Giorgio Iemmolo and Giorgio Francesco Arcodia

Differential Object Marking and identifiability of the referent: A study of Mandarin Chinese

Abstract: This paper examines the interaction of DOM with information structure in Mandarin Chinese. Despite the large amount of works on this topic, much remains to be explained, in particular with respect to some alternations that do not easily fit the explanations proposed so far in terms of affectedness, animacy and definiteness. Through the analysis of text excerpts taken from the Corpus of Modern Chinese of the Center for Chinese Linguistics (CCL) of Peking University, we argue that, in addition to previously identified constraints, DOM in Mandarin Chinese performs another important function in discourse, namely that of signalling the high identifiability of the marked referent.

Keywords: Differential Object Marking, Mandarin Chinese, Cantonese, topic, information structure

Giorgio Iemmolo: Seminar für Allgemeine Sprachwissenschaft, Universität Zürich, Plattenstrasse 54, CH-8032 Zürich. E-mail: giorgio.iemmolo@uzh.ch
Giorgio Francesco Arcodia: Facoltá di Scienze della Formazione, U6, Università degli Studi di Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano. E-mail: giorgio.arcodia@unimib.it

1 Introduction*

This paper examines Differential Object Marking (henceforth, DOM) in Mandarin Chinese. As is well known, DOM is quite common crosslinguistically. In many languages some direct objects (DOs) are marked overtly, while others are not. In the literature, the presence of DOM is usually ascribed to the animacy and/or definiteness of the direct object referent. The more animate or definite a direct

---

* Although this work is the outcome of a joint project, Sections 1 and 2 were written by Giorgio Iemmolo, Sections 3, 4 and 5 by Giorgio Francesco Arcodia. The glosses follow the general guidelines of the Leipzig Glossing Rules.

object is, the more likely it is to be overtly marked (cf. Aissen 2003; Bossong 1985; Comrie 1979, 1989).

Although DOM in Chinese has been extensively studied both from a synchronic and from a diachronic point of view, much remains to be explained, particularly with respect to the issue of the interaction of the phenomenon with information structure. Mandarin Chinese is an SVO language. DOM appears when the DO is human/animate or definite/generic (Li and Thompson 1981; Yang 2008), and causes the direct object to be moved to the preverbal position[1]. DOs high on the animacy or definiteness hierarchy are nearly always marked by means of the preposition *bǎ* (or by the preposition *jiāng*, mainly in the written language)[2], as in (1), whereas for DOs characterised by low animacy or definiteness differential marking is optional (2). Moreover, when a DO is marked, it is moved to the preverbal position:

(1) *Lǐ Yǒu dāngchǎng bǎ tā sha-si*
Li You on.the.spot BA 3SG.M kill-dead
'Li You killed him on the spot'

(2) *wǒ dǎ-sùi-le bōli-bēi*
1SG hit-broken-PFV glass-cup
'I smashed the glass'

According to various studies, DOM-sentences express disposal, i.e., how the entity denoted by the DO object is affected by the subject (Li and Thompson 1981: 465), and are influenced by the event boundedness (Liu 1997). Disposal analyses have been proposed within the Transitivity theory framework (Hopper and Thompson 1980), in which *bǎ* is analyzed as a marker of high transitivity. However, examples may be found in which transitivity seems to play no role in the use of the marker, as in (3), where the predicate is a psych verb:

---

1 Our analysis is restricted only to cases in which *bǎ* marks direct objects. As a matter of fact, *bǎ* can be used to introduce an indirect object, an instrumental, a locative NP, a possessor of an object NP (often in a part-whole relationship with it) and finally the subject of a clause expressing the result of an action (Yang 2008: 68).

2 The status of *bǎ* is quite controversial in the literature. Whilst most linguists agree that *bǎ* is synchronically a preposition (Chao 1968; Yang and van Bergen 2007; Yang 2008, among others), many studies have proposed an analysis of *bǎ* as a verb (Bender 2000), as a dummy case that fills the head of a causative phrase when there is not verb raising (Sybesma 1992), or also as a coverb (Li and Thompson 1981). In this paper, *bǎ* will be considered as a preposition, due to the fact that *bǎ* does not behave as a verb, as demonstrated by Sun (1996) and discussed by Li (2006).

(3) *wǒ   bǎ   tā   de   míngzi wàngjì-le*
     1SG BA 3SG.F DET name   forget-PFV
     'I forgot her name'

Other current approaches, mainly within the framework of *Optimality Theory*, consider *bǎ*-marking of human and animate DOs as motivated by the need for distinguishing between subjects and DOs, while the optional marking of definite objects is motivated by the unmarkedness of these DOs in the preverbal position, usually occupied by definite referents (Yang 2008; Yang and van Bergen 2007). While the distinguishing function might be one of the factors triggering DOM, again we have examples in which such an explanation does not fit the data. For instance, *bǎ*-marking is found in imperative constructions, in which the need for disambiguation is not so crucial, given that the imperative forms are distinguished both formally and in the intonation, by means of elements such as the particle *bié* in negative imperatives:

(4) *bié   bǎ tā   è-sǐ*
     IMP.NEG BA 3SG.M starve-die
     'Don't starve him'

In addition, diverse approaches view Chinese DOM as related to information structure. For example, according to Tsao (1987), DOs marked by the preposition *bǎ* are instances of secondary topics. More recently, Liu (2007) has demonstrated that the variation in DOM is mainly conditioned by the information status of the *bǎ*-marked NP. In her view, DOM is likely to appear either with NPs carrying old information or with heavy NPs carrying new information (Liu 2007). Liu's approach presents some theoretical problems, insofar as it is highly unlikely that the same marker signals both old and new information. We will discuss this in Section 4.

    In the present paper, we would like to further investigate the pragmatics of DOM marking in Chinese. We try to demonstrate that DOM in Mandarin Chinese has another important function: in addition to the aspectual constraints, we argue that *bǎ*-marking signals the identifiability of referents in discourse.

    Through the examination of text excerpts, we show that marked DOs are readily identifiable by the hearer/reader, because they are accessible (in terms of degrees of identifiability, as put forward by Lambrecht 1994) either from the physical/textual context or from various available frames (e.g., discourse or lexically-evoked frames, Chafe 1987; Givón 1990).

2 Preliminaries

2.1 Chinese DOM

As mentioned earlier, DOM in Mandarin Chinese seems to be governed by the two dimensions of animacy and definiteness. Moreover, other semantic and syntactic conditions must be satisfied in order for DOM to occur without leading to ungrammaticality.

The first required condition is the boundedness of the event represented by the verb governing the *bǎ* NP. According to Liu (1997), *bǎ*-marking is used only when the event expressed by the sentence is bounded, as shown by the ungrammaticality of Example (5), in which the event expressed by the verb is not yet completed:

(5) **wo   zai   ba wo de   lunwen   xie*
    1SG PROG BA 1SG DET dissertation write
    int: 'I'm writing my dissertation' (Liu 2007; glosses adapted)

With respect to the animacy parameter, DOM is compulsory when the DO is human or animate, as shown in (6), while the marking is optional when the DO is low in the animacy hierarchy, as in (7) and (8), as mentioned before (Yang and van Bergen 2007: 1622; glosses adapted):

(6) *Ta   *(ba) laoshi   tuidao   le*
    3SG.M   BA   teacher push.over PRF
    'He pushed over the teacher'

(7) *Ta   (ba) pingguo chi   le*
    3SG.M BA   apple   eat   PRF
    'He ate the apple/apples'

(8) *Wo (ba) na-ge   qiu   fang jin   lanzi   li   le*
    1SG BA   that-CLF ball put   into basket inside PRF
    'I put that ball into the basket'

However, some inanimate DOs are obligatorily marked, if they are indefinite or at least non-specific, as in (9) and (10):

(9) *Ta   *(ba) yi-ge   pingguo chi le*
    3SG.M   BA   one-CLF apple   eat PRF
    'He ate an apple'

(10) *Ta   *(ba) yi-pen   yifu   xi   le*
    3SG.M   BA   one-CLF   clothes   wash   PRF
    'He washed a basin of clothes'

Definiteness closely interacts with animacy in determining the marking of DOs. *bǎ*-marking is commonly used with definite or generic NPs, as in (7) and (8). However, as observed by Yang and van Bergen (2007: 1626), definiteness in Mandarin Chinese DOM plays a different role than in other DOM systems. In fact, the marker is optional when the DO is definite or specific, as in (7) and (8), whereas it becomes obligatory when the DO is indefinite or non-specific, as shown in (9) and (10).

According to Yang and van Bergen (2007), such seemingly anomalous behavior is easily explained if we take into account that the preverbal position (recall that marked DOs must be moved to preverbal position) requires its argument to be definite. Thus, if a DO in this position is already definite, the marking is optional. On the contrary, when an indefinite or non-specific DO is found in the preverbal position, there is a mismatch between the syntactic position, which requires its argument to be definite, and the NP properties. To put it another way, definite DOs in Mandarin Chinese are optionally marked because they are not marked with respect to their preverbal position. Conversely, indefinite or non-specific DOs obligatorily get the case marker.

However, this explanation does not account for the cases in which DOs with the same semantic features (i.e., definite or specific) can be either marked or unmarked. We show that such cases can be explained by reference to the pragmatic status of the DO in the discourse, in particular identifiability.

2.2 Identifiability in Chinese

Following Lambrecht (1994; see also Chafe 1976, 1987, 1994), we assume that all discourse referents (syntactically expressed as arguments)[3] within a discourse can be classified as identifiable or unidentifiable. Identifiability can be defined as "a speaker's assessment of whether a discourse representation of a particular referent is already stored in the hearer's mind or not" (Lambrecht 1994: 76). Thus, a referent is identifiable when both the speaker and the addressee have a mental

---

3 Arguments in this sense include here also adjuncts, various kinds of subordinate clauses and adverbial phrases (Lambrecht 1994: 75).

representation of that referent, whereas a referent is unidentifiable when only the speaker has a representation of it.

There are different resources according to which the speaker can consider a referent as identifiable by the hearer (Lambrecht 1994: 88–89). First of all, a referent can be "inherently" identifiable because there is only one referent that can be referred to with that expression. This is the case of proper names, NPs like *the moon*, *the Queen*, *dad*, and generic NPs where the entire class of all of the entities can be referred to with a NP that identifies only one referent, such as *the kids* in the sphere of family. Referents may also be made identifiable de-ictically or anaphorically (Lambrecht 1994: 88–89). In the case of deictic reference, the referent is identifiable because it is spatially or temporally close to the speech context, or it is highly salient within the discourse. In case of anaphoric reference, the referent is identifiable because it was previously mentioned in the discourse. Usually, once a referent has been introduced into the discourse, it is referred to anaphorically with a pronoun or a definite NP. Finally, identifiability can be established indirectly through the mention of another entity. When an entity is introduced into the discourse, a frame of identifiable referents is activated. For example, if a war is introduced into the discourse, then the speaker can easily refer to weapons, enemies, etc. as identifiable referents. Likewise, inalienable possessed entities, such as body parts, are usually signaled as identifiable because they are usually inferred via a frame that is associated with a previously introduced referent, i.e., the possessor (see Prince 1981).

Lambrecht (1994: 79–85) observes that it is worth to keep the pragmatic/cognitive category of identifiability distinct from the grammatical/formal category of definiteness. Prototypically, definiteness signals identifiability of the discourse referent associated with the NP, whereas the use of indefinite expressions signal the unidentifiability of the referent. As a matter of fact, the mapping between these two categories is rather imperfect: for example, consider the distinction between specific and non-specific as to indefinite expressions. In a sentence like *I am looking for a book*, *a book* can be interpreted as specific or non-specific depending on whether the speaker is looking for a particular book or not. This ambiguity is worked out within discourse by means of anaphoric reference; if the book is specific, a definite anaphoric expression must be used, like *I found it/the book*. If the book has non-specific reference, the anaphor must necessarily be an indefinite pronoun or NP, as in *I found one/a book*. In a similar vein, generic NPs, i.e., NPs whose referents identify either the whole class of all entities singled out by it or some representative set of members of this class, can be assumed to be identifiable (Lambrecht 1994: 82). By using these kinds of NPs, the speaker requires the hearer to identify the entire semantic class, not a specific referent, as *a camera* in the sentence *having a camera on a phone is useful for those quick*

unexpected captures, but they do not substitute a camera. Furthermore, the *camera* can be anaphorically referred to either via another indefinite NPs, as in the previous example, or a definite pronoun.

What discussed so far shows that the correlation between the cognitive (and presumably universal) category of identifiability and its formal coding, that is definiteness, is far to be one-to-one. Furthermore, the grammatical encoding of definiteness (by means of phonological, morphological, lexical or positional devices) displays a great amount of variability across languages. There are languages in which the grammaticalization of definiteness may not be fully developed (Lyons 1999: 278).

The distinction between the formal category of definiteness and the cognitive category of identifiability is fundamental for languages like Mandarin Chinese in which the interplay between identifiability and definiteness is a very complex issue. Even if it lacks definite articles, Mandarin Chinese employs other strategies to convey the definite/indefinite and identifiable/unidentifiable distinctions. Chen (2004: 1151) recognizes three major types of linguistic devices that are employed to signal (un)-identifiability, namely lexical, morphological, and positional ones.

As for the lexical strategies, identifiability is indicated by means of demonstratives, possessives and universal quantifiers. The demonstratives *zhè–zhèxiē* 'this–these' and *nà–nàxiē* 'that–those' are the primary means used to signal identifiability, serving some of the functions characteristic of definite articles in languages like English or Italian. However, both demonstratives still have strong deictic features that do not allow considering them as definite articles (Chen 2004: 1154). Quite interestingly, the indefinite determiner *yī* 'one' + classifier has developed a special use with definite referents, "serving as a backgrounding device marking entities as of low thematic importance and unlikely to receive subsequent mentions in ensuing discourse" (Chen 2003: 1182)[4]. For this reason, NPs introduced by *yī* + classifier will be treated as neutral with respect to the identifiability parameter in our analysis. All the NPs introduced by possessives and universal quantifiers, such as *dōu, suǒyǒu, yīqiè* 'all', are considered as definite in

---

4 A clear example of the pragmatic use of *yī* plus a classifier is given below. In this example, *yī* modifies a definite DO introduced by *bǎ*.

(i) *Ta   bei   pengyou   ba   yi   ge   taitati   gei   pian   zou   le*
    he   BEI   friend   BA   one   CLF   wife   PP   cheat   away   PFV
    'He was cheated by his friend out of his wife'
    (Chen 2003: 1173; glosses adapted)

Here the presence of *yī* does not indicate indefiniteness, but rather serves to signal the low thematic importance of the referent in the following discourse.

our analysis, following Chen (2004: 1157), who states that they are allowed only in definite positions (see below). As for the morphological devices, definiteness may be marked by means of reduplicated classifiers or nouns, which acquire the same meaning of universal quantifiers (see Chen 2004: 1163).

Moreover, some cases do exist in which NPs are neutral with respect to the identifiability parameter, such as bare NPs and NPs modified by cardinal numerals or a quantifier like *jǐ* 'how many/some'. Chen claims that whether these NPs are to be interpreted as identifiable or unidentifiable is suggested by the position of such NPs in the sentence. In fact, it has been recognized for a long time that sentence positions, and arguments which usually occur in those positions, are strongly influenced by the definiteness (and the degree of identifiability) of the NP (Keenan and Comrie 1977; Givón 1990, *inter alia*).

For instance, preverbal positions in SVO languages typically display a strong preference for definite expressions. By contrast, post-verbal positions are more likely to be occupied by indefinite expressions (Givón 1990). This pattern applies to Mandarin Chinese as well (Chen 2004: 1166). Chen (2004: 1168–1175) puts forward that Chinese has two positions with respect to definiteness: the definiteness-inclined position and the indefiniteness-inclined position. The latter is typically occupied by post-verbal NPs, NPs introduced in presentative constructions, etc. The former is likely to be filled by definite NPs serving as subject, and preverbal or *bǎ*-marked direct objects. The restrictions on the possibility for a NP to occur in a certain position may be quite severe. For instance, lexically or morphologically definite NPs may not be allowed in some post-verbal positions. In the following sections, we will show how the pragmatic status of identifiability strongly correlates with the occurrence of *bǎ*-marking: this hypothesis explains well the observed optionality of the marker, as well as its occurrence with indefinite or non-specific direct objects.

3 Parameters for the analysis of the data and structure of the corpus

Following Lambrecht and Chafe's approach, all direct objects in the data were classified according to their identifiability status[5]. We classified all the examples according to the following parameters (see Section 2.2):

---

5 In addition to identifiability status, Chafe (1976, 1994: 73; cf. Lambrecht 1994) suggests that a referent may be in different activation statuses, namely active, semi-active or inactive. An active (or given) referent is one that is considered in the focus of interest of the hearer because it has

(A) inherently identifiable;
(B) anaphoric or deictic identifiability;
(C) frame-established identifiability;
(D) not identifiable or unclear.

As mentioned in the introduction, this study is based on a sample of 201 occurrences of *bǎ* from Mandarin text excerpts, taken from the corpus of Modern Chinese of the Center for Chinese Linguistics (CCL) of Peking University. The CCL corpus is a free-access non-tagged corpus, containing around 477 million characters[6] (in the latest version, July 2009); for all excerpts, the source text is indicated; types of texts include spoken language, practical texts, comedies, novels, etc.

To build our own corpus, we first looked up *bǎ* in the CCL corpus, and we found 374.807 occurrences; among those, we randomly[7] chose a sample from the seven types of texts in which *bǎ* could be found. The composition of the corpus is given in Table 1.

As one may notice, the number of occurrences for different types of texts is rather unbalanced. For instance, we selected 40 occurrences of *bǎ* from the "spoken language" texts, but only ten from the categories of "practical texts" and of "newspapers and periodicals". The reason for this lies in the kind of data we needed for our analysis: to assess whether a referent is activated or not, one

---

been previously mentioned or is inherently present in discourse. A semi-active (or accessible) referent is present in the universe of discourse (for example, a referent evoked in discourse situation) but is not currently under discussion. An inactive (new) referent is one that has not been previously mentioned in the discourse, therefore being out of the focus of interest of the hearer. As Chafe notes, although identifiability and activation are closely interrelated, they should be kept distinct: while the former has to do with the *status* of referents in the mind of speakers, the latter concerns the *status* of referents in the mind of the hearer.

Although identifiability status and activation status are independent variables, they are still closely connected. First, when a referent is deictically activated, it is also (virtually always) active/given since, by definition, a deictically activated referent must have been previously mentioned in the discourse. Second, if a referent is not identifiable, then it is inactive (new). Also, if a referent is evoked by the frame, it will also be accessible, which means that, from the point of view of activation, it can be neither given nor new. Due to reasons of space, we shall take into consideration only the identifiability status in our analysis.

6 In the Chinese linguistic tradition, the character ($zì$) rather than the word ($cí$) is used as the basic unit for determining the size of a text (see Chao 1968:138). While the word has an uncertain status in the Chinese speakers' perception of the units of their own language, the character, a graphic unit, poses no problems for segmentation and, also, is almost always associated with a meaning, just like words in the familiar Indo-European languages of Europe.

7 For each category of texts, the first X occurrences were chosen; thus, for instance, for "spoken language" the first 40 hits were analyzed, for "biographies" the first 25 hits, etc.

Table 1: The structure of our corpus

Type of text	Occurrences of *bǎ*
Spoken language	40
Biographies	25
Practical texts	10
TV and Cinema	41
Newspapers and periodicals	10
Translated works	39
Literary works	36
Total	201

requires an excerpt which is sufficiently long; unfortunately, the excerpts were not all of the same length and, therefore, we privileged those in which there was "enough context" to determine whether the referent had been previously activated or not.

Admittedly, we included also a small number of very short excerpts (2–3 sentences); this is because they were part of dialogic texts, in which the usage of pronouns and demonstratives makes explicit the activation status of referents marked by *bǎ*. Also, although not all types are equally represented, we nevertheless believe that we have a fair number of occurrences from all the types, in order to obtain the greatest possible variety of texts.

However, in order to establish a firm correlation between identification status and DOM, it is necessary to compare the figures for *bǎ*-marked objects with corresponding data from unmarked post-verbal objects, as in (2). If we expect differentially marked objects to be identifiable, the prediction for unmarked object in the canonical position is that they tend to be non-identifiable. Data on the identifiability status of post-verbal objects can be compared to those of *bǎ*-marked objects only if the other conditions for the use of *bǎ* are met; above all, the event expressed by the sentence must be bounded (see 2.1).

We analyzed the same number of occurrences of unmarked post-verbal objects as for differentially marked ones in our sample; however, since there were not enough instances of post-verbal objects in sentences which met the criteria mentioned just above, we had to expand our corpus to obtain 35 more occurrences of post-verbal objects (11 for the category "spoken language" and 24 for the category "literary works"), again drawn from the CCL corpus. The number of occurrences for each type of texts is exactly the same for *bǎ*-marked and for unmarked objects (as illustrated in Table 1).

4 The identifiability status of *bǎ* objects *vs.* unmarked post-verbal objects

As outlined in the two preceding sections, we analyzed a sample of 201 occurrences of *bǎ* in text excerpts and classified each occurrence according to the identifiability status marked by *bǎ*; we repeated the analysis on an equal number of unmarked post-verbal objects appearing in predicates compatible with *bǎ*-marking (i.e., where DOM could have been used, but actually it was not). Some referents could not be easily recognized as for those parameters and, therefore, we had a few cases that had to be classified as unclear.

In order to illustrate our treatment of the data, let us provide a few (uncontroversial) examples of excerpts from our sample, one for each of the identifiability statuses mentioned in 3.1.

Inherently identifiable referent:

(11) *tā*   *xiǎng* *bǎ* *Zhū* *Dé* *gǎn-chū*   *bùduì*
     3SG.M want BA Zhu De drive-out army
     'He wanted to drive Zhu De out of the army'

Here, the proper name *Zhū Dé* can identify only one individual. Proper names, as personal pronouns, are intrinsically definite because they refer to an entity (usually animate, but it is not always the case) without describing them, e.g., they do not describe any quality of the entity designated by the noun (Croft 1991: 126–127). Thus, it is highly probable that both the speaker and the hearer have a mental representation of the referent designated by the proper name, and therefore they get marked by *bǎ*.

Deictic or anaphoric identifiability:

(12) a. *māma*   *bǎ* *jiǎo* *gěi*   *wǎi-le*
        mother BA foot give sprain-PFV
        'Mother sprained her foot'
     b. *Huímín* *sǐ-le*   *jiùshì* *èrshíbā* *chǐ*   *báibù*
        Huis   die-PFV just 28   1/3.of.metre plain.white-cloth
        dài-zǒu-le.   *Āi,*   *dào-le*   *kēngr* *lǐtou,* *bǎ* *zhè-ge*
        take-away-PFV Hey arrive-PFV hole   inside BA DEM-CLF
        báibù   *dǎkāi*   *ā* (...)
        plain.white.cloth unfold ah

'After a (person from the) Hui (nationality) has died, they take away 28
chi of plain white cloth. Oh, when they are inside the hole, they unfold
this plain white cloth, (. . .)'

In (12a), *jiǎo* 'foot' is a inalienably possessed entities, and is thus regarded as
anaphorically identifiable. In (12b), *báibù* 'plain white cloth' is introduced in the
first sentence as a new referent; in the second part of the excerpt, it is referred to
by the demonstrative *zhè-ge* 'this (one)', a marker of identifiability (see 2.2), and
it is marked by *bǎ*. We may conclude that *báibù* is an anaphorically identifiable
referent.

Frame-established identifiability:

(13) *wǒ   bù   shì   nǐ   qīzi,   wǒ   yǐjīng   bǎ   líhūn-xiéyìshū*
    1SG not COP 2SG wife 1SG already BA divorce-agreement
    jì-guòqu . . .
    send-away
    'I am not your wife, I have already sent the divorce agreement . . .'

The *líhūn-xiéyìshū* 'divorce agreement' marked by *bǎ* is evoked by the frame,
which we may call 'marital status', since the topic of the dialogue is, obviously,
the relationship between a divorced couple.
    A marginal phenomenon which we encountered in our analysis of the sample
is the differential marking of two coordinated objects (say, A and B), of which only
one is identifiable (A):

(14) (. . .)   *Zhōng-Gòng   Zhōngyāng* (. . .)   *zhàokāi   zhèngzhì-jú*
        Chinese-CP   Central.committee   convened   political-bureau
    kuòdà-huìyì,   *juédìng   bǎ   hǒng-yī-fāngmiànjūn   yǔ*
    enlarged-meeting   decide   BA   red-one-army   with
    Jūn-wěi   *zòngduì   zhěngbiān   wéi   Zhōngguó*
    military-commission   column   regroup   into   China
    gōng-nóng   *Hóng-jūn   Shǎn-Gān   zhīduì*
    worker-peasant   Red-Army   Shaanxi-Gansu   detachment
    'The Central Committee of the Chinese Communist Party (. . .) convened an
    enlarged meeting of the Politburo and decided to regroup the First Army of
    the Red Army and the column of the Military Commission into the Shaanxi-
    Gansu detachment of the Workers' and Peasants' Red Army'

The *bǎ*-marked referents are the *hǒng-yī-fāngmiànjūn* 'First Army of the Red Army' and the *Jūn-wěi zòngduì* 'column of the Military Commission'; only the *hǒng-yī-fāngmiànjūn* was mentioned earlier in the text. In such cases, we counted the object as identifiable.

As stated in the introduction, the aim of our study was that of providing evidence in favor of a strong correlation between the use of *bǎ* and the identifiability status of the referents; more precisely, we proposed that a *bǎ*-marked referent will be most likely to be readily identifiable by the receiver (hearer/reader). The only case when *bǎ*-marking of the object is not expected is when the object is not identifiable, i.e., when a "discourse representation" of a referent is not present in the hearer's mind (cf. the quotation from Lambrecht 1994: Section 2.2), when the referent has not been mentioned before and it is not present in the universe of discourse. However, some occurrences of unidentifiable, referents as *bǎ*-marked objects are attested in our sample. The general figures are given in Table 2.

So, 15 out of 201 occurrences of *bǎ*-marked objects were unidentifiable, new referents (see Example (15)). This is a very small number, which amounts to less than 7.5% of the total occurrences; in around 86.6% of the cases the marked object was identifiable. We also have 12 dubious cases (ca. 6%) as far as the identifiability status is concerned.

What about unmarked post-verbal objects? As stated before (3.2), the opposite prediction would seem logical: post-verbal objects should tend to be unidentifiable, new referents. Since the most relevant distinction here is that between unidentifiable and identifiable referents, whatever the basis for their identifiability may be, we divided the occurrences of post-verbal objects between those two categories, simplifying the picture. In Table 3. we summarize our quantitative data on the interaction of activation status with differential *vs.* canonical object marking, both in absolute and in percentage terms:

We used both Fisher's test and chi-square (with Yates' correction) to assess the statistical correlation between identifiability and post-verbal *vs.* preverbal *bǎ*-marked objects. We indeed found an extremely significant statistical correla-

Table 2: The identifiability status of *bǎ*-marked objects in our sample

Inherently identifiable	19
Deictically / anaphorically identifiable	126
Frame-established identifiability	29
Unidentifiable	15
Unclear	12
Total	201

Table 3: The identifiability of *bǎ*-marked and post-verbal objects in our sample

	Identifiable		Unidentifiable		Unclear	
	Occurr.	%	Occurr.	%	Occurr.	%
bǎ-marked	174	86.57	15	7.46	12	5.97
post-verbal	79	39.30	96	47.76	26	12.94

tion (i.e., χ^2 [1, $N = 401$] = 92.18, *p-value* = <0.0001) between differentially marked objects and identifiability on the one hand and post-verbal objects and unidentifiability on the other hand.

Identifiability and *bǎ*-marking in Mandarin thus show a strong, albeit not absolute, correlation, i.e., differential marking of identifiable objects is not a rule, strictly speaking. As mentioned above (2.1), the occurrence of DOM in Chinese is conditioned by other factors as well, such as the boundedness of the event denoted by the verb, animacy of the referent, etc. Probably some of those factors account for the few instances of unidentifiable *bǎ*-marked objects. For instance, in (15) we have a very heavy *bǎ*-marked object, which appears to be unidentifiable:

(15) *Zhū Dé bǎ cóng guówài xué-dào de xiānjìn jūn-shì*
     Zhu De BA from abroad study-RES DET advanced military-affairs
     sīxiǎng, guòqù dài bīn dǎzhàng de jīngyàn yǔ
     idea   in.the.past lead soldier fight   DET experience and
     Jǐnggāng-shān-qū gémìng dòuzhēng de shíjì jǐnmì
     Jinggang-mountain-area revolution struggle   DET reality close
     jiéhé-qilai
     integrate-RES
     'Zhu De closely integrated the advanced military ideas he had learned from abroad, his own experience leading soldiers into battle and the reality of the revolutionary struggle in the Jinggang mountains area'

The *bǎ*-marked object in (15) is made up of a series of coordinated noun phrases. As mentioned in the introduction, Liu (2007) found that *bǎ* marks NPs carrying old information as well as NPs carrying new information, provided that the latter are heavy; such an analysis would explain the usage of *bǎ* with a new referent in (15).

What about the high number of identifiable post-verbal objects? Firstly, since post-verbal collocation of objects is the unmarked option, it is not surprising that

both identifiable and unidentifiable objects are found in that construction; it is the need to mark identifiability, among other factors, which triggers the marked construction, i.e., *bǎ*-marking. In their discussion of the fronting of relative clauses in Cantonese, Matthews and Yeung (2001:90) suggest that when a relative clause modifies the object of the main verb in a sentence, topicalization (i.e., fronting) of the object is common for processing reason (16a), but the 'canonical' SVO sentence (16b) is nonetheless perfectly grammatical (glosses adapted)[8]:

(16) a. *Ting¹ jat⁶   gong² go²   di¹ je⁵   lei⁵   zeon² bei⁶ hou²   mei⁶   aa³*
          tomorrow talk   that CLF stuff you   prepare   finish not   PRT
     b. *Lei⁵   zeon² bei⁶ hou²   ting¹ jat⁶   gong² go²   di¹   je⁵   mei⁶   aa³*
          you   prepare   finish tomorrow talk   that CLF   stuff   not   PRT
          'Have you finished preparing the stuff you're talking about tomorrow?'

Matthews and Yeung report the results of an experimental study according to which sentences with a heavy NP object such as *ting¹ jat⁶ gong² go² di¹ je⁵* 'the stuff you're talking about tomorrow' are read and understood faster by native speakers of Cantonese if such heavy NP is moved before the verb; such order may be preferable, but it is in no way the only choice, as the corresponding unmarked SVO structures are anyway possible. The same may be said about DOM and OV order in Mandarin: identifiable referents may trigger *bǎ*-marking, but this does not entail that they cannot appear as post-verbal objects in an unmarked SVO sequence; we just expect that when DOM occurs, the object be identifiable, more often than not, and this expectation is borne out by our data.

Secondly, it is very difficult to judge whether *bǎ*-marking is plausible for a certain object in a specific predicate: apart from identifiability status, as we have seen, it is not entirely clear what factors come into play in Chinese DOM, and there is a lot of variation among different speakers as to which predicates admit *bǎ*-marking, when such construction is odd (dispreferred) or ungrammatical, etc. For instance, it is possible that some of the unidentifiable post-verbal objects which we found in our sample did not get *bǎ*-marking (and fronting) because they were not highly topical; LaPolla (1995: 310) proposed a "Principle of word order in Chinese" according to which "[t]opical or non-focal NPs occur preverbally and focal or non-topical NPs occur post-verbally". Hence, according to this principle, a focal object would be located after the verb, independently from its identification status; here is one example from our corpus:

---

8 The examples are provided by Matthews and Yeung in the Jyutping romanization system; tones are marked by superscript numbers (1–6).

(17) *Sòng   Jiànpíng* [. . .] *jiù   yòu   bǎ tā   bāla-dào-le   yībiān,*
     Song   Jianping   then   again BA 3SG.F push-arrive-PFV one-side
     rǎnhòu lākāi   mén (. . .)
     then   pull-open door
     'Song Jianping [. . .] then pushed her again aside, then pulled the door
     open (. . .)'

In this excerpt, *tā* 'she/her' is an identifiable non-focal object which is differen-
tially marked, as expected; *mén* 'door' is a again an identifiable referent (it had
been mentioned before in the passage), but it is non-topical and it appears in the
post-verbal position, as predicted by LaPolla's principle stated above. This is but
a simple example of how other factors may intervene in the positioning of the
object in a given sentence.

   To conclude our discussion, we believe that it is also worth discussing some
unclear cases we found in our sample. In the following excerpt, the *bǎ*-marked
referent is taken by us to be evoked by the frame; however, its activation status
is unclear:

(18) *tā   zǒngjié-chū* [. . .] *zhàn-shù   yuánzé.* [. . .] *Lín Biāo bǎ*
     3SG.M   summarize-RES   military-tactic principle   Lin Biao BA
     fùzá   de   zhàn-shù   wèntí yòng jǐ   gè   zì
     complicated DET military-tactic issue use   a.few CLF   word
     jiǎnliàn-de   gàikuò
     succint-ADV   summarize
     'He [Lin Biao] summarized [. . .] principles of military tactics. [. . . .] Lin Biao
     summarized complicated tactical issues using just a few words'

Firstly, a series of 'principles of military tactics' (*zhànshù yuánzé*) is introduced;
later on, we find the object *zhànshù wèntí* 'tactical issues' marked by *bǎ*. We could
regard *zhànshù wèntí* as a given referent, if it were used in the text as a near-
synonym of *zhànshù yuánzé*; otherwise, we should say that it is accessible, being
evoked by the frame. The latter analysis appears as more appropriate, but we
cannot be certain about it, given the context.

   The case below is somehow trickier:

(19) *cǐ   zhàn zhī qián,   qīn   Huá Rì-jūn   bìngwèi*
     this battle DET before invade China Japanese-army not.yet
     yù-dào   shénme dǐkàng.   Tā-men ba jǐ-shíwàn
     meet-RES any   resistance 3-PL   BA some-hundred.of.thousands

Huā-běi	*Guómíndǎng*	*jūnduì*	*qūgǎn*	*de*	*wàngfēng'értáo*
China-north	KMT	army	drive.away	ADV	flee.at.the.sight.of

'Before this battle, the Japanese China Expeditionary Army had not yet met any resistance. They drove away some hundreds of thousands from the North China Kuomintang army, who fled at their sight.'

The *bǎ*-marked object is *jǐ-shíwàn Huā-běi Guómíndǎng jūnduì* 'some hundred of thousands from the North China Kuomintang army', a referent which was not mentioned before. Since the whole excerpt is about the anti-Japanese resistance in China, reference to troops, generally speaking, is most certainly evoked by the frame and, thus, may be regarded as present in the universe of discourse. However, here reference to a specific army is made (the North China Kuomintang army); it is not clear whether such referent can be frame-activated. One alternative analysis could be that the *Huā-běi Guómíndǎng jūnduì* be regarded as a unique referent and, therefore, inherently identified; however, the object is not the army as a whole, but some hundreds of thousands (*jǐ-shíwàn*) of troops from it, casting doubts on the uniqueness of the referent.

In the following excerpt, it is again unclear whether the differentially marked object may be regarded as inherently identifiable, despite the fact that we are dealing with a proper name:

(20) | *zài* | *tā* | *zǎo-qī* | *cóngshì* | *diànyǐng* | *gōngzuò* | *shí*, |
|---|---|---|---|---|---|---|
| at | 3SG.M | early-period | be.engaged.in | film | work | time |

tā	*bǎ*	"*Mǎbùsī*	*bóshì*"	(*Fúlǐcí*	*Lǎnggé*	*dǎoyǎn*,	1922	*nián*
3SG.M	BA	Mabuse	doctor	Fritz	Lang	direct	1922	year

shèzhì)	*zhè*	*bù*	*yǐngpiàn*	*gǎo-chéng*	*Sūlián-bǎn*
produce	DEM	CLF	film	do-RES	USSR-edition

'In the early days of his engagement in cinema, he [S. Ejzenštejn] made the Soviet edition of "Dr. Mabuse" [Fritz Lang 1922]'

Here the *bǎ*-marked object is the 1922 film "Dr. Mabuse", which is indeed a unique referent and, therefore, could be understood as inherently identifiable. As seen above, is it *highly* probable that the sender and the receiver of the message share a mental representation of the referent designated by a proper name but, nevertheless, it is not to be given for granted. As a matter of fact, one could introduce a new participant in discourse by a proper name (although, perhaps, this is less often the case for proper names). In (20), additional information is given to identify the marked object (namely, the director of the film and the year of production); this could suggest that the writer regarded "Dr. Mabuse" as a possibly new referent for the reader. In such kind of written texts, where there is no interaction

with the receiver, it is not easy to figure out whether a referent designated by a proper name is always to be understood as inherently identifiable.

5 Conclusion

In the literature on DOM in Chinese, several factors have been called into play to explain the constraints on the occurrence of *bǎ*-marking in Mandarin, but no satisfying account has been proposed up to date, as discussed earlier in Sections 1 and 2.1. In the present paper, we have put forward the hypothesis that the occurrence of DOM in Mandarin Chinese is best explained by analyzing the identifiability and activation status of the *bǎ*-marked referent: DOM is to be expected when the referent is identifiable and, from the point of view of activation, either given or accessible; new referents are unlikely to be introduced as DOs. We analyzed a corpus of text excerpts of Modern Chinese and we showed that only 15 out of 201 occurrences of *bǎ*-marked objects were unidentifiable, new referents (less than 7.5% of the total occurrences). Therefore, it appears that there is a strong correlation between identifiable/active referents and *bǎ*-marking; differential marking of new referents, albeit not impossible, appears to be very uncommon. To test our hypothesis, we also searched for post-verbal unmarked objects in predicates compatible with *bǎ*-marking in the same sample (see Section 3 for the details), expecting to have more unidentifiable objects. What we actually found is that about 39% of those post-verbal objects were actually identifiable, and as many as 47.76% were unidentifiable, thus showing a less clear correlation with identifiability status; nevertheless, the *p*-value we obtained for the correlations is indeed statistically significant, and thus our hypothesis stands the test of data.

What remains to be investigated are the correlations between the identifiability/activation status of the DO and the other factors which constrain the occurrence of *bǎ*, as e.g., the kind of event described by the verb phrase (bounded *vs.* unbounded), animacy, etc. For instance, we remarked that the heaviness of the object might play a role, triggering *bǎ*-marking for inactivated referents, when heavy (see Example 15), as suggested by Liu (2007), even though it is unclear how a single marker may signal both new and old information (see Matthews and Yeung [2001] for a cognitive processing account). Moreover, there might be other "triggers" for the occurrence of DOM which have not been considered yet in the literature, and which could help account for the few *bǎ*-marked referents which we found in our corpus.

Also, since we built and analyzed a corpus of texts containing DOM, we did not consider in detail the problem of optionality, i.e., we did not discuss the acceptability of the very same objects which we had in the excerpts if used with-

out *bǎ*. A *continuum* ranging from "impossible" to "obligatory" for *bǎ*-marking has been proposed in the literature (Li and Thompson 1981: 487), based on other criteria, i.e., definiteness and the so-called "disposal" of the object. As we said above, *bǎ*-marking of new referents is not impossible, but rather unlikely; we can therefore hypothesize that the likelihood of DOM is higher for identifiable / activated DOs, thus explaining the skewed distribution of *bǎ*-marking for referents according to their activation status. Also, following LaPolla's (1995) principle of word order in Chinese, we remarked that the constituents occurring before the verb are typically topical or non-focal, whereas those occurring after the verb are focal or at least non-topical; such general principle may explain the many cases in which non-topical or focal identifiable object NPs are found in post-verbal position and do not get DOM, although other conditions for *bǎ*-marking are met. In short, identifiability appears as a trigger for DOM in Mandarin, interacting with other constraints and tendencies both at the syntactic and at the pragmatic level.

References

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory* 21(3). 435–483.

Bossong, Georg. 1985. *Empirische Universalienforschung: Differentielle Objektmarkierung in der neuiranischen Sprachen*. Tübingen: Narr.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li (ed.), *Subject and topic*, 25–55. New York: Academic Press.

Chafe, Wallace. 1987. Cognitive constraints on information flow. In Russell Tomlin (ed.), *Coherence and grounding in discourse*, 21–51. Amsterdam & Philadelphia: John Benjamins.

Chafe, Wallace. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Chao, Yuen Ren. 1968. *A grammar of spoken Chinese*. Berkeley & Los Angeles: University of California Press.

Chen, Ping. 2003. Indefinite determiner introducing definite referent: A special use of *yi* 'one' + classifier in Chinese. *Lingua* 113(12). 1169–1184.

Chen, Ping. 2004. Identifiability and definiteness in Chinese. *Linguistics* 42(6). 1129–1184.

Comrie, Bernard. 1979. Definite and animate objects: A natural class. *Linguistica Silesiana*. 13–21

Comrie, Bernard. 1989. *Language typology and linguistic universals: Syntax and morphology*. Oxford: Blackwell.

Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago: Chicago University Press.

Givón, Talmy. 1990. *Syntax: A functionaltypological introduction*, vol. 2. Amsterdam & Philadelphia: John Benjamins.

Hopper, Paul & Sandra Thompson. 1980. Transitivity in grammar and discourse. *Language* 56(2). 251–299.

Keenan, Edward & Bernard Comrie. 1977. Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry* 8(1). 63–99.

Lambrecht, Knud. 1994. *Information structure and sentence form: topic, focus, and the mental representation of discourse referents*. Cambridge: Cambridge University Press.

LaPolla, Randy J. 1995. Pragmatic relations and word order in Chinese. In Pamela Downing & Michael Noonan (eds.), *Word order in discourse*, 297–329. Amsterdam & Philadelphia: John Benjamins.

Li, Yen-hui Audrey 2006. Chinese *ba*. In Martin Everaert & Henk van Riemsdijk (eds.), *The Blackwell companion to syntax*, vol. 1, 374–468. Malden, MA: Blackwell.

Li, Charles & Sandra Thompson. 1981. *Mandarin Chinese: A functional reference grammar*. Berkeley, CA: University of California Press.

Liu Feng-Hsi. 1997. An aspectual analysis of *ba*. *Journal of East Asian Linguistics* 6(1). 51–99.

Liu Feng-Hsi. 2007. Word order variation and *ba* sentences in Chinese. *Studies in Language* 31(3). 649–682.

Matthews, Stephen & Louisa Y. Y. Yeung. 2001. Processing motivations for topicalization in Cantonese. In Kaoru Horie & Shigeru Sato (eds.), *Cognitive-functional linguistics in an East Asian context*, 81–102. Tokyo: Kurosio Publishers.

Yang, Ning. 2008. *The indefinite object in Mandarin Chinese: Its marking, interpretation and acquisition*. Utrecht: LOT Publishers.

Yang, Ning & Geertje van Bergen. 2007. Scrambled objects and case marking in Mandarin Chinese. *Lingua* 117(9). 1617–1635.