Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop, 5–8 September 2003, Göteborg University, Sweden.
Robert Eklund (ed.), Gothenburg Papers in Theoretical Linguistics 90, ISSN 0349–1021, pp. 73–76.

Repairs and repetitions in spontaneous Mandarin

Shu-Chuan Tseng

Institute of Linguistics, Academia Sinica, Taipei

Abstract

246 overt repairs, 653 complete repetitions and 475 partial repetitions were identified in an annotated corpus of spontaneous Mandarin conversations. On the basis of the data, this paper investigates Mandarin repairs and repetitions by segmenting them into the reparandum part, the editing part and the reparans part and by tagging them using the CKIP automatic word segmentation and tagging system. Results of the use of editing term, the distribution of part of speech and syllables in the reparandum are presented. Semantic differences and similarity in the discrepancy of tagging results of the reparandum and the reparans are also discussed.

1. Introduction

Speech repairs and repetitions are typical phenomena in disfluent spontaneous speech. Different from other disfluency such as pauses and fillers, repairs and repetitions have relatively complex context in syntax, semantics and lexis. Psycholinguistic studies and conversational analyses have long noticed repairs and repetitions in narratives and conversations ([5] and [7]). Repairs and repetitions show regular syntactic patterns and reflect pragmatic functions. From another point of view, most of the established speech recognition and parsing systems nowadays can successfully process well-formed and well-spoken utterances, for instance clearly read speech. But for the rest of "ill-formed" and "not properly spoken" utterances found in spontaneous conversation, no satisfying solutions have been found yet. All these fragmentary, incomplete or sometimes even regarded as incorrect speech stretches are by no means marginal phenomena. A number of empirical studies on repairs, pauses and repetitions for different languages have been done in the past years on different spoken corpora such as the Map Task Corpus, the TRAIN Corpus and the Switchboard Corpus ([1, 3, 4, 8]).

This paper studies features of repairs and repetitions produced in spontaneous Mandarin on the basis of eight hours of conversations. In modern Mandarin, the number of monosyllabic words decreases, whereas di- and trisyllabic words clearly increase. Our spoken data supports this notion; the average number of syllables per word is 1.65. The process of making new words in Mandarin involves compounding and abbreviating; lexical components of words may come from different syntactic categories first and new words are then created by abbreviating the compounded words. For instance, in the partial repetition **gao1 [pause] gao1yi1**, **gao1** means *high* and **gao1yi1** means *the first year in senior high school*, an abbreviated form of **gao1zhong1** (*senior high school*) **yi1nian2ji2** (*the first year*). The former **gao1** is an adjective independently, where the latter **gao1** in "**gao1yi1**" is a morphemic component of a noun. The position in which the process of compounding and abbreviating takes place could be possibly the position where the restart of a repetition may prefer. Mandarin examples are written in Pinyin and the numerals following each syllable are lexical tones. 1, 2, 3, 4, and 5 represent high flat, rising, contour, falling tones and the neutral tone.

2. Repairs and word repetitions in spontaneous Mandarin

The corpus data we used in this paper is extracted from the Mandarin Conversational Dialogue Corpus. This section will briefly introduce the corpus. Then the criteria of identifying repairs and word repetitions will be clarified and some statistics of the data will be presented and discussed.

2.1. Mandarin Conversational Dialogue Corpus

Mandarin Conversational Dialogue Corpus was collected from 2000 to 2001 at the Institute of Linguistics in Academia Sinica. It consists of 30 digitized conversational dialogues of a total length of 27 hours. 60 subjects were randomly chosen from Taipei, the capital city of Taiwan. Eight conversations spoken by nine female and seven male speakers were annotated by adopting a taxonomy scheme of four groups of spontaneous speech phenomena: 1) disfluency, 2) sociolinguistic phenomena, 3) particular vocalisation and 4) unintelligible and non-speech sounds. Disfluency includes for instance prosodic discontinuity such as breaks and word fragments, constructions not in agreement with the standard grammatical rules such as sentence fragments and speech repairs. Sociolinguistic phenomena are code switching (use of a foreign language or a Chinese dialect) and invented new words. Phonemic assimilations, syllable reductions, lengthening are some of the typical particular pronunciations in rapid and casual speech. Five human annotators transcribed the conversations in Chinese characters and in Pinyin, aided by interface "TransList" [9] to insert annotation tags and convert the horizontally arranged transcripts to a character-based and vertically presented database in Access format. Eventually, 53,225 annotation tags were used to annotate totally 140,579 transcribed syllables.

2.2. Repairs

Repairs must have a clearly identifiable reparandum item and a reparans item. That is to say, only disfluent sequences in which we can clearly identify what is to be corrected and what is the correction are annotated as repairs. In Example 1, **jin4kou3** is the reparandum and **chu1kou3** is the reparans. Also found in this example, **EN** (In our transcription system, all discourse particles are written in capital Latin letters) is an editing term, often used to bridge the gap between the reparandum and the reparans.

Example 1: overt repair

shi4	**jin4kou3**	**EN**	**chu1kou3**	**ma1?**
is	import	[discourse particle]	export	[interrogative particle]

Do you import uhn export products?

2.3. Repetitions

Repetitions in Mandarin are in a lot of cases perfectly legal syntactic constructions to put emphasis on particular components or to express subtle semantic nuance, for instance **da4da4de5** and **da4de5** both mean *big*, but having different discourse implications (**da4da4de5** has an emphasized effect).

Repetitions in this context are disfluent repetitions, which cannot be explained or justified by Mandarin grammatical rules. Complete repetitions are defined as fully repeated word sequences, for instance the repetition of the disyllabic word **yin1wei4** (*because*) in Example 2. Quite often, complete repetitions repeat words more than once. In partial repetitions only part of a word sequence is repeated, e.g. **kan4dian4 kan4dian4shi4** (*watch tele- watch television*) in Example 3.

Example 2: complete repetition
yin1wei4 yin1wei4 ta1 you3 jian4shen1 zhong1xin1
because because it has fitness center
Because because it has a fitness center.

Example 3: partial repetition
kan4 dian4 kan4 dian4shi4 zui4jin4 you3
watch electricity watch television recently has
xin1 dian4ying3
new movie
On the tele- on the television, there is a new film recently.

2.4. Repair and Repetition Sequences

We introduce the concept of a quasi-phrase to help the annotators dealing with spontaneous Mandarin. As mentioned above, due to some essential features of Mandarin such as free word order, no morphological markings at the surface level and the large number of variations of compounding and abbreviating words, it is hard to have a clear cut between morphology and syntax in Mandarin. A quasi-phrase is a part of a sentence representing a piece of information which itself is an undividable unit for listeners irrespective of syntactic structures. This is a more or less subjective judgment, but it is necessary, because there are no clear morphological markings helping the annotators dividing sentences consistently. Besides, the human annotators have detailed discussions prior to the annotations, so that the perceptual judgment should be to a certain extent consistent.

Chu1kou3 (export) is a verb. **Cong2tou2dao4jiao3** (from head to feet, meaning every part of the body) is an idiom. They are both quasi-phrases, although they contain different numbers of morphemes and they involve different syntactic levels of a sentence. A repair sequence is annotated from the site of the nearest quasi-phrasal boundary before the reparandum item to the site of the nearest quasi-phrasal boundary after the reparans item. Example1 is a simple case; the annotated repair sequence is **jin4kou3ENchu1kou3**. Similarly a repetition sequence is annotated from the site of the nearest quasi-phrasal boundary before the to-be-repeated item to the site of the nearest quasi-phrasal boundary after the repeated item. So in the idiom case above, **cong2tou2dao4 cong2tou2dao4jiao3** is annotated as a partial repetition sequence.

2.5. Data

Annotated data is summarized in Table 1. One thing to note is that the discourse marker **dui4** (*right*) is often repeated in spoken discourse. Sometimes it functions as a hesitation marker; sometimes it is repeated several times to win time for the speaker. They are used very often and it will influence the interpretation of the statistics. So we excluded all repetitions of **dui4** in the statistics. Furthermore, in Table 1 the number of the to-be-repeated syllables is not equivalent to that of the repeated syllables, because some of the occurrences were repeated more than once.

Table 1 shows that complete repetitions are the least likely to be accompanied by an editing term because of the smallest ratio of occurrences with an editing term over occurrences without an editing term. This result illustrates that complete repetitions, where no new information is produced, do not need an editing phase that much as partial repetitions and repairs do, where new information is uttered after the to-be-repeated and –repaired parts. It is surprising that more than the half (54%) of all occurrences of repairs and repetitions were identified together with an editing term. The editing terms counted here include perceivable paralinguistic sounds such as breathing, inhalation or short break. It will be interesting to examine the acoustic measurements of these editing terms to determine whether occurrences of repairs and repetitions are actually identifiable in terms of their acoustic features [6].

Table 1: Repairs and Repetitions in Mandarin Conversations.

	Repair	Complete Repetition	Partial Repetition	Total
Occurrences	246 (17.9%)	653 (47.5%)	475 (34.6%)	1374 (100%)
Occurrences with an Editing Term	157 (63.82%)	298 (45.64%)	287 (60.42%)	742 (54%)
With / Without an Editing Term	17.6:10	8.4:10	15.3:10	
Reparandum Syll. Involved in Repair	663			
Reparans Syll. Involved in Repair	1039			
To-be-Repeated Syll. Involved in Repetition		1043	680	
Repeated Syll. Involved in Repetition		1149	1590	
Total Involved Syll.	1702	2192	2270	6164

Moreover, within identified repair and partial repetition sequences the number of syllables involved in the reparandum part is about the half of the number of syllables involved in the reparans part. This implies that after correcting/repeating the actual reparandum item a continuation directing the utterance to a meaningful unit is needed. In our definition of repair and repetition sequences, a meaningful unit is a quasi-phrase. We will examine the length of quasi-phrases in later analysis.

3. Tagging experiment

In order to obtain consistent tagging results, we adopted the automatic word segmentation system developed for modern Mandarin by CKIP at Academia Sinica [2] to tag all identified occurrences of repair and repetition sequences by their part of speech (POS). Manual corrections of tags were necessary after the automatic tagging system was executed due to two main reasons. The tagging program was originally designed for written Mandarin, so a certain number of usages in spontaneous speech utterances are actually unknown words to the program. And our data of repairs and repetitions for parsing are themselves irregularities relative to standard Mandarin grammars, so some of the wrong parsing results were to be expected.

3.1. Tagged Results

After human annotators marked up all sequences of repairs, complete and partial repetitions, they furthermore segmented the sequences into three phases: the reparandum part, the editing part and the reparans part. The reparandum part contains all items before the editing term and the reparans part contains all items after the editing term. The editing term forms the editing part itself. The data was then processed by the CKIP tagging program. Because the sequences are quasi-phrases, they cannot directly reflect what exactly is repeated or repaired. Thus, we narrowed down the items to the first tagged POS of the reparandum item and the reparans item. The reason why only the first POS was considered is that the majority of repetitions involve only one POS (details cf. Section 4.2.). Table 2 shows the numbers of occurrence and the percentage of 1) the first reparandum and reparans POS in complete repetitions, 2) the first reparandum POS in partial repetitions, 3) the first reparans POS in partial repetitions, 4) the first reparandum POS in repairs, and 5) the first reparans POS in repairs. The POS categories are based on the CKIP tagging system. Predicative adjectives are included in the verb category. Foreign words and unrecognizable words are put into the category "foreign word". In the case of repairs, word fragments are tagged by "foreign word", too.

Table 2: POS in Repetitions and Repairs.

	Completely Repeated POS	Partially Repeated POS	Partially Repeated Target POS	Repaired POS	Repaired Target POS
Verb	102 (15.62%)	**126 (26.53%)**	**101 (21.26%)**	**61 (24.80%)**	**66 (26.83%)**
Preposition	82 (12.56%)	30 (6.32%)	21 (4.42%)	13 (5.28%)	9 (3.66%)
Noun	**274 (41.96%)**	**189 (39.79%)**	**226 (47.58%)**	**107 (43.50%)**	**117 (47.56%)**
Adverbial	152 (23.28%)	**113 (23.79%)**	**106 (22.32%)**	**48 (19.51%)**	**49 (19.92%)**
Conjunction	41 (6.28%)	15 (3.16%)	19 (4.00%)	6 (2.44%)	3 (1.22%)
Non-Predicative Adjective	2 (0.31%)	2 (0.42%)	2 (0.42%)	0 (0%)	0 (0%)
Foreign Word	0 (0%)	0 (0%)	0 (0%)	11 (4.47%)	2 (0.81%)

The first to note in Table 2 is that the distribution of the repeated POS in complete repetitions is very different from the other four cases. Prepositions are more often identified in complete repetitions (12.56%) than in partial repetitions (6.32% and 4.42% respectively) and repairs (5.28% and 3.66% respectively). Verbs are much less often completely repeated (15.62%) than partially repeated (26.53% and 21.26% respectively) and repaired (24.8% and 26.83%). However, partial repetitions and repairs show a symmetric similarity across reparandum and reparans.

Excluding complete repetitions, nouns are the most frequently repeated and repaired POS, then verbs and adverbials. More specifically, nominal, verbal and adverbial reparans makes up about 90% of the overall occurrences. Prepositions, conjunctions and non-predicative adjectives together are less than 10%. We notice that a number of the reparandum in repetitions and repairs are tagged differently from the reparans. This can reflect the discrepancy of the syntactic structure in disfluency and the morphological preferences of restarting in repetitions and repairs. In Section 5, we will look into the discrepancy by taking nouns as an example.

3.2. Number of POS

Figure 1 shows the numbers of POS involved in the reparandum part in repairs, complete and partial repetitions. It looks like the distributions of complete and partial repetitions are quite similar, where between 70% and 80% of both types of repetitions involve only one POS. While repetitions are uttered, no matter complete or partial, preferably only one POS is repeated. But of what morphological length is the preferred POS? We will look into the syllabic length of POS in the next section. Here we also observed that repetitions of a length of more than two POS are fairly rare (less than 5% of the overall repetitions).

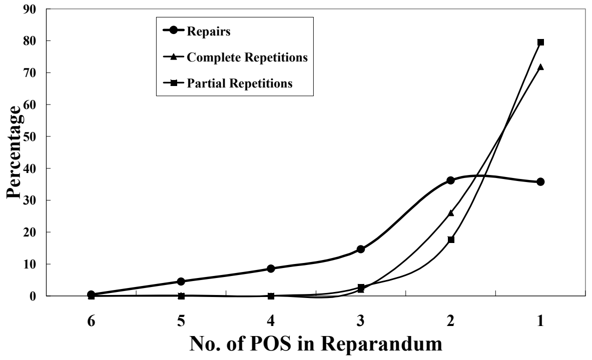

Figure 1: Number of POS of Reparandum Part in Repairs, Complete and Partial Repetitions.

For repairs, the curve shown in Figure 1 is different from those for repetitions. The reparandum part is most frequently composed of one and two POS, 35.7% and 36.2% respectively. As Figure 1 illustrates, up to six POS can be included in the reparandum part according to the definition of a quasi-phrase. Presumed that the reparandum part of repetitions and repairs needs to be a completely meaningful information unit, not in an arbitrary way, our data supports the notion that repetitions prefer short quasi-phrases and repairs prefer long quasi-phrases.

3.3. Number of Syllable

A single part of speech may include more than one morpheme. In Mandarin, morphemes are not as clearly defined as syllables, because a syllable is represented by a character in the written form. Thus, we examined the data from the perspective of syllabic size. And interestingly, the distribution is quite different from that of the POS size, as illustrated in Figure 2.

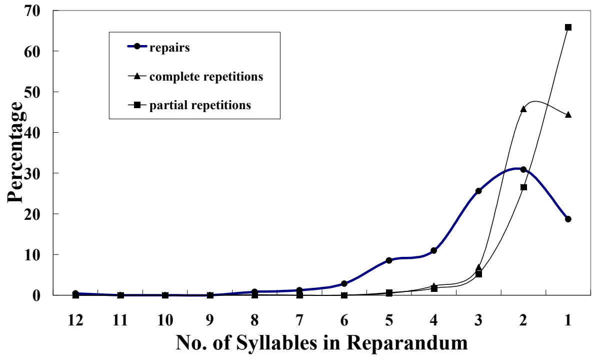

Figure 2: Number of Syllables of Reparandum Part in Repairs, Complete and Partial Repetitions.

Figure 2 shows that di- and monosyllabic words, though tagged by one POS, are completely repeated equally frequently. The number of the most frequently repeated syllables in partial repetitions is the same as that of the POS, which is one. For repairs, the number of monosyllabic reparandum items drops, whereas the number of trisyllabic reparandum items increases. On the whole, Figure 2 demonstrates that the preferred positions to restart a repetition or a repair is after one repeated syllable in the case of partial repetitions; after one and two repeated syllables in the case of complete repetitions; after two and three repaired syllables in the case of repairs.

4. Repetitions and repairs in nouns

Nouns are the most often repeated and repaired content word POS, so we present some preliminary observations regarding the POS discrepancy of the reparandum POS and the reparans POS.

4.1. Differently Tagged Reparandum of Nominal Reparans

In partial repetitions, among those reparans characters tagged as nouns (226 occurrences), only 168 of their reparandum characters were tagged as nouns (30 verbs, 9 prepositions and 17 adverbials etc.). Similarly, 97 of 117 targeting nouns in repairs were tagged as nouns (10 verbs, 3 prepositions and 3 adverbials etc.). In the examples given below, the POS tags in brackets are CKIP tags. Due to the lack of space, for further details please refer to [2].

Repeated	Repeated Target
hua4(VC)	**hua4mian4**(Na)
to draw	*picture*
ai4(VL)	**ai4xin1**(Na)
to love	*kindness, sympathy*
jiang1(P)	**jianglai2**(Nd)
with, by means of	*future*
na4(Dk)	**na4**(Nep) **ge1**(Nf)
therefore, then	*that CLASSIFIER*

Repaired	Repaired Target
kong1(VHC)	**fei1xing2yuan2**(Na)
to be empty	*pilot*
dao4(P)	**xia4**(Ncd)
(arriving) at	*the lower side*
ben3lai2(D)	**yi3qian2**(Nd)
originally	*the past*

Syllables are written in characters in Mandarin, so almost every character can be assigned a POS, because they all have certain meaning. Therefore, to detect repair or repetition patterns by means of syntactic categories may not be the ideal solution. As shown by our results, a mess of POS pattern is the unavoidable consequence.

4.2. Semantic Relation of Nominal Repairs

Examining the reparandum items which were also tagged as nominal ones, we obtained some interesting clues of semantic relations of the reparandum and the reparans. On the basis of our data on nominal repairs, the following semantic relations can be preliminarily found: 1) substitutions of hyponyms: the reparans specifies the reparandum, 2) substitutions of similar denotations: the reparans and the reparandum share similar information domain and 3) substitutions of antonyms: the reparandum and the reparans are antonyms.

na4bian1 (Ncd)	**mei3guo2**(Nc)	
over there	*USA*	*(sub. of hyponyms)*
kao3shi4 (Na)	**lian2kao3** (Na)	
exam	*entrance exam*	*(sub. of hyponyms)*
sui4 (Nf)	**nian2** (Nf)	
years-old	*years*	*(sub. of shared deno.)*
jiao4shou4 (Na)	**bo2shi4** (Na)	
professor	*PhD*	*(sub. of shared deno.)*
zi1xun4(Na)	**dian4nao3**(Na)	
information	*computer*	*(sub. of shared deno.)*
xian4shi2 (Na)	**shi4shi2** (Na)	
reality	*fact*	*(sub. of shared deno.)*
she4hui4ke1(Na)	**zi4ran2ke1**(Na)	
social sciences	*natural sciences*	*(sub. of antonyms)*
zheng4fu3 (Na)	**ren2ming2** (Na)	
government	*people*	*(sub. of antonyms)*

The examples above show that semantic relations can be empirically observed by means of real spoken data of repairs. It is especially interesting in the case of Mandarin because of the wide variety of morphological compounding of words.

5. Conclusion

This paper presented numeral preliminary results on Mandarin repairs and repetitions. Editing terms were found very frequently used in Mandarin repairs and repetitions. POS and syllabic features reflect different respects of the production. Disyllabic words are frequently repeated, although they are often tagged as one POS. The role POS plays in the production of repairs and repetitions does not imply that POS is also important in the detection. The discrepancy of tagged POS in the reparandum and the reparans due to the character-morpheme-syllable relation in Mandarin seems to prohibit detecting approaches based on POS patterns. Semantic differences of nominal reparandum and reparans were found in the data. It is an interesting issue worth further works, which can shed light on semantic relationships of Mandarin from a different point of view.

6. Acknowledgements

Part of the study presented in this paper has been supported by the National Science Council (91-2411-H-001-043) and the Ministry of Education (91-E-FA-06-4-4) of Taiwan.

7. References

[1] Carletta, Jean, Richard Caley & Stephen Isard. 1993. A Collection of Self-Repairs from the Map Task Corpus. *Tech. Rep*. University of Edinburgh.

[2] Chen, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang & H.-L. Hsu. 1996. ACADEMIA SINICA BALANCED CORPUS: Design Methodology for Balanced Corpora. *PACLIC 11*, pp. 167–176.

[3] Den, Yasuharu & Herbert Clark. 2000. Word Repetitions in Japanese Spontaneous Speech. *Proc. ICSLP'00*, 16–20 October 2000, Beijing, China, vol. 1, pp. 58–61.

[4] Heeman, Peter & Allen James. 1999. Speech Repairs, Intonational Phrases and Discourse Markers: Modelling Speakers' Utterances in Spoken Dialogue. *Computational Linguistics*, vol. 25, no. 4, pp. 527–571.

[5] Levelt, Willem J. M. 1983. Monitoring and Self-Repair in Speech. *Cognition*, vol. 14. pp. 41-104.

[6] Nakatani Christine & Julia Hirschberg. 1994. A Corpus-Based Study of Repair Cues in Spontaneous Speech. *Journal of the Acoustical Society of America*, vol. 95. pp. 1603–1616.

[7] Schegloff, Emanuel, Gail Jefferson & Harvey Sacks. 1977. The Preference of Self-Correction in the Organization of Repairs in Conversation. *Language*, vol. 53, no. 2, pp. 361–382.

[8] Shriberg, Elizabeth. 1996. Disfluencies in SWITCHBOARD. *Proc. of the International Conference on Spoken Language Processing*, 3–6 October, 1996, Philadelphia, Pennsylvania, USA, addendum, pp. 11–14.

[9] Tseng, Shu-Chuan & Yi-Fen Liu. 2002. Annotation Manual of Mandarin Conversational Dialogue Corpus. *Tech. Rep. CKIP-02-01*. Academia Sinica.