A Comparative Study of Subject Pro-drop in Old Chinese and Modern Chinese

Zhiyi Song

1 Introduction

Chinese is a Subject Pro-drop language in that the subject of a clause need not be overt. Thus a Chinese speaker has the choice of using either a null subject or an overt pronoun in the subject position of a sentence, as in

ta kanjian yige nuhaizi, Ø/ta daizhe yiding xiaohongmao.
he see one-classifier girl, Ø/she wear one-classifier small red hat.
'He saw a girl; she is wearing a red hat.'

Chinese differs from other Pro-drop languages such as Italian or Turkish in that the language has no inflections to mark subject-verb agreement. Huang (1984, 1989) argued from a syntactic perspective, that for languages like Chinese, a null subject is identified by an NP in the superordinate clause, due to the lack of Agr. Tsao (1979) and Li (1981) observed that subject Pro-drop in Chinese was actually Topic-NP deletion, which is an optional process, alternating with the use of overt pronoun in the subject position. Li (1985) and Chen (1986) proposed from a discourse perspective that null subject in Chinese is more likely to occur in cases of topic continuity, in which the information represented by the subject is the component of a series of related actions, events or states. Crosslinguistically, DiEugenio (1998) and Prince (1999), in their studies of Italian and Yiddish respectively, tried to address subject pro-drop in terms of Centering Theory (Grosz, Joshi, and Weinstein 1995).

All previous studies are based on Modern Chinese (MC), which has a significant amount of increase in the use of pronominal and nominal anaphors, as compared to Old Chinese (OC), where subject pro-drop is more frequent, as can be seen in the two following parallel texts:

OC	MC
jian yuren	*tamen kandao yuren*
Ø see fishman	'**they** see-perf fisherman'
'they saw the fisherman'	
nai dajing	*juede shifen yiwai*
Ø be surprised	'Ø feel very surpised'
'they were very surprised'	
wen suo cong lai	*wen ta cong na lai*
Ø Ask Ø from where come	'Ø ask **him** from where come'
'they asked him where he came from'	*yuren yiyi zuo le huida*
ju da zhi	'**fisherman** one-one gave-le answer'
Ø all answer them	
'the fisherman answer all the questions'	*jiu you ren yaoqing yuren dao jiali qu*
bian yao huan jia	'then have **someone** invite **fisherman** go home to'
then Ø invite Ø back home	
'then they invited him home'	

In this light, this paper poses two questions on the subject pro-drop phenomena of OC and MC: 1) Can we explain the subject Pro-drop of both OC and MC in terms of Centering Theory? 2) If so, then in which way does the subject Pro-drop in OC differ from that in MC?

2 Centering Theory

Centering Theory efficiently captures conversation participants' attentional state and hence is a main component of local discourse coherence. Thus, reference tracking and pronoun resolution have been areas of active investigation under the framework of Centering Theory. There are two types of centers: Forward-looking Centers (henceforth, Cf) which are a partially ordered set of discourse entities that each utterance evokes and Backward-looking centers (henceforth, Cb) which are the links from the current utterance to the previous utterance. Constraints with regard to the centers are summarized in Prince and Walker (1996) as follows:

For each utterance U_i in a discourse segment $U_1 \ldots U_m$:

a. There is at most one Backward-looking Center, Cb.
b. Every element of the Forward-looking centers list of U_i, $\{Cf(U_i)\}$, must be realized (explicitly or implicitly) in U_i.

c. The Backward-looking center of U_i, $Cb(U_i)$, is the highest-ranked element of $\{Cf(U_{i-1})\}$ that is realized in U_i.
d. The highest-ranked element of $\{Cf(U_i)\}$ is Preferred Center or Cp of U_i.

The version of Centering Theory that I adopt is that of Grosz, Joshi, and Weinstein (1995), Prince (1999), and Walker, Joshi and Prince (1998). The typology of transitions is based on two factors:

1. Whether the Cb is the same from previous utterance to current utterance, namely Cb(Ui) = Cb(Ui-1),
2. Whether this discourse entity is the same as the CP of the current utterance, namely Cb(Ui) = Cp(Ui).

The algorithm for Centering transitions that applies here is demonstrated in Table 1:

	Cb(Ui) = Cb(Ui-1)	Cb(Ui) ≠ Cb(Ui-1)
Cb(Ui) = Cp(Ui)	Continue	Smooth-shift
Cb(Ui) ≠ Cp(Ui)	Retain	Rough-shift

Table 1: Algorithm for Centering Theory transitions

3 The Corpus and Coding

The corpus for this study includes 16 writings from Gu Wen Guan Zhi, composed by Wu Diaohou and Wu Chucai in Qing Dynasty, with the Modern Chinese parallels provided by Zang Hanzhi. The writings, however, date from 770 BC to 900 AD. For this study, five variables which are closely related to Centering Theory for each clause were coded: Subject type (full NP, pronouns, null subject), whether or not the Cb of the clause is the Cp of previous clause, the syntactic position of the Cb (subject, object, possessive), whether the subject is the global focus[1] of the writing and Centering transition state (continuation, retaining, smooth shift, rough shift), with the subject type as the dependent variable while others as independent variables.

It has been discussed that the usual ordering for the Forward Looking Center (Cf) for western languages is:

SUBJECT > OBJECT2 > OBJECT > OTHERS

[1]I am very grateful to Ellen Prince, Uri Horesh, Jinyoung Choi for their aid and encouragement of this paper.

In their study of Japanese discourse, Walker, Iida and Cote (1994) proposed that discourse topic is more salient and should be ranked higher on the Cf. Chinese is a topic comment language (Li 1981), hence the topics should be ranked higher than the grammatical subjects in the Cf list:

TOPIC > SUBJECT > OBJECT2 > OBJECT > OTHERS

But the ranking that I employ in this study is the former one, and it is due to the following reasons. First, 'topic' is a very ambiguous term, and it is unclear whether topic in Chinese is the left dislocated entity in a clause or any canonical entity; second, there is no comparable morphological marker, like Japanese *–wa*, to mark topic in Chinese, and it is therefore hard to judge whether the canonical entity of a clause is subject or topic; third, Chinese not only drops subjects, but also other syntactical components, like objects, possessives, preposition phrases, etc., which makes it more vague to decide whether the dropped component in canonical position is the subject or something else.

Another coding concern is related to the problem of segmentation. In Grosz, Joshi, and Weinstein (1995), centering is a local mechanism that is strictly restricted in discourse segments. However, Walker (1996) argues that such restrictions pose problems as centers are clearly carried over segment boundaries and proposed to integrate centering with the cache model of attentional state. In this study, I do not segment the writings but rather assume that each writing is one flat discourse. Therefore, if there is no Cb in Ui, the transition in Ui is coded as Rough-shift and that of Ui+1 is coded as Continue, the initial clause of each writing however, is excluded from the analysis.

4 Findings

After the exclusion, the corpus consisted of a total of 407 main clauses in OC and 385 in MC. Diachronically, MC subjects are more likely to be full NPs while OC subjects are more likely to be dropped: OC drops 59% of the subjects while in MC only 44% of the subjects are zero. Interestingly, in both OC and MC, there were very few tokens of overt pronominal subjects, with 7 in OC and 22 in MC; the causes are not the focus of this paper, but will be worthwhile for future study. In the Varbrul analysis, I have excluded all the pronoun tokens for Varbrul analysis, as the occurrences are rare and combining them with either zeros or full NPs or excluding them would not make much difference in results. Table 2 presents the counts and frequencies

of the subject type in OC and MC. The chi-square test shows that the difference between OC and MC in subject type is significant.

	OC	MC
NP	0.39/158	0.51/208
Zero	0.59/242	0.44/177
Pronoun	0.02/7	0.05/22

Table 2: Frequencies of Subject Type in Old and Middle Chinese
$x^2=24.7$, $p \leq 0.001$

variables	factors	weight for OC	weight for MC
1. Cb(Ui) = Cp(Ui-1)	Cb is CP of u-1	**0.665**	**0.673**
	Cb isn't CP of u-1	0.229	0.188
2. Transition state	continuation	**0.711**	**0.736**
	retaining	0.111	0.177
	smooth shift	**0.594**	**0.601**
	rough shift	0.367	0.356
3. Gram. Status of Cb	subject	**0.598**	**0.612**
	object	0.274	0.218
	possessive	0.124	0.237
4. Focus	global focus	**0.625**	n.s.
	non-global focus	0.401	n.s.

Table 3: Varbul results for all factors in Old and Middle Chinese

Varbrul analyses of the coded data reveal some interesting findings. Most strikingly, for OC, all four independent variables are significant, while for MC, whether the subject is the global focus of the writing is not significant for application (it will be further discussed in Section 4). For the first variable, if Cb(Ui) = Cp(Ui-1) holds, for both OC and MC, the subject is more likely to be a zero anaphor, while when it does not hold, the subject tends to be a full NP. As for the transition states, Continue and Smooth-shift transitions favor null subject while Rough-shift and Retain transitions disfavor it, i.e. when the subject is also the Cb of the current clause, it is more likely to be dropped, but not if the Cb is the object or possessive. The transition state ordering is different from the rule that is proved by empirical work

of other languages (Walker, Joshi, and Prince 1998). The Continue transition is preferred to the Smooth-shift transition, which is preferred to the Rough-shift transition, which is preferred to the Retain transition, which surprisingly, is least likely to favor null subject, even less likely than Rough-shift transition. The possible reason will be discussed further in Section 4. In OC, if the subject is the global focus of the discourse, it favors zero anaphor in subject position, with 76% of the global focus tokens being dropped. But in MC, only 55% of them are dropped. The figures for the effects of all factors are presented in Table 3.The highlighted figures show the factors which favor null subject.

		OC		MC		difference between OC and MC
		Zero	NP	Zero	NP	
Total	N	242	158	177	208	
	%	60	39	45	54	
CbUi = CpUi-1	N	49	87	18	109	**p ≤ 0.0001**
	%	36	63	14	85	
CbUi ≠CpUi-1	N	193	71	159	99	**p ≤ 0.001**
	%	73	26	61	38	
Continue	N	158	18	131	33	**p ≤ 0.01**
	%	89	10	79	20	
Rough-shift	N	28	65	8	86	**p ≤ 0.0001**
	%	30	69	8	91	
Smooth-shift	N	48	19	33	32	**p ≤ 0.01**
	%	71	28	50	49	
Retain	N	8	56	5	57	p = 0.41
	%	12	87	8	91	
Subject	N	221	75	169	114	**p ≤ 0.0001**
	%	74	25	59	40	
Object	N	20	68	7	79	**p ≤ 0.01**
	%	22	77	8	91	
Possessive	N	1	15	1	15	p = 1
	%	6	93	6	93	
Non-global	N	111	118	87	136	p = 0.043
	%	48	51	39	60	
Global	N	131	40	90	72	**p ≤ 0.0001**
	%	**76**	23	**55**	44	

Table 4: Differences between Old and Middle Chinese for all factors

Even though OC and MC overall show the same tendency in alternation between zero anaphors and full NP in subject position, there is slight difference in each factor of the variables. Table 4 presents the figures for all variables collapsed in factors and both subject types collapsed. Except for factors of Retain transition, Cb being possessive and subject being non-global focus, the difference between OC and MC in all other factors are significant, with OC having higher frequencies of using zero anaphor rather than full NPs in subject position.

5 Discussion

The two puzzles are: First, why is the ordering of transition states in OC Continue > Smooth-shift > Rough-shift > Retain rather than the usual ranking Continue > Retain > Smooth-shift > Rough-shift? And secondly, why is the global focus factor significant for subjects in OC, but not in MC?

Grosz, Joshi, and Weinstein (1986) proposed that a Retain actually signals that the speaker is intending to shift onto a new entity in the next utterance and hence the current center is realized in a lower ranked position on the Cfs. The corpus for this study has altogether 64 tokens of Retain transition, only 8 tokens of which are null subject, because the subject in Retain transition state is not an entity from the previous utterance, but a new entity (new to the previous utterance, not necessarily to the discourse) introduced to the current utterance. In Smooth-shift transition, however, the subject is the Cb, which is one of the entities realized in the previous utterance and hence is more likely to be dropped

As shown in Table 5, in both OC and MC, if the transition is Continue, the subjects are more likely to be zero; if the transition is Retain, the subjects are unlikely to be zero. When the transition is Smooth-shift or Rough-shift, OC tends to have zero anaphors more often than MC does in that OC drops 51% while MC only drops 17%. The difference between OC and MC for these two transitions is significant with a chi-square of 18.4 and p-value of 0.001. A possible explanation is that OC and MC treat global focus differently and that in OC, global focus is treated as discourse old information, and thus can be dropped even though it is not present in Cfs of the previous utterance. In MC, however, a subject is more likely to be treated as discourse-new information if it is not present in the previous utterance even if it is the global focus.

As shown in the Varbrul analysis, variable 5, namely whether the subject is the global focus, functions differently in OC and MC. In OC, the subject being the global focus of the writing favors drop, but in MC, this is not a significant factor for subject being zero anaphor. Table 5 is the cross tabula-

tion of subject type and Centering transition states for tokens of global subject.

	OC		MC		
	Zero	NP	Zero	NP	
Continue	97	6	78	16	p = 0.01
Retain	3	10	2	11	p = 0.62
Smooth-shift	18	6	8	15	p = 0.006
Rough-shift	14	18	2	31	p = 0.0004
sum	132	40	90	73	p = 3E-05

Table 5: Cross tabulation of subject type and transition state

The relatively high frequencies of null subject in Rough-shift transitions is actually correlated with variable 5. When the null subject tokens in Rough-shift transitions are collapsed into variable 5, in OC, half of the 28 zero tokens in Rough-shift transitions are global focuses. As discussed before global focuses are discourse-old information, hence the global focus tokens are excluded from the Rough-shift transition in OC, but not in MC. Then there are only 14 null subject tokens remaining in OC. The difference between OC and MC in Rough-shift transitions therefore is not significant, as shown in Table 6:

	OC	MC
zero	14	8
NP	65	86

Table 6: Difference between Old and Middle Chinese in Rough-shift transistions

$x^2=3.28$, $p \leq 0.1$

Therefore, Rough-shift transition can not rank higher than Retain transition in the transition states ordering in OC if we exclude the global focus tokens from consideration.

Interestingly, among the 14 tokens which are zero in OC for Rough-shift transition as shown in table 5, two are zero in MC and 11 are presented as overt pronouns in MC. This, on one hand, proves that global focus should be treated as discourse-old information in OC, on the other hand, shows that MC tends to be more discourse coherent in that zero pronouns are so rare in Rough-shift transition.

6 Conclusion

The analysis results show that Subject Pro-drop in both OC and MC was constrained by Centering Theory in that Continue and Smooth-shift transitions favor null subject while Rough-shift and Retain transitions disfavor it. OC had a higher rate of null subject than MC in terms of all the variables that were considered in this paper. These results also seem to exhibit a different hierarchical transition order of preference for subject Pro-drop, whereby Smooth-shift was more likely to favor null subject than Retain, because the subjects in Smooth-shift transition are more likely to be discourse-old, while those in Retain transition are more likely to be new entities introduced into the discourse. Null subject is much more likely in OC when the subject is a global focus. This corpus study therefore provides evidence that focus type should be taken into account for discourse study of OC.

References

Chen, Ping. 1986. Referent introducing and tracking in Chinese narrative. Doctoral dissertation, University of California, Los Angeles.

Di Eugenio, B. 1998. Centering in Italian. In *Centering in discourse*, ed. Marilyn Walker, Aravind K. Joshi, and Ellen Prince, 115-38. Oxford: Oxford University Press.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1986. Towards a computational theory of discourse interpretation. Unpublished manuscript.

Grosz, Barbara J., Aravind K. Joshi and Scott Weinstein. 1995. Towards a computational theory of discourse interpretation. *Computational Linguistics* 21:203-225.

Huang, C-T. James. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry* 15: 531-574.

Huang, C-T. James. 1989. Pro-drop in Chinese: a generalized control theory, In *The null subject parameter*, ed. Osvaldo Jaeggli and Ken Safir, 185-214. Boston: Kluwer Academic Publishers.

Li, Charles N. 1981. *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

Li, Cherry Ing. 1985. Participant anaphora in Mandarin Chinese. Doctoral dissertation, University of Florida.

Prince, Ellen F. 1999. Subject pro-drop in Yiddish. In *Focus: Linguistic, cognitive, and computational perspectives (studies in natural language processing)*, ed. Peter Bosch and Rob van der Sandt. Cambridge: Cambridge University Press.

Tsao, Feng-fu. 1979. *A functional study of topic in Chinese*. Taipei: Student Book Co.

Walker, Marilyn A., Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics* 20(2):193-232.

Walker, Marilyn A., Aravind K. Joshi, and Ellen F. Prince. 1998. Centering in naturally-occurring discourse: An overview. In *Centering in discourse*, ed. Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince. Oxford University Press.

Department of Linguistics
619 Williams Hall
University of Pennsylvania
Philadelphia, PA 19104-6305
zhiyi@babel.ling.upenn.edu