Extending and Scaling up the Chinese Treebank Annotation

Xiuhong Zhang, Nianwen Xue
Brandeis University
Waltham, MA 02453, USA
{xhzhang,xuen}@brandeis.edu

Abstract

We discuss on-going efforts to scale up the Chinese Treebank annotation and extending Chinese treebanking to informal genres like conversational speech, news groups and weblogs, as well as discussion forums. The original Chinese Treebank annotation scheme was designed for formal genres such as newswire and magazine articles, where the language is very formal and each document is carefully edited. When moving to informal genres, we can no longer assume that the data is error-free and we have to extend the annotation scheme to account for disfluencies. We show that the disfluencies can be characterized into a finite set of categories, consistent with what has been reported in theoretical linguistic literature. Treebanking is also a time-consuming process that requires extensive linguistic training from annotators, and the limited pool of qualified treebankers is a major obstacle for large-scale treebanking efforts. To address bottleneck, we implemented a procedure that decomposes the treebanking process into five self-contained steps. In so doing, we reduced the cognitive load on the annotators at each step and thus enlarged the annotator pool, and we show that we are able to increase the throughput by 30%.

1. Introduction

Large-scale treebanks [13,16] have proved to be instrumental in advancing the state of the art in syntactic parsing, a fundamental technology in Natural Language Processing. Early treebanking efforts started with the annotation of carefully edited textual data such as Wall Street Journal articles (Penn Treebank) and Xinhua newswire articles (Chinese Treebank) where the data can be assumed to be error-free. There is a growing need, however, for annotated data in informal genres, which include conversational speech, news groups, web blogs, and online discussion forums. Annotation of such informal genres requires substantial extension to the original annotation guidelines to cover new linguistic (and sometimes non-linguistic) phenomena. We show that while these new linguistic phenomena are diverse, they have clear patterns that can be characterized and classified, a pre-requisite to successful annotation.

Treebanking is a time-consuming process and scaling up treebanking efforts while maintaining annotation quality is always a challenge. This is because it takes a long time to train new treebankers and they have to have significant prior formal linguistic training to be able to understand the grammatical formalisms and make the necessary linguistic distinctions between different types of linguistic structures. These requirements severely limit the pool of qualified treebankers, making it difficult to scale up an annotation effort simply by hiring more qualified annotators, even if cost is not a factor. In reality, cost is another factor that has to be considered.

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 27–34,
Tianjin, China, 20-21 DEC. 2012

We address this challenge by decomposing the treebanking process into smaller, self-contained tasks, which reduces the cognitive load on the annotators so that more annotators can participate without having to understand all aspects of the treebanking annotation efforts. This is in keeping with the trend of using crowd-sourcing to quickly collect large amount of annotated data using platforms such as Mechanic Turk, although we did not go as far, as there has been no evidence thus far for successful treebanking effort by using a large number of minimally trained annotators, to the best of our knowledge. What we sought is a middle ground between crowd sourcing and the traditional treebanking practice of using highly trained annotators. The rest of the paper is organized as follows. In Section 2, we give a brief overview of the Chinese Treebank annotation scheme. Section 3 describes characteristics of informal genres and how the new phenomena are treated in our revised annotation scheme. In Section 4 we present our new workflow that decomposes our annotation task into smaller, self-contained tasks. We also discuss advantages of such an approach and problems that still exist. Section 5 presents some relevant statistics and Section 6 discusses related work. Section 7 concludes our paper.

2 An overview of the existing Chinese Treebank annotation framework

The Chinese Treebank (CTB) is a fully segmented, part-of-speech (POS) tagged, and syntactically bracketed Chinese corpus annotated in a phrase structure framework [16]. The CTB adopts the same architectural and representation framework used by the Penn Treebank [13], as is natural given the success of the Penn Treebank annotation style and the affinity of the research groups. Just like the Penn Treebank, the CTB has three layers of annotation: word segmentation / tokenization, part-of-speech (POS) tagging, and syntactic bracketing. There are three sets of guidelines [17,18,19], one for each layer, and the syntactic bracketing guidelines are by far the most complex among the three. At the part-of-speech tagging layer, each word token in the corpus is assigned one of the 34 tags in the CTB POS tagset. At the syntactic bracketing level, the CTB annotation framework uses three types of formal devices to represent the syntactic structure of a sentence. They include labeled brackets for representing constituents (See Appendix 1 for a list of phrase labels), function categories for representing the grammatical functions in the form of dash tags attached to the phrase label, and empty categories and traces that represent phonological null elements and long-distance dependencies. An example taken from the Penn Chinese Treebank is presented below, and this example has all three elements.

(1)
(IP-HLN (NP-SBJ (NN 经济/economics)
(NN 专家/expert))
(VP (VV 提出/propose)
(NP-OBJ (CP-APP (IP (NP-SBJ (-NONE- **pro**))
(VP (ADVP (AD 进一步/further))
(VP (VV 扩大/expand)
(NP-OBJ (NP-PN (NR 海南/Hainan))
(PP (P 对/toward)
(NP (NN 外/outside)))
(NP (NN 开放/open))))))
(DEC 的/DE))
(NP (NN 系列/series)
(NN 建议/recommendation)))))
"Economic experts proposed a series of recommendations to further expand the opening of Hainan to the outside."

The original CTB was annotated in two stages. The first stage is the word segmentation/POS tagging stage where Chinese sentences are segmented into words and each word token is assigned a POS tag. The second stage is the syntactic bracketing stages, where each constituent is grouped together and assigned a phrase label. Where appropriate, one or more functional tag is appended to the phrase label and empty categories are added.

3 Extending the Chinese Treebank annotation to informal genres

The original CTB annotation scheme [2] was designed for genres such as newswire and magazines, where the language is very formal and each document is carefully edited. As we move to informal genres such as forum discussions, web blogs, online instant chatting, telephone phone conversations and so on, we encounter many new phenomena that have to be accounted for. These include typographic errors, incomplete sentences, non-speech elements such as background noises that are recorded in transcriptions of speech, disfluent (and yet understandable) utterances. These new phenomena fall into two broad categories: non-linguistic phenomena such as typographical errors that are introduced due to haste and carelessness, and linguistic phenomena such as disfluencies in conversational speech where a speaker

has to repair the utterance s/he produced under the time pressure. We discuss these broad categories and how they are treated in our annotation framework in the next two subsections. As of this writing, we have annotated over 400,000 words in the informal genre based on the extended annotation guidelines.

3.1 Typographical errors and non-speech elements

Typographical errors do not have a linguistic explanation, and they are produced due to carelessness, fatigue, or haste on the part of the authors or transcribers. Because we adhere to the practice of "not altering the source data and only adding annotation" in the annotation process, we add tags at both the part-of-speech tagging and syntactic bracketing levels to mark up these errors where appropriate.

The first type of typographic error is mis-spelled Chinese characters. Since words with this type of typographic error usually can still be interpreted, we segment and POS-tag them as if we were annotating their correct counterpart. For example, we annotate 幸口开河 as if it were 信口开河. We treat it as one word and label it as VV at the POS level. We do NOT change the original characters in the text, as a matter of principle.

(2) 幸(信)口开河/VV
talk irresponsibly "talk irresponsibly"

The second type of typographic error is characters written in the wrong order. It is different from the first type in that the word boundaries are messed up and cannot be segmented and POS-tagged as if it were correct. In this case we add a new POS tag NOI ("Noise") to tag the messed up parts and group the entire string as TYPO, a phrase label:

(3) (TYPO 事/NOI 类/NOI 各/NOI 故/NOI)

? type each ?
Correct:各/DT 类/M 事故/NN

every type accident

"all sorts of accidents"

The mechanical errors are random and cannot be fully anticipated, so broad encompassing categories such as NOI (POS tag), TYPO (phrase label) are used to label them.

We also added a phrase label SKIP to mark up sequences of non-speech elements, indicating that this portion can be ignored when the text is interpreted. Non-speech elements include background noises recorded in speech transcripts, boundary markers and so on.

3.2 New linguistic phenomena in informal genres

There are also a large number of new linguistic phenomena that cannot be accommodated by the original annotation framework, and these include incomplete sentences, embedded speech, fillers and other types of disfluencies. These are linguistic issues whose cognitive processes and pragmatic effects have been widely discussed in the literature [3] [4] [5] [6] [7]. Based on the studies of these issues in the literature, we added 4 phrase labels and 2 functional tags to account for them.

Incomplete utterances (INC)

In informal genres, especially in conversational speech, there are often incomplete utterances. To label such utterances, we added the phase label INC to the original annotation scheme. INC is a label for root nodes only, similar to FRAG, IP, CP in the original guidelines. It is different from FRAG in that the latter is semantically complete even though it does not have the typical structure of a sentence. Utterances marked INC are incomplete both in its syntactic structure and in its semantic interpretation. (4) is an example.

(4)

```
(INC (CP-CND (ADVP (CS 如果/if))
                (IP (NP-SBJ (PN 他们/they))
                  (VP (VV 来/come))))
     (PU ，)
     (ADVP (AD 那/then)
     (NP-SBJ (PN 我/I)
     (VP-UNF (ADVP (AD 就/then))))))
```

"If they come, then I will …"

Fillers (FLR)

In conversational speech, the speaker often needs to think about what s/he wants to say and use fillers to buy her/him some time. The linguistic devices s/he uses for this purpose are called fillers. Fillers do not have a significant role to play in the syntactic structure of a sentence and they do not add to the semantic content of a sentence either. Fillers form a close set because there are only a finite number of them, but there is little restriction on where they can occur in the sentence. Fillers in

Chinese include “嗯/um, uh-huh” , “呃/Ugh”, “唔/oh”, “啊/Ah”, “这个/Eh”, “那个/Eh”, etc.

(5)

```
(IP (NP-SBJ (PN 你/you))
    (VP (ADVP (AD 多/more))
        (FLR (INF 那个/that one))
        (VP (VV 长/grow)
            (NP-OBJ (QP (CLP (M 个/CL)))
                        (NP (NN 心眼儿/mind))))))
```

“You should be more mindful.”

Disfluency (DFL)

In conversational speech, a speaker often has to repeat what s/he has just said, or abandon what s/he just said and restart with revised content. This is a phenomenon called *repair* in speech literature. There is extensive literature on speech repairs [8][9][13]. Typically, a speech repair instance can be characterized as a template that consists of a *reparandum* and an *alteration* [13]. The reparandum is the speech sequence that is erroneous or inappropriate, while the alteration represents the correction of the problematic sequence. The alteration can delete from, add to, substitute for, or repeat the problematic sequence. Or it can be a fresh restart that has little resemblance to the problematic sequence. The alteration is essential to the completeness of the syntactic structure of a sentence, while the reparandum, like fillers, can be considered to be “extra” material. We label such extra material with the phrase label DFL. The idea is that when such extra material is stripped, the remaining structure is a syntactically well-formed sentence.

(6a) Repetition

```
(IP  (PP-TMP (P 到/up to)
             (NP (NT 现在/now)))
     (FLR (SP 啊/Ah))
     (PU ，)
     (NP-SBJ (-NONE- pro))
     (VP (DFL (VP (ADVP (AD 已经/already))
                  (VP (VE 有/have))))
         (PU ，)
         (ADVP (AD 已经/already))
         (VP (VE 有/have)
             (IP-OBJ (NP-SBJ  (DNP  (DNP (QP (CD 七百多万
/more than 7 million))                    (DEG 的))
                                    (DNP (NP (NN 个人/individual travelling
游))
                                         (DEG 的))
                                    (NP (NN 旅客/visitors))))
                             (VP (VV 来/come)
                                 (NP-PN-OBJ (NR 香港/Hongkong)))))))
```

“Up to now, there have been, have been more than 7 million individual visitors visiting Hongkong.”

(6b) Substitution

```
( (NP-Q (SPK [Speaker_A1])
        (DFL (NT 昨天/yesterday))
        (FLR (IJ 哎/ah))
        (PU ,)
        (NP (NT 今天/today))
        (SP 啊/Ah)
        (PU ?)))
```

“Yesterday, (you mean) today?”

(6c) Restart

```
( (CP-Q (SPK [Speaker_A])
        (INTJ (NN 咯/um))
        (PU ,)
        (DFL (ADVP (INF 那/then))
           (NP-SBJ (PN 它/it))
           (VP-UNF (ADVP (AD 怎么/how come))))
        (PU ,)
        (IP (NP-SBJ (-NONE- pro))
          (VP (ADVP (AD 不/not))
                (VP (VV 知道/know)
                  (NP-OBJ (DP (DT 怎么/how))
                              (QP (CLP (M 回/Classifer)))
                              (NP (NN 事儿/matter))))))
        (SP 啊/Ah)
        (PU .)))
```

“Uh, then how come it, I don’t know what the matter is.”

Embedded utterances (MBD)

Embedded utterances are cases where the utterance of one speaker is embedded in the utterance of another speaker. This happens when one speaker interrupts when another speaker has not finished his/her sentence. The embedded utterances are usually short comments that indicate consent, etc.

(7)

```
(SPK [Speaker_A])
 (CP (IP (CP-ADV (IP (NP-SBJ (-NONE- pro))
                                (VP (ADVP (CS 一/at first))
                                    (VP (VV 开始/begin))))
                             (SP 吧/ba))
                     (PU ,)
                     (MBD (INTJ (SPK [Speaker_B])
                                (IJ 啊/Ah)
                                (PU ,)))
                     (SPK [Speaker_A])
                     (CP (IP (NP-SBJ (PN 他/he))
                             (VP (VV 要/want)
                               (VP (VP (VV 做/do)
                                       (NP-OBJ (NN 科学
/science)
                                               (NN 研究
/research)))
                                   (VP (VV 用/use)))))
                       (SP 的/DE))
                     (PU ,)
                     (DFL (IP (NP-SBJ (-NONE- pro))
                              (VP-UNF (VC 是/BE)
                                      (PP (P 用/by means of)
                                        (NP (PN 我/I)))))
                       (DEG 的/DE))
                     (PU ,)
                     (CP (IP (NP-SBJ (-NONE- pro))
                             (VP (VC 是/BE)
                               (IP-PRD (NP-SBJ (PN 我/I))
```

(VP (MSP 去/go)
(VP (VV 申请/apply))))))
(SP 的/DE))))
"Speaker A: 'At first'
Speaker B: 'Ah'
Speaker A: 'He wanted to use it for scientific research. He used mine, it's me who applied for it.'"

In addition to the new phrase labels above, we have also added two new functional tags (-DIS,-UNF). –DIS represents discourse markers and – UNF denotes incomplete phrases in a syntactic parse. –UNF is different from INC in that INC is a root node label (label for the entire sentence) while –UNF is functional tag indicating a non-root node label is incomplete. In general, functional tags can be attached to any phrase label to provide additional information. A constituent bearing the -UNF tag can be a NP, VP, etc.. A constituent bearing the –DIS tag is usually an adverbial phrase (ADVP), although it can be other types of phrases.

-DIS: functional tag indicating discourse marker

In spoken discourse, some lexical items demonstrate the discourse function of linking two stretches of discourse, with their original semantic meanings weakened or 'bleached' [10] [11]. They serve to indicate that an adverbial phrase functions as a discourse marker rather than an indicator of time, location, manner, reason and so on. The following is an example of discourse markers:
"*就是说/that is to say*" （sometimes for further clarification, but often indicates that the speaker has got something to say)

(8)
(CP (IP (CP-CND (IP (ADVP (AD 所以/so))
(NP-SBJ (PN 你/you))
(VP (ADVP (AD 要是/if))
(VP (VV 回来/return))))
(SP 的话/if))
(NP-SBJ (PN 你/you))
(VP (ADVP (AD 就/then))
(VP (VV 可以/can)
(VP (VV 知道/know)
(PU ,)
(IP-OBJ (ADVP-DIS (AD 就是说/that's to say))
(PU ,)
(FLR (INF 这/this))
(NP-SBJ (DP (PN 那些/those))
(NP (NN 东西/stuff)))
(FLR (SP 啊/Ah))
(PU ,)
(VP (PP-ADV (P 跟/with)
(NP (PN 他/him)))
(VP (VV 对路/fit))))))))
(SP 啦/la)
(PU .))

"So if you come back, then you know, that's to say, those stuff fit him.'"

-UNF: Functional tag indicating unfinished constituent

(9)
(INC (CP-CND (ADVP (CS 如果))
(IP (NP-SBJ (PN 他们))
(VP (VV 来))))
(PU ，)
(ADVP (AD 那)
(NP-SBJ (PN 我)
(VP-UNF (ADVP (AD 就))))))
"If they come, then I will …"

For the sake of completeness, the revised tagsets (phrase labels and functional tags) for the Chinese Treebank are presented in Tables 1 and 2 respectively, with new tags marked by *.

Label	Description	Label	Description
ADJP	Adjective phrase	LCP	Localizer phrase
ADVP	Adverb phrase	LST	List marker
CLP	Classifier phrase	*MBD	Embedded utterance
CP	Clause headed by a complementi Zer	IP	Simple clause
*DFL	Disfluency	NP	Noun phrase
DNP	Phrase formed by "XP+DEG"	PP	Prepositiona l phrase
DP	Determiner Phrase	PRN	Parenthetica l
DVP	Phrase formed by "XP+DEV"	QP	Quantifier phrase
*FLR	Filler	*SKIP	Skip
FRAG	Fragment	*TYPO	Typographic error
*INC	Incomplete	UCP	Unlike coordination
IP	Simple sentence	VP	verbphrase
LCP	Localizer Phrase		

Table 1: revised phrase labels. * indicates new labels

Function tags			
Tag	**Description**	**Tag**	**Description**
ADV	Adverbial	MNR	Manner
APP	Appositive	OBJ	Direct object
BNF	Beneficiary	PN	Proper noun phrase
CND	Condition	PRD	Predicate
DIR	Direction	PRP	Purpose or reason
*DIS	Discourse connective	Q	Question
EXT	Extent	SBJ	Subject
FOC	Focus	TMP	Temporal
HLN	Headline	TPC	Topic
IJ	Interjective	TTL	Title
IMP	Imperative	*UNF	Incomplete phrase
IO	Indirect Object	VOC	Vocative
LGS	Logical subject	WH	Wh-phrase
LOC	Locative		

Table 2: Revised functional tags. * indicates new tags

4 Scaling up the CTB annotation by broadening the annotator pool

The original Chinese Treebank was annotated in two stages: the word segmentation/POS tagging stage and the syntactic bracketing stage. In the word segmentation/POS-tagging stage, an annotator adds word boundaries and POS tags to words in a corpus. In the bracketing stage, an annotator groups the constituents and organizes them into a hierarchical structure, adding functional categories and empty categories to the syntactic structure of a sentence, following a set of treebanking guidelines that are close to 200 pages [20].

Moving to informal genres and scaling up the annotation effort magnify two challenges in Chinese Treebanking. The first one is that in informal genres, the rules for using punctuation marks are very loose, and in conversational speech, punctuations are of course not used at all and they are added later on by transcribers. These lead to unreliable sentence boundaries if we follow the standard practice of using periods, question marks and exclamation marks as markers of sentence boundary. Another challenge is that as we increase the volume of annotation, we need more trained treebankers. Training a treebanker takes a long time and treebankers have to come with extensive formal linguistic training to begin with.

To meet these challenges, we implemented a new workflow that consists of five stages, illustrated graphically in Figure 1. The new workflow decomposes the treebanking process into five self-contained steps, namely, sentence boundary detection, word segmentation/POS tagging, constituent grouping, functional category and empty category annotation, and post-processing and validation. Compared with the original Chinese Treebank workflow, we added a sentence boundary detection stage where we perform sentence segmentation. More importantly, we decomposed the bracketing stage, the most difficult aspect of treebanking, into two steps. The first step is to group the constituents of a sentence into a hierarchical structure. This step produces a bare-bone syntactic parse for a sentence. In the second step, we add functional tags and empty categories to the bare-bone structure to produce a full parse.

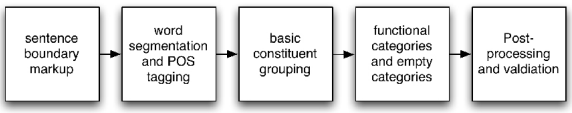

Figure 1: Annotation work flow

The purpose of the new workflow is to reduce the cognitive load of the annotators and thus increase the pool of qualified annotators. Treebankers now do not all have to understand all aspects of treebanking. Some treebankers can concentrate on grouping the constituents correctly and others can focus on the functional tags and empty categories. This is in keeping with the spirit of crowdsourcing [12], the essence of which is to design annotation tasks in a way that increases the annotator pool so that minimally trained annotators can work on them. Our new workflow can be viewed as a small step in that direction. As a result of this new workflow, four treebankers rather than two can work on this project. We did an internal performance evaluation about the amount of data we are able to annotate per week, and compared to our work rate prior to the introduction of the new work flow, our speed accelerated by 30% with more consistency and accuracy.

The new workflow also allow cross-checking between different layers of annotation. Treebankers working on the bare-bone structure can check er-

rors in word segmentation and POS tagging, and treebankers working on functional tags and empty categories can check the bare-bone structures. The new workflow also opens up more opportunities for automation. Automatic pre-processing was performed at each step. Sentence-segmented data is automatically word segmented and POS-tagged using a word segmenter/POS-tagged we developed in-house [14] before they are manually corrected. Word segmented and POS-tagged data is then automatically parsed using the Berkeley parser retrained on available Chinese Treebank data. Finally, we developed a simple rule-based tool that automatically adds functional tags and empty categories to the bare-bone parses before they are corrected.

5 Some relevant statistics

Our raw texts include newswire, magazine articles, broadcast news, broadcast conversations, and weblogs. As of this writing, we have annotated over 400,000 words in the informal genre based on the extended annotation guidelines. Here is some statistics based on an analysis of 461 files with 396,874 words:

label	occurrences
DFL tags	2819
FLR tags	1854
INC tags	637
TYPO tags	13
SKIP tags	281
MBD tags	167
-DIS tags	150
-UNF tags	924

6 Related work

The success of the Penn Treebank [15] has spurred the development of a large number of treebanks in many different languages, but most of the early treebanking efforts are directed at the formal genres. Specific to Chinese, there are number of significant treebanking efforts (Sinica Treebank and Tsinghua Treebank), but the Chinese Treebank is one of the early ones. There are relatively few efforts directed at annotating informal genres. The Switchboard Corpus is one notable exception [17]. It is a speech corpus annotated following guidelines that extend the Penn Treebank annotation guidelines. To the best of our knowledge, there is no similar annotation in Chinese.

7 Conclusion

We presented our effort to extend the Chinese Treebank annotation to informal genres, and in the process, we extended the Chinese Treebank annotation guidelines to account for new linguistic phenomena, which include typographic errors and disfluent speech. We also presented a new workflow aimed at scaling up the current treebanking effort. The new workflow decomposes the complex treebanking into more manageable subtasks. In doing so, it reduces the cognitive load on treebankers and thus increases the annotator pool.

Acknowledgements
We gratefully acknowledge the effort of our annotators. This work is funded by the DAPRA via contract HR0011-11-C-0145 entitled "Linguistic Resources for Multilingual Processing". All opinions expressed here are those of the authors and do not necessarily reflect the views of DARPA.

References

[1] Nancy Ide, Laurent Romary, International standard for a linguistic annotation framework, SEALTS '03 Proceedings of the HLT-NAACL 2003 workshop of Software engineering and architecture of language technology systems – Volume 8 Pages 25-30

[2] Nianwen Xue, Fei Xia, Fu-Dong Chiou and Martha Palmer. 2005. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, Volume 11 / Issue 02 / June 2005 , pp 207-238
-

[3] Yan Wang, 2011. A Discourse-Pragmatic Functional Study of the Discourse Markers Japanese *Ano* and Chinese *Nage*. Intercultural Communication Studies XX: 2 (2011)

[4] Arnold, J. E., and M. K. Tanenhaus. in press. Disfluency effects in comprehension: How new information can become accessible. The processing and acquisition of reference, ed. by E.Gibson and N. Perlmutter. MIT Press

[5] Blackmer, E. R., and J. L. Mitton. 1991. Theories of monitoring and the timing of repairs in spontaneous speech. Cognition 39.173-194.

[6] Bortfeld, H., S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan. 2001. Disfluency rates in spontaneous speech: Effects of age, relationship, topic, role, and gender. Language and Speech 44.123-147.

[7] Shriberg, E. 1996. Disfluencies in Switchboard. Proceedings, International Conference on Spoken Language Processing, Addendum, 11-14. Philedephia

[8] Lin, Che-kuang, et al., 2005 Important and New Features with Analysis for Disfluency Interruption Point (IP) Detection in Spontaneous Mandarin Speech. Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop

[9] Georgila, Kallirroi, et al., 2010 Cross-Domain Speech Disfluency Detection. ACL

[10] Biq, Yung-O. 2001. The Grammaticalization of Jiushi and Jiushishuo in Mandarin Chinese. Concentric: Studies in English Literature and Linguistics 27. 53-74.

[11] Su, Lily I-wen. 1998. Conversational coherence: The use of ranhou in Chinese spoken discourse. In Shuanfan Huang (ed.), Selected papers from the second international symposium on languages in Taiwan, 167-181. Taipei: Crane

[12] Crowdsourcing Translation: Professional Quality from Non-Professionals. Omar Zaidan and Chris Callison-Burch, 2011. In Proceedings ACL-2011.

[13] Tseng, S.-C. 2006. Repairs in Mandarin conversation. *Journal of Chinese Linguistics* 34(1): 80-120.

[14] Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. International Journal of Computational Linguistics and Chinese Language Processing, 8(1):29-48.

[15] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the penn treebank. *Comput. Linguist.* 19, 2 (June 1993), 313-330.

[16] Nianwen Xue, Fu-Dong Chiou and Martha Palmer, 2002, Building a Large-Scale Annotated Chinese Corpus. In Proceedings of the 19th. International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, 2002.

[17] Godfrey, J. J., Holliman, E. C., and McDaniel, J., "SWITCHBOARD: Telephone speech corpus for research and development," Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc., pp. 517-520, 1992.

[18] Xia, Fei. 2000a. The Segmentation Guidelines for the Penn Chinese Treebank 3.0. University of Pennsylvania Technical Report, IRCS00-06

[19] Xia, Fei. 2000b. The POS Tagging Guidelines for the Penn Chinese Treebank 3.0. University of Pennsylvania Technical Report, IRCS00-07

[20] Xue, Nianwen and Fei Xia. 2000. The Syntactic Bracketing Guidelines for the Penn Chinese Treebank 3.0. University of Pennsylvania Technical Report, IRCS00-08