Investigating Non-Sentential Utterances in a Spoken Chinese Corpus

Kwong-Cheong Wong, Jonathan Ginzburg

Investigating Non-Sentential Utterances in a Spoken Chinese Corpus

Kwong-Cheong Wong   Jonathan Ginzburg
Université Paris-Diderot (Paris 7)

wongkwongcheong@gmail.com
yonatan.ginzburg@univ-paris-diderot.fr

Abstract

This paper describes a preliminary investigation into Chinese non-sentential utterances (NSUs) in a corpus of spoken Mandarin. It presents, with examples, a corpus-based taxonomy of Chinese NSUs. This taxonomy builds on the one by Fernández and Ginzburg for English NSUs in the British National Corpus (BNC) [1]. Partly due to the distinctiveness of spoken Chinese, eight new classes are added and their reasons for addition are explained. The paper concludes with discussions for future work.

Key Words- Dialogue, Non-sentential utterances, Corpus analysis, Spoken Chinese

1. Introduction

Dialogue is full of fragmentary utterances which are incomplete sentences but nevertheless convey a complete sentential meaning in the given context. These utterances are called *non-sentential utterances* (NSUs) in the literature (see, e.g., [1], [2], [3] and [4]). In recent years, these frequently occurring NSUs, whose meaning being highly context dependent, have drawn the attention of researchers, especially those in the field of computational dialogue systems, for the reason that these NSUs can provide us with various clues to the structure of context. In fact, based on an investigation into NSUs in the British National Corpus (BNC), Ginzburg has developed a theory of meaning for spoken interaction which is basically a theory of dialogue context [3]. Building on the work by Fernández and Ginzburg on English NSUs in the BNC [1], this paper describes a preliminary investigation into Chinese NSUs in a corpus of spoken Mandarin. This paper is structured as follows: Section 2 introduces the NCCU Corpus of Spoken Mandarin, the corpus we used in this research. Section 3 presents, with examples from the corpus, our corpus-based taxonomy of Chinese NSUs and explains the reasons why several new classes which are not in the taxonomy by Fernández and Ginzburg for English NSUs in the BNC (called, henceforth, the FG

taxonomy) are added. Section 4 concludes the paper with discussions for future work.

2. The NCCU Corpus of Spoken Mandarin

While there are many spoken English corpora which have been established for different research and educational purposes, there are few spoken Chinese corpora. The NCCU Corpus of Spoken Chinese, which is established and maintained by the National Chengchi University (NCCU), Taiwan, is a project of language documentation whereby open online access to data of three major dialects of Chinese, namely, spoken Mandarin, spoken Hakka, and spoken Southern Min, are provided for non-profiting-making research [5]. In the research reported in this paper, we used only the Mandarin sub-corpus of this NCCU Corpus, which is called, henceforth, the NCCU Corpus of Spoken Mandarin (NCCUCSM) and is available on the web at http://140.119.172.200/. The spoken data of NCCUCSM are mostly in the form of spontaneous face-to-face conversations. The data we used for this research are the first 10 conversations which are available for free. These 10 conversations are among college students who knew each other, totaling about 3.5 hours. Table 1 below, from [5], is a list of these 10 conversations (named M001-M010) with the information about the participants (F = Female, M = Male) and the length of the texts (in minutes, in number of turns, and in number of Chinese characters).

Table1 Details of the first 10 conversations of NCCUCSM

	Participants		Length of Conversation		
	F	M	Minutes	#turns	#characters
M001	2	0	14	422	8688
M002	2	1	24	568	13117
M003	3	0	21	466	8724
M004	3	0	21	405	9869
M005	2	0	40	730	22258
M006	1	1	20	440	9707
M007	2	1	20	386	8463
M008	2	2	20	307	6584
M009	2	1	19	341	8908
M010	2	0	20	394	9163
Total:	21	6	219	4459	105,481

3. A Corpus-based Taxonomy for Chinese NSUs

Our taxonomy of Chinese NSUs is built on the FG taxonomy of English NSUs [1]. However, due to the fact that Chinese is distinctly different from English (see, e.g., [6] and [7]), it is reasonable to expect that our taxonomy of Chinese NSUs contains modifications to the FG taxonomy of English NSUs. The issues here concern the definition of "predicate" in these two different languages. As defined in [3], non-sentential utterances (NSUs) are "utterances without an overt predicate" (p. 2). In contrast to English, in which only verb phrases can act as predicates, there are non-verb predicates, besides verb predicates, in Chinese. Specifically, a Chinese predicate can be either a verb phrase, a noun phrase, an adjective phrase or even a subject-predicate phrase itself (see, e.g., [7]). Consequently, a Chinese utterance containing no verb cannot, by itself, be a sufficient condition for being counted as an NSU, for it may have other kinds of predicates. For example, the following two Chinese utterances are not NSUs as each of them contains an overt (non-verbal) predicate.

A Chinese utterance with a nominal predicate:
我　　　二十二歲
wo3　　　er4shi2er4sui4
1SG　　　twenty-two-years-old
'I was twenty two.'　　　　[NCCUCSM 6: 506]

A Chinese utterance with an adjectival predicate:
二十歲　　　　　很　年輕
er4shi2sui4　　　hen3　nian2qing1
twenty-years-old　very　young
'Twenty years old is very young.'　[NCCUCSM 6: 453]

Here we use "NCCUSM C:L1-L2" or just "C:L1-L2" to indicate that the example is taken from lines L1-L2 of Conversation M00C of NCCUSM. On the other hand, there is a much more critical issue concerning whether Chinese modals are verbs. This issue is still highly controversial, as Huang *et al.* [8] point out, "While the basic distinction between nouns and verbs is universally recognized in modern literature on Chinese syntax, scholars differ, sometimes drastically, on other categories" (p. 9). In particular, there is still no consensus on whether Chinese modals are verbs. In the literature, there are three common views on the categorization of Chinese modals: (1) Chinese modals are verbs; (2) Chinese modals are adverbs; and (3) Some Chinese modals are verbs, some are adverbs and some are both [9]. In this research, we follow [8], [10], [11], [12] and others in treating Chinese modals as verbs. Note that this is a critical decision as it leads to the addition of new NSU classes to our taxonomy

which are not in the FG taxonomy. The reason for this addition is that in Chinese, to answer a polar question positively is by way of repeating the verb in the question (or negatively by repeating the verb with a negation adverb). Since we treat Chinese modals as verbs, there are two kinds of verbs: lexical verbs and modal verbs. In case that the verb being repeated is a *lexical* verb, the answer is *not* an NSU because the lexical verb (or the verb plus the negation adverb) itself constitutes a *complete* (verbal) predicate. For example,

A: 你　　還　　　記得　　　　喔
　　ni3　hai2　　ji4de2　　　o
　　2SG　still　　remember　　QST
　　'Do you still remember?'
B: 記得　　　啊
　　ji4de2　　a
　　remember　PRT
　　'I remember.'　　　　[NCCUCSM 5: 210-211]

By contrast, in case that the verb being repeated is a *modal* verb, the answer *is* an NSU because the modal verb, with or without a negation adverb, is an *incomplete* (verbal) predicate. For example,

A: 那個　真的　　會　　　睡著　　　　嗎
　　na4ge　zhen1de　hui4　　shui4zhao2　　ma
　　that　really　　will　　sleep:RESULT　QST
　　'Does it really make people sleepy? '
B: 會
　　hui4
　　right
　　'Right.'　　　　　　[NCCUCSM 6: 109-110]

It therefore follows that not containing a verb can neither be a sufficient nor a necessary condition for a Chinese utterance to be classified as an NSU. Not sufficient, because in Chinese there are non-verbal predicates; not necessary, because an utterance containing a modal verb can also be classified as an NSU. In order to distinguish these verb-containing NSUs from those which contain no verb, two new classes, namely *Verbal Affirmative Answer* and *Verbal Rejection* are added to our taxonomy. Apart from these two, we have also added six more new classes, namely, *Helpful Acknowledgement*, *Helpful Affirmative Answer*, *Correction*, *Noun Phrase*, *Re-Affirmation* and *Interjection*. Empirical evidence, i.e., examples from the corpus, for these NSUs is provided in the following subsections. Table 2 below is our taxonomy of Chinese NSUs. There are totally 23 classes of NSUs which are grouped under four categories: (1) Acknowledgement, (2) Questions, (3) Answers, and (4) Extension Moves. In the following, we describe each NSU class and illustrate it with an example from NCCUSM.

Table 2 A Taxonomy for Chinese NSUs in the NCCU Corpus of Spoken Mandarin (NCCUSM)

	NSU Class
	A. Acknowledgement
1	Plain Acknowledgement
2	Repeated Acknowledgement
3	Helpful Acknowledgement
	B. Questions
4	Classification Ellipsis
5	Sluice
6	Noun Phrase
7	Check Question
8	Filler
	C. Answers
9	Short Answer
10	Affirmative Answer
11	Repeated Affirmative Answer
12	Verbal Affirmative Answer
13	Helpful Affirmative Answer
14	Re-Affirmation
15	Rejection
16	Verbal Rejection
17	Helpful Rejection
18	Correction
19	Interjection
20	Propositional Modifier
	D. Extension Moves
21	Factive Modifier
22	Bare Modifier Phrase
23	Conjunction + Fragment

3.1 Acknowledgement

There are 3 classes of Acknowledgement.

(1) *Plain Acknowledgement* signals understanding or acceptance. E.g.,

A: 沒辦法
   mei2ban4fa3
   NEG:way
   'Nothing we can do.'
B: 對呀
   dui4 ya
   right PRT
   'Right.'   [1: 4-5]

(2) *Repeated Acknowledgement* is an acknowledgement containing verbatim repetition of a previous utterance. E.g.,

A: 所以   是   一個   馬屁 大會
   suo2yi3 shi4 yi2ge ma3pi4 da4hui4
   so   COP one   fawn-conference
   'So, it's a fawning conference.'

B: 對   馬屁大會
   dui4 ma3pi4da4hui4
   yes   fawn-conference
   'Yes. It is a fawning conference.'   [4: 190-191]

(3) *Helpful Acknowledgement* is an acknowledgment providing additional information. E.g.,

A: 像   我們 那個 班級 才 幾 個 而已
   xiang4 wo3men ne4ge ban1ji2 cai2 ji3 ge er2yi3
   like   1PL   that CL class   just few CL only
   'Like our class, there are only a few students.'
B: 對 呀   才 三 個   ei
   dui4 ya   cai2 san1 ge ei
   right PRT   just three CL PRT
   'Right. There're only three students.'   [1: 114-116]

3.2 Questions

There are 5 classes of Questions.

(4) *Classification Ellipsis* is a reprise fragment used to clarify not fully comprehended utterance. E.g.,

A: 那 個 要   涮 .. 涮   就 好   了
   na4 ge yao4   shan4 shan4 jiu4 hao3 le
   that CL have-to boil   boil   then OK PRT
   'That one has to be boiled. Just boil it.'
B: 涮   喔
   shan4 o
   boil   PRT
   'Boil?'   [5: 197-198]

(5) *Sluice* is a bare question-denoting *wh*-phrase. E.g.,

A: 我 就 共 你 偷   囥
   gua2 to7 ka7 li2 thau1 khng3
   1SG will give 2SG stealthily put
   'I will stealthily put things into the pot.'
B: 啥   物
   siann2 mih1
   what
   'What are you putting?'   [5: 2-3]

(6) *Noun Phrase* is a question containing only noun phrase(s). E.g.,

A: 看 到   兩 個 學生   手   牽 手
   kan4 dao4 liang3 ge xue2sheng1 shou3 qian1 shou3
   see:RESULT two CL student   hand hold hand
   'I saw two students in my class walking hand in hand.'
B: 男   的 女 的
   nan2 de nv3 de
   boy   girl

'Boys or girls?' [1: 242-243]

(7) *Check Question* requests explicit feedback. E.g.,

A: 我 也 覺 得 這 樣 靠 著 很 舒 服
wo2 ye3 jue2 de zhe4 yang4 kao4 zhe hen3 shu1 fu2
1SG also think like-this lean PROG very comfortable
B: 真的 喔
zhen1de o
really PRT
'Really?' [9: 255-256]

(8) *Filler* completes an unfinished utterance. E.g.,

A: 可 是 很 多 家 丹 堤 的 都 是
ke3 shi4 hen3 duo1 jia1 dan1ti2 de dou1 shi4
but many CL Dan-Ti ASSC all COP
'But a lot of the Dan-Ti shops are all'
B: 很 暗 hon
hen3 an4 hon
very dim PRT
'very dim.' [2: 237-238]

3.3 Answers

There are 12 classes of Answers.

(9) *Short Answer* is an answer-conveying NSU to a *wh*-question. E.g.,

A: 附 中 現 在 第幾 志願
fu4zhong1 xian4zai4 di4ji3 zhi4yuan4
affiliated-high-school now what-ranking choice
'What is the ranking of the Affiliated High School?'
B: 第 三
di4 san1
third
'It's the third of the top schools.' [1: 366-367]

(10) *Affirmative Answer* is a positive answer to a polar query. E.g.,

A: 他們 覺 得 那個 漂 亮 喔
ta1men jue2 de ne4 ge piao4 liang4 o
3PL think that CL beautiful QST
'Do they think that she's beautiful?'
B: 對
dui4
right
'Right.' [2: 350-351]

(11) *Repeated Affirmative Answer* is an affirmative response with repetition. E.g.,

A: 是 這 張
shi4 zhe4 zhang1
COP this CL
'This one?'
B: 這 張
zhe4 zhang1
this CL
'This one.' [2: 28-29]

(12) *Verbal Affirmative Answer* is an affirmative response which contains just a modal verb. E.g.,

A: 那個 真的 會 睡著 嗎
na4ge zhen1de hui4 shui4zhao2 ma
that really will sleep:RESULT QST
'Does it really make people sleepy?'
B: 會
hui4
right
'Right.' [6: 109-110]

(13) *Helpful Affirmative Answer* is an affirmative response providing additional information. E.g.,

A: 你們 實習 老師 也 會 化
ni3men shi2xi2lao3shi1 ye3 hui4 hua4
2PL intern-teacher also will draw
'You intern teachers will also use it?'
B: 很多
hen3duo1
very
'Many of them.' [2: 422-423]

(14) *Re-Affirmation* affirms the truth of a previous utterance. E.g.,

A: 我 不 太 記得 了
wo3 bu2 tai4 ji4de2 le
1SG NEG quite remember PRF
'I don't quite remember it.'
B: 你 少 裝
ni2 shao3 zhuang1
2SG NEG pretend
'Don't pretend that you don't remember.'
A: 真的
zhen1de
really
'Really.' [10: 3-5]

(15) *Rejection* is a negative answer to a polar query or an assertion. E.g.,

A: 你 有 加重 他們 的 負擔 嗎
ni2 you3 jia1zhong4 ta1men de fu4dan4 ma

2SG have increase   3PL POSS burden QST
'Did you increase their burden?'
B: 沒有
mei2you3
NEG
'No.'   [10: 55-56]

(16) *Verbal Rejection* is a negative answer to a polar query or an assertion which contains just a modal verb with a negation adverb. E.g.,

A: 會 不 會 是   你   老 母
hui4bu2hui4   shi4   ni2   lau7bu2
will:NEG:will COP   2SG   mother
'Could it have been your mother?'
B: 不   會   吧
bu2 hui4   ba
NEG will   PRT
'No.'   [4: 345-346]

(17) *Helpful Rejection* is a negative answer with an alternative to a polar query or an assertion. E.g.,

A: 你們 只 有 一 次 總 複習 喔
ni3men zhi3 you3 yi2 ci4 zong3fu4xi2   o
2PL   only have one time overall-review PRT
'You only do an overall review once, right?'
B: 兩   次 啦 兩   次
liang3 ci4 la liang3 ci4
two   time PRT two   time
'Twice. Twice.'   [6: 62-63]

(18) *Correction* is the corrected version of the faulty part of a previous utterance. E.g.,

A: 他 不怒 不怒 自 [威]
ta1 bu2nu4 bu2nu4 zi4 wei1
3SG REPAIR NEG:angry:SELF:majestic
'He's not angry...charismatic without being angry.'
B: [而威   ]
er2wei1
but-majestic
'but charismatic.'   [1: 90-91]

(19) *Interjection* is an abrupt expression for sudden sensations and emotions. E.g.,

A: 就   沒辦法   啊
jiu4 mei2ban4fa3   a
PRT NEG:way   PRT
'Nothing I can do.'
B: 喔
o
BC

'Oh.'   [1: 41-42]

(20) *Propositional Modifier* is a propositional adverb that can convey a complete message in the context. E.g.,

A: 喔 應該   都 有   問題   啦
o ying1gai1 dou1 you3 wen4ti2   la
PRT should   all have problem   PRT
'They should have problems.'
B: 可能   吧
ke3neng2 ba
probably   PRT
'Probably.'   [4: 298-301]

3.4 Extension Moves

There are 3 classes of Extension Moves.

(21) *Factive Modifier* is an evaluative adjective. E.g.,

A: 然後   NB 可以 帶 回   家
ran2hou4 NB ke2yi3 dai4 hui2 jia1
then notebook can take back home
'They can take the laptop home with them.'
B: 好好
hao2hao3
good
'That's really good.'   [3: 308-309]

(22) *Bare Modifier Phrase* is an adjunct that can modify a previous utterance in the context. E.g.,

A: 我 現在   別人 大聲 我 會 比
他 更 大聲   的 那 一 種
wo3 xian4zai4 bie2ren2 da4sheng1 wo3 hui4 bi3
ta1 geng4 da4sheng1   de na4 yi4 zhong3
1SG now other-person loud 1SG will COMPARE
3SG even loud   ASSC that one kind
'Now, I am the kind of person who, when other person is loud, will be even louder.'
B: 人生   磨練   之後
ren2sheng1 mo2lian4 zhi1hou4
life   train   after
'After training in life.'   [6: 546-547]

(23) *Conjunction + Fragment* consists of a connective which introduces a fragment. E.g.,

A: 蠻   瘦   的
man2 shou4   de
quite thin   PRT
'He's quite thin.'
B: 而且 高高   的
er2qie3 gao1gao1   de

and   tall   PRT
'And he's tall.'   [2: 53-54]

4. Conclusions and Future Work

This paper describes a preliminary investigation into Chinese non-sentential utterances (NSUs) in the NCCU Corpus of Spoken Mandarin. In particular, it presents, with examples from the corpus, a corpus-based taxonomy of Chinese NSUs. Although this taxonomy is built on the FJ taxonomy - the taxonomy by Fernández and Ginzburg for English NSUs in the British National Corpus (BNC), several new classes are added. The reasons for the addition have been explained by appealing, in part, to the distinctiveness of spoken Chinese. In the future, we plan to test this taxonomy with other Chinese corpora and on other languages. A recent work in this direction is by Floriane Guida who tests the FG taxonomy on French [13]. Another future work is to use this taxonomy as the basis for an interpretation module of a dialogue system building on, e.g., GoDiS [14], [15].

5. References

[1] Fernández, R., and Ginzburg, J., "Non-sentential utterances: a corpus study," *Traitement automatique des languages. Dialogue*, 43(2), 2002, pp. 13-42.

[2] Fernández, R., *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College, London, 2006,

[3] Ginzburg, J., *The Interactive Stance: Meaning for Conversation*, Oxford : Oxford University Press, 2012.

[4] Schlangen, D., A *Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D thesis, University of Edinburgh, 2003.

[5] Chui, K., and Lai, H., "The NCCU Corpus of spoken Chinese: Mandarin, Hakka, and Southern Min," *Taiwan Journal of Linguistics*, 6(2), 2008, pp.119-144.

[6] Xiao, R., and McEnery, T., *Corpus-based Contrastive Studies of English and Chinese*, New York: Routledge, 2010.

[7] Pan, W., and Tham, W. M., *Contrastive Linguistics: History, Philosophy and Methodology*, London; New York: Continuum, 2007.

[8] Huang, C.-T. James, Li, Y.-H. Audrey, and Li, Yafei, *The Syntax of Chinese*, New York: Cambridge University Press, 2009.

[9] Song, Y. K., *A Study on the Negation of Modals in Modern Chinese* (in Chinese), China Social Sciences Press, 2007.

[10] Lin, J. and Tang, C. Jane, "Modals as verbs in Chinese: a GB perspective," *Bulletin of the Institute of History and Philology, Academia Sinica,* 66(1), 1995, pp. 53-105.

[11] Xiang, E., *A Syntactic Comparison of Chinese and English Auxiliary Verbs* (in Chinese), Ph.D thesis, Huazhong Normal University, 2011.

[12] McCawley, J. D., "Justifying part-of-speech assignments in Mandarin Chinese," *Journal of Chinese Linguistics*, 20, 1992, pp. 211-45.

[13] Guida, F., "Les phrases sans verbes," Ms., Université Paris-Diderot, 2013.

[14] Larson, S., Ljunglöf, P., Cooper, R., Engdahl, E., and Ericsson S., "GoDiS - An accommodating dialogue system," *Proceedings of ANLP/NAACL2000 Workshop on Conversational Systems*, 2000, pp. 7-10.

[15] Larsson, S., *Issue-based Dialogue Management*, Ph.D. thesis, Göteborg University, 2002.