chinese speech repair structure is developed. As a matter of fact, the template-based method can be applied to Chinese speech repairs and we use the notation [**Rep**(arandum), **Ed**(iting Term), **Alt**(eration)]. Within this notation, the internal relationship of Chinese speech repairs is specified.

4.1 Infrequent Editing Terms

Only 27 speech repairs contain an editing term. In total, there are eight editing terms found in the 373 speech repairs. These are 哎(9 occurrences), 啊(5), 呵(5), 呃(3), 嗷(2), 呢(1), 嘿(1), 啦(1). It makes up only 7.2% of the overall speech repairs. This result illustrates the differences between the micro-discourse structure in English and in Chinese with respect to the appearance of edit signals in repairs. The approach Hindle [7] proposed in his deterministic parser for spontaneous speech is based on the edit signal hypothesis (speech repairs are often marked by edit signals in the form of particles, pauses or lexically marked sequences). Our result clearly shows the infrequency of editing terms in Chinese repairs, i.e. specific approaches to detecting and correcting Chinese repairs are necessary.

4.2 Word-Based or Phrase-Based?

There are two reasons for the choice of morpheme as the basic component in the structural description of Chinese repairs. First, each morpheme in Chinese is at the same time a potential word and all morphemes have somehow their semantic function. Secondly, in the speech repairs we investigated, significant regularity was found in relation to the morphemic retracing within repairs, i.e. the morphemic distance between **Rep** and **Alt** is at most three in more than 87% of the overall repairs. Therefore, three morphemic repair sequence patterns are determined:

A ***M_rep1 M_alt1*** **B**,

A ***M_rep1 M_rep2 M_alt1 M_alt2*** **B**, and

A ***M_rep1 M_rep2 M_rep3 M_alt1 M_alt2 M_alt3*** **B**,

where **A** and **B** are the rest of the utterance containing repair sequences ***M_X***. ***M_X***s are morphemes and ***X***={*rep, alt, rep1, rep2, rep3, alt1, alt2, alt3*}. The relationship of **A**, **B**, and ***M_X*** could not be determined by our corpus analysis. However, the internal structure within ***M_X*** can

be specified as follows. ***M_rep1*** and the corresponding ***M_alt1*** are usually at the phrasal-initial position. However, the repair sequences ***M_rep***s ***M_alt***s do not necessarily have the same syntactic category, because ***M_rep***s are in many cases fragmentary words, so are ***M_alt***s. Therefore, it is not very useful to compare the syntactic categories of these two component parts of speech repairs. Besides, ***M_rep***s and ***M_alt***s often contain more than one phrase. In the case of Chinese repairs, this clearly illustrates the deficiency of approaches suggested for the Indo-European languages such as English [7], which make use of the mapping between corresponding phrases in the reparandum and in the repair. Although the beginning of the reparandum and the alteration is usually located at the phrase-initial position, word boundary is an even more likely position to detect Chinese repairs. However, according to the results of the corpus analysis, no significant evidence could be obtained to give a clear answer to the question: is the production of speech repairs in Chinese word-based or phrase-based?

5. CONCLUSIONS

This paper has shown that recent research tools and methodology in the computational linguistic field can be applied to obtain new perspective on traditional linguistic issues. In spite of the fact that speech repairs have long history in the psycholinguistic research context, this kind of research has not been applied to other linguistic frameworks yet. From this viewpoint, this paper has shown that the production of speech repairs reflects the interaction of morphemes, words, and phrases in Chinese. These three linguistic levels are all related to the semantic and orthographic characteristics because the Chinese characters function as morphemes and words. The corpus analysis of Chinese speech repairs has determined the frequent repair patterns by examining the morphemic retracing phenomenon and the structural relations within repairs. Furthermore, the occurrence of speech repairs at different constituent boundaries was investigated in detail, so that probabilities can be derived from the empirical results.

Another particularity of Mandarin Chinese is that it has lexical tones. The interaction between syntax and prosody in Mandarin is a linguistically and computationally interesting issue. How syntactic and prosodic functions co-operate and compensate in discourse is worth investigating. In the case of Chinese, approaches to the use of acoustic cues to detect and correct speech repairs have been suggested [8], however, specific investigations into the role the lexical tones play in the production of Chinese speech repairs have not been initiated yet.

6. ACKNOWLEDGEMENTS

Results presented in this paper are part of the project supported by the National Science Council in Taiwan, NSC 89-2411-H-212-003. And I'd like to sincerely thank the colleagues in the Industrial Research Technology Institute (IRTI) who kindly supported me with the TWPTH corpus speech data.

7. REFERENCES

1. Bear, J., Dowding, J., and Shriberg, E. ``Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog", *ACL*, 56-63, 1992.

2. Chao, Y.-R. "A Grammar of Spoken Chinese", Berkeley: University of California Press, 1968.

3. Chui, K.-W. "Organization of Repair in Chinese Conversation. *Text*, 16/3:343-372, 1996.

4. CKIP, "Sinica Balanced Corpus" Technical Report no. 95-02/98-04. (in Chinese)

5. Eklund, R., and Shriberg, E. "Cross-Linguistic Disfluency Modeling: A Comparative Analysis of Swedish and American English Human-Human and Human-Machine Dialogs", *ICSLP 98*, 2631-2634, 1998.

6. Heeman, P., and Allen, J. ``Speech Repairs, Intonational Phrases and Discourse Markers: Modelling Speakers' Utterances in Spoken Dialogue", *Computational Linguistics*, 25/4:527-571, 1999.

7. Hindle, D. ``Deterministic Parsing of Syntactic Non-fluencies", *ACL*, 123-128, 1983.

8. Lee, Y.-S., and Chen, H.-H. "Using Acoustic and Prosodic Cues to Correct Chinese Speech Repairs", *EUROSPEECH 97*, 2211-2214, Rhodes, Greece, 1997.

9. Levelt, W. ``Monitoring and Self-Repair in Speech", *Cognition*, 14: 41-104, 1983.

10. Li, C., and Thompson, S. "Mandarin Chinese: A Functional Reference Grammar", Berkeley: University of California Press, 1981.

11. Lickley, R.J., and Bard, E.G. ``When Can Listeners Detect Disfluency in Spontaneous Speech", *Language and Speech*, 41/2:203-226, 1998.

12. Nakatani, C., and Hirschberg, J. ``A Corpus-Based Study of Repair Cues in Spontaneous Speech", *Journal of the Acoustical Society of America*, 95: 1603-1616, 1994.

13. Tseng, S.-C. "Modelling Speech Repairs in German and Mandarin Chinese Spoken Dialogues", *COLING 2000*, 2000. (to appear)

14. Tseng, S.-C. "Grammar, Prosody and Speech Disfluencies in Spoken Dialogues", PhD Thesis, University of Bielefeld, 1999.