# Modelling Speech Repairs in German and Mandarin Chinese Spoken Dialogues

Shu-Chuan Tseng
Da-Yeh University
112 Shan-Jiao Rd. Da-Tsuen
Changhua, Taiwan 515
tseng@aries.dyu.edu.tw

**Abstract**

Results presented in this paper strongly support the notion that similarities as well as differences in language systems can be empirically investigated by looking into the linguistic patterns of speech repairs in real speech data. A total of 500 German and 325 Mandarin Chinese overt immediate speech repairs were analysed with regard to their internal phrasal structures, with particular focus on the syntactic and morphological characteristics. Computational models in the form of finite state automata (FSA) also illustrate the describable regularity of German and Mandarin Chinese speech repairs in a formal way.

## Introduction

Spontaneous speech analysis has recently been playing a crucial role in providing empirical evidence for applications in both theoretical and applied fields of computational linguistics. For the purpose of constructing more salient and robust dialogue systems, recent analyses on speech repairs, or more generally speaking, on speech disfluencies in spoken dialogues have tried to explore the distributional characteristics of irregular sequences in order to develop annotation systems to cope with speech repairs (Heeman and Allen 1999, Nakatani and Hirschberg 1994). This new research direction, nevertheless, has until recently merely focused on the surface structure of speech repairs on the one hand. On the other hand, except for very few investigations starting to deal with speech repairs across several languages (Eklund and Shriberg 1998), most of the studies on speech repairs have investigated only single languages. In addition, studies have shown that syntactic and prosodic features of spontaneous speech data provide empirical evidence with regard to reflecting the speaking habits of speakers, and also help to develop better parsing strategies and natural language processing systems (Heeman and Allen 1999, Hindle 1983). These systems should understand and react to the language use of human users (Lickley and Bard 1998, Tseng 1998).

This paper presents results of a comparative study of speech repairs with the goal of examining and modelling repair syntax by looking into empirical cross-linguistic speech data. In this paper, the phenomena of speech repairs are introduced first, followed by an empirical cross-linguistic analysis of speech repairs in German and Mandarin Chinese, which have different language typologies. Speech data, therefore, were collected to look for linguistic sequences and particularities of spontaneous speech, which usually cause difficulties for language dialogue systems. Syntactic patterns found in the comparative analysis have subsequently been formalised to make clear the internal structures of speech repairs. Formal modelling in FSA should finally show the formal characteristics of repair sequences in these two language systems.

## 1 Related Work

This section summarises previous results related

to speech repairs. First, a generally adopted template model of describing repairs is introduced, followed by a brief summary of recent studies on speech repair processing in German and Mandarin Chinese.

### 1.1 Template Model of Repairs

Most models of repair structures (Levelt 1983) apply a template-based approach. In principle, a template model is composed of three parts: reparandum (Rep), editing terms (Et) and alteration (Alt). The reparandum denotes the speech stretch, which needs to be repaired, whereas the alteration is the repair itself. Editing terms are sequences produced between the reparandum and the alteration, which often appear in form of silent or filled pauses and can also be absent, depending on the speaking situation. A classification system of repairs can be derived from the structural relations between the reparandum, the editing term and the alteration:

- *addition repairs*

Example: 會 影響到 整個 的 (Rep) 整個 企業 的 (Alt) 投資 意愿 (TWPTH Corpus)[1]

- *substitution repairs*

Example: Und unten runten ist halt die gelbe Mutter (Rep) äh (Et) die orange Mutter (Alt) (Sagerer *et al.* 1994)[2]

- *repetition repairs*

Example: En aan de rechterkant een oranje stip (Rep) oranje stip (Alt). (Levelt 1983)[3]

- *abridged repairs*

Example: I think that you get - it is more strict in Catholic schools. (Hindle 1983)

### 1.2 Grammar-Oriented Production of German Speech Repairs

German, an Indo-European language, is a language with a strong emphasis on grammatical flexion. Phrases with congruence in gender, number and case are important from syntactic and morphological viewpoints. Thus, phrasal boundaries may play a role in the production of German repairs. Results provided by Tseng (1999) empirically support the significant role of phrasal boundaries in German by examining German speech repairs. Phrasal boundaries seem to be the positions to start as well as to end speech repairs. The following utterance in which a German repair is produced clearly illustrates this phenomenon: "Ich habe einen Würfel *mit einer mit einem Gewinde*[4]", where **mit einer** is a phrasal fragment and **mit einem Gewinde**, starting from the phrasal beginning, is a complete phrase repairing the previous phrasal fragment. In her conversation analysis on self-repairs in German, Uhmann (1997) also mentions that repairs tend to appear at constituent boundaries in most cases, i.e., deleting problem sequences involved in repairs will result in the utterances containing speech repairs becoming well-formed.

### 1.3 Lexis-Oriented Production of Chinese Speech Repairs

One way to illustrate the differences in languages is to examine and to compare the types of speech repairs in the languages respectively. The modern description methodologies of grammar structures in German and Chinese (Chao 1968, Li and Thompson 1981) originated from similar theoretical backgrounds. However, Chinese has a great variety of compound words, but lacks grammatical markings at the morphological level. To be more specific, the word formation in Chinese is accomplished by combining morphemes, where each morpheme has its own lexical content and orthographic character. This is essentially different from the syntactic-morphological derivation as well as compounding in German.

Lee and Chen (1997) classified Chinese speech repairs in patterns and developed a language model for their language recognition system to

---

[1] Verbatim translation: will influence whole POSSESSIVE-particle whole industry POSSESSIVE-particle investment interests. Sentential translation : It will influence the whole the whole industrial investment interests.

[2] And beneath that is the yellow nut eh the orange nut.

[3] And at the right-side an orange dot orange dot.

[4] I have one cube *with a*[feminine, singular, dative, indefinite] *with a*[neuter, singular, dative, indefinite] *bolt*.

cope with speech repairs. However, they did not carry out any further investigations on the structure of repairs. Different from the production of German speech repairs, Chui (1996) proposed, in her studies on repairs in Chinese spoken conversations, that syntax seems to play a less important role than the lexical complexity and the size of words in the production of Chinese speech repairs. For instance, not the constituent boundaries, but the completeness of the lexical content and the scope of the lexical quantity of the words **should** (應該) and **engineer** (工程師) in the utterance *他 應該 不 應該 升 工程 工程師 那麼 快 的*[5], are the major factors which influence the production of repairs.

## 2 Data and Corpus

In order to examine the production of speech repairs in different languages, the German corpus BAUFIX and the Chinese corpus TWPTH were chosen to carry out further comparative analyses.

### 2.1 German Data: BAUFIX

The BAUFIX corpus (Sagerer *et al.* 1994) consists of 22 digitally recorded German human-human dialogues. 44 participants co-operated in pairs as instructor and constructor, where their task was to build a toy-plane. Because of the limited visual contact between dialogue partners in some given cases, subjects had to rely on their verbal communication to a great extent. This corpus setting was especially constructed to force subjects to repair their speech errors. For the purpose of this paper to investigate repair syntax, the corpus analysis is mainly concerned with immediate self-repairs. They were identified and hand-annotated by the author. In total, 500 speech repairs were classified according to their syntactic attributes such as categories and parts of speech. They were subsequently analysed with respect to the location of interruption and their repair structure.

### 2.2 Mandarin Chinese Data: Taiwan Putonghua Corpus (TWPTH)

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 dialogues and 30 monologues. Three dialogues were analysed for the study in this paper and each is about 20 minutes long. In total, 325 immediate speech repairs were identified in these three dialogues and they were annotated according to the POS system developed for the Sinica Corpus (CKIP 1995).

### 2.3 Comparison of Repair Data

Some central statistics on BAUFIX and TWPTH data are summarised in Table 1:

Table 1: Summary Statistics

| | BAUFIX | TWPTH |
|---|---|---|
| Language | German | Mandarin Chinese |
| total no. of words | 35036 words | 9168 words<br>47655 characters |
| total no. of repairs | 500 | 325 |
| no. words involved in repairs | 1823 words | 950 words<br>1622 characters |
| % repair-words of total words | 5.2 % | 10.4 % (word)<br>3.4 % (character) |
| % of phrases involved in repairs | PP 34.8 %<br>NP 38 % | VP 35.7 %<br>NP 41.2 % |

Table 1 shows that the percentage of problem words (words involved in speech repairs) is similar in both BAUFIX and TWPTH corpora. With regard to the number of words (i.e. lexical items) 10.4% of overall words in TWPTH are involved in repair sequences, whereas only 5.2% of words in BAUFIX are found in repair sequences. However, the statistics show a pattern, which is more closely related, 3.4% and 5.2% respectively, if we consider the number of characters instead of words in Chinese. Chinese

[5] Verbatim translation: He should NEGATION-particle should promote engineer(word fragment) engineer so quickly DISCOURSE-particle. Sentential translation: He *should should not* be promoted to *engineer*(word fragment) *engineer* so soon.

words can be mono- or multi-syllabic. In Chinese, lexical items are composed of characters, where each character is an independent meaningful monosyllabic morpheme. This study can possibly provide insights into the role of characters in Chinese at syntactic and morphological levels.

Other interesting results that can be noted from Table 1 are the types of phrases involved in repair sequences. In BAUFIX, because of the task-oriented corpus setting, few verbs were used. Instead, the focus is more on NPs and PPs, since the speakers had to express exactly what the parts look like and where to place them. Different from BAUFIX, the TWPTH speakers did not have to give exact descriptions. Therefore, a considerable number of verbs were used, which we can observe from the high percentage of VPs involved in repair sequences. However, in both corpora, NPs make up a high percentage, 38% and 41.2% respectively. For this reason, NPs will be further investigated for their syntactic structures.

## 3 Analysis of Repair Syntax in NPs

This section is concerned with the distribution and patterns of NPs in the context of repair syntax in German and Mandarin Chinese.

### 3.1 Regular Patterns

Among 190 NPs involved in repair sequences in BAUFIX, there are 147 NPs for which the internal structure within the NPs can be given exactly as follows (Tseng 1999),

NP => N
NP => DET + N
NP => DET + ADJ
NP => DET + ADJ + N
NP => DET + ADJ + ADJ + N
NP => so + DET + N
NP => so + DET + ADJ + N
NP => so + DET + ADJ + ADJ + N

where the other 43 NPs in repairs are abridged repairs, therefore, their internal structures cannot be determined.

Compared with German NP-repairs, Chinese speakers produce rather simple repair sequences in NPs. Only 62.7% (84 out of 134) of Chinese repairs found in the corpus are single NP phrases. The rest of repair sequences in which NPs are involved, contain other phrasal categories such as verb phrases or adverbials. Since these dialogues are concerned with normal and everyday conversations, no complicated noun phrases were used. These NP-repairs have the following structures:

NP => N
NP => DET
NP => DET + N
NP => ADJ + N
NP => QUAN + CLASS
NP => QUAN + CLASS + N

where QUAN denotes numbers and CLASS means classifiers in Chinese.

### 3.2 Syntactic Formalization

83.4% out of 147 specific NP repairs in German start at phrase-initial positions and end at phrase-final positions. In the Chinese data, only three NP-repairs among the 84 single NP-repairs were not traced back to the phrase-initial position. Phrasal boundaries play a role while speech repairs are produced in both languages, especially phrase-initial positions before the reparandum. The syntactic structure of the majority of German and Chinese repairs in NPs can be formally described by means of phrasal modelling.

Figure 1: Phrasal Modelling of German NP-Repairs

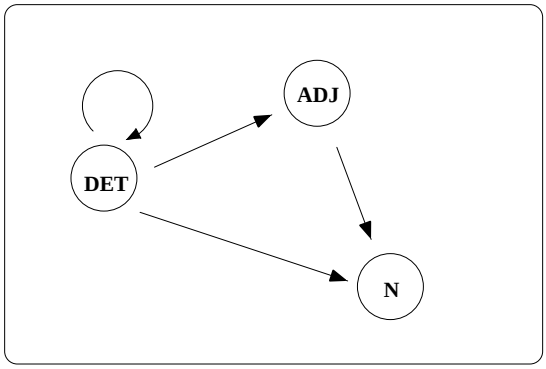



Figure 1 models 50% of NP repair sequences of the type DET ADJ N in BAUFIX, where the reflexive arrow on DET designates the sequence

DET DET. The first DET can be a fragmentary or a false determiner, whereas the second DET is supposed to be the corrected word accordingly. The initial element **DET** in a German noun phrase, i.e. the phrase-initial boundary is the most frequent location at which a repair is restarted. In other words, while producing repairs, speakers tend to go back to the determiner to repair NPs.

Although the data investigated here is not necessarily representative for most Chinese speakers, this result, does not empirically confirm Chui's conclusion (1996) that syntax should play a less important role than the lexical complexity and the quantity constraint of the to-be-repaired lexical items. Instead, the phrase-initial position seems to be the location to restart repairs in Chinese. Therefore, the results indicate that the lexical content of the to-be-repaired items tends to play a less important role than syntax in both languages.

### 3.3 Cross-Linguistic Differences

In contrast to the similarities between German and Chinese speech repairs mentioned in the sections above, differences can also be identified. Some differences can be noted through a comparison of repair syntax in German and Mandarin Chinese. It is more common for NPs in German to be repaired directly within NPs, whereas in Chinese NPs are often repaired within a more complex syntactic context, i.e. Chinese repairs are composed of more than one phrasal category. To investigate the syntactic and morphological distribution of speech repairs in both languages, the length of retracing in both languages is examined. The results are presented in Table 2.

Table 2: Distribution of Retracing

| retraced words or characters | German (words) | Chinese (characters) |
|---|---|---|
| 0 | 22.5% | 3.6% |
| 1 | 62.9% | 61.9% |
| 2 | 12.9% | 27.4% |
| 3 | 1.7% | 6% |
| 4 | 0 | 1.2% |

No similarity between German and Chinese was obtained by checking the number of retraced words in Chinese, because the majority of "the retraced parts" in Chinese are word fragments. But it is clearly shown in Table 2 that German words and Chinese characters play a similar role in the production of speech repairs. Whether it has to do with the syllabic weighting in both languages or the semantic content of characters in Chinese needs further linguistic investigation.

## 4 Formal Modelling

With regard to relations of repair syntax and the editing structuring in repairs, instead of only looking into their surface structure, the syntactic regularity in German and Chinese NP-repairs can be modelled in the form of finite state automata. We again take German as example.

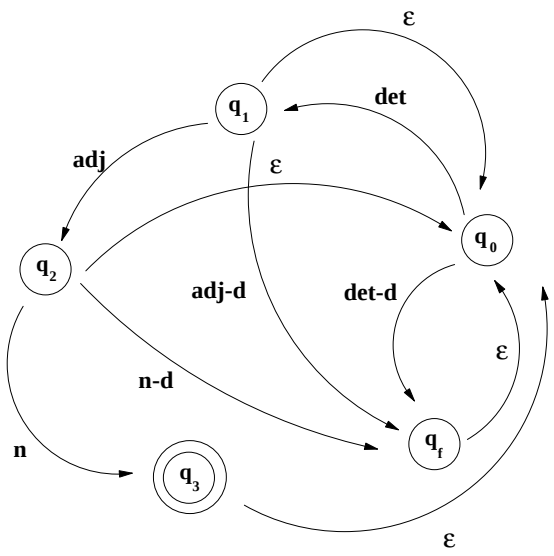
### 4.1 Finite State Automata

Finite state automata similar to ***M*** with ε-transitions denoted by a quintuple <Q, Σ, δ, q0, F> defined as follows can model more than 80% of overall German NP-repairs:
Q = {q0, q1, q2, q3, qf},
Σ = {det, adj, n, det-d[6], adj-d, n-d, ε},
q0 is the initial state,
F ={q3} and
δ(q0, det)=q1, δ(q1, adj)=q2, δ(q2, n)=q3,
δ(q0, det-d)=qf, δ(q1, adj-d)=qf, δ(q2, n-d)=qf,
δ(qf, ε)=q0, δ(q1, ε)=q0, δ(q2, ε)=q0,
δ(q3, ε)=q0

***M*** is graphically illustrated in Figure 2. Several particularities are described in this automaton. First, when NP-repairs are produced, no matter where the real problem word is located (It can be det-d, adj, adj-d, n or n-d), speakers tend to go back to the phrase-initial position to restart their speech. It the case of NPs, the determiner is the most frequent location for re-initiating a correct speech. The final position is in most cases phrase-final. Therefore, in ***M***, there is only one final state q3. This models the coherence within NP phrases in German that speakers usually complete phrases, after they have started them.

[6] Det-d, adj-d, and n-d denote fragmentary (or false) determiners, adjectives and nouns respectively.

Figure 2: Finite State Automaton ***M***




### 4.2 Discussion

The FSA ***M*** suggested above is suitable for the syntactic characteristics of speech repairs in both German and Chinese. Repair syntax has been taken into consideration from a procedural point of view, instead of simply describing the sequential structures. In this model, probabilities (for instance, word frequency or acoustic features) on the arcs can be implemented to operate a parsing system, which can deal with speech repairs. However, speech data of appropriate size are needed to obtain significant probabilities.

For more linguistic insights into the word-character relations in Chinese or across languages, i.e. the overlapping syntactic and morphological role of phrasal boundaries, further modification is needed to make the repair processing and detection in the Chinese case more realistic.

## Conclusion

This paper has shown that speech repairs not only play a decisive role in speech processing technology systems, they also provide empirical evidence and insights into the inherent linguistic characteristics of languages. Based on the results of corpus analysis, similar syntactic features of speech repairs in German and Chinese were identified and the repair syntax was formally modelled by means of phrasal modelling and finite state automata. Discrepancy at the morphological level of both languages was shown and more detailed investigations are necessary. Further analyses on acoustic-prosodic features of cross-linguistic data are currently being carried out.


## Acknowledgements

I'd like to thank the Sonderforschungsbereich (SFB 360) colleagues in Bielefeld who collected and pre-processed the BAUFIX data as well as the colleagues in the Industrial Research Technology Institute (IRTI) in Chu-Dong who kindly supported me with the TWPTH corpus data. Without them the investigation described in this paper would not have been carried out and this paper could not possibly have been written.


## References


Chao Y.-R. (1968) *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

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

CKIP (1995) *Sinica Balanced Corpus*. Technical Report no. 95-02/98-04. (in Chinese)

Eklund R. and Shriberg E. (1998) *Cross-Linguistic Disfluency Modeling: A Comparative Analysis of Swedish and American English Human-Human and Human-Machine Dialogs*. In: Proceedings of ICSLP'98. Sydney, Australia. pp. 2631-2634.

Heeman, P. and Allen, J. (1999) *Speech Repairs, Intonational Phrases and Discourse Markers: Modelling Speakers' Utterances in Spoken Dialogue*. Computational Linguistics 25/4. to appear.

Hindle, D. (1983) *Deterministic Parsing of Syntactic Non-fluencies*. In: ACL'83. Philadelphia, USA. pp. 123-128.

Lee Y.-S. and Chen H.-H. (1997) *Using Acoustic and Prosodic Cues to Correct Chinese Speech Repairs*. In: Proceedings of EUROSPEECH'97. Rhodes, Greece. pp. 2211-2214.

Levelt W. J. (1983) *Monitoring and Self-Repair in Speech*. Cognition 14. pp. 41-104.

Li C. and Thompson S. (1981) *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Lickley, R. J. and Bard, E. G. (1998) *When Can Listeners Detect Disfluency in Spontaneous Speech?* Language and Speech 41/2. pp. 203-226.

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

Sagerer G. and Eikmeyer H. and Rickheit G. (1994) *"Wir bauen jetzt ein Flugzeug": Konstruieren im Dialog. Arbeitsmaterialien*, Technical Report. SFB360 "Situierte Künstliche Kommunikation. University of Bielefeld, Germany.

Tseng S.-C. (1999) *Grammar, Prosody and Speech Disfluencies in Spoken Dialogues*. PhD Thesis. University of Bielefeld, Germany.

Tseng S.-C. (1998) *A Linguistic Analysis of Repair Signals in Co-operative Spoken Dialogues*. In: Proceedings of ICSLP'98. Sydney, Australia. pp. 2099-2102.

Uhmann, S. (1997) *Selbstreparaturen in Alltagsdialogen: Ein Fall für eine integrative Konversationstheorie*. In: Syntax des gesprochenen Deutschen. Ed. Schlobinski. Westdeutscher Verlag. pp. 157-180.