Hindawi
Complexity
Volume 2022, Article ID 2156978, 8 pages
https://doi.org/10.1155/2022/2156978

Research Article

Expletive *ta* in Mandarin Chinese: A Quantitative Study

Yue Yu and Qiurong Zhao

School of Foreign Studies, University of Science and Technology Beijing, Beijing 100086, China

Correspondence should be addressed to Qiurong Zhao; qiurong.zhao@ustb.edu.cn

Received 31 October 2021; Revised 22 December 2021; Accepted 30 December 2021; Published 22 February 2022

Academic Editor: Ning Cai

Expletive pronouns have posed significant challenges to both linguistic studies and natural language processing in Chinese. Based on the data from the Chinese corpus at the Center for Chinese Linguistics of Peking University (CCL), this paper investigates the automatic dependency parsing errors of expletive *ta* ("it/he") and relative constructions and provides a detailed quantitative analysis of the characteristics of the constructions in question. The findings not only provide evidence to show that even in radical pro-drop languages such as Mandarin, pleonastics can occur in both object and subject positions but also offer important insights for Chinese natural language processing.

1. Introduction

Expletive pronouns, which have no antecedent or referent, are also called "dummy," "empty," or "nonreferential" pronouns in the literature and have received considerable attention from lingustic studies in recent decades (see [1–4]; Rothstein [5–8]; Wu and Matthews [9]). Taking the third person singular pronoun *ta* ('he/it/she') in Mandarin as an example (in spoken Chinese, the third person pronoun *ta* does not make a distinction between feminine and masculine; while in written Chinese, it can be orthographically differentiated into three forms, namely, 它 ('it'), 他 ('he'), and 她 ('she')), it is canonically construed as referential in the sense that it is anaphorically or cataphorically related to a singular nominal element in the preceding or the following utterances. When used expletively, it has neither an antecedent nor a referent, as shown in (1) and (2).

(1) a. *Wo yao haohao shui ta san-tian.*

I want good good sleep it three day (English word-by-word translation)

'I want to have a sleep for three days' (English sense translation)

b. *He ta gegou.*

drink it classifier (CL)enough

'Drink to one's satisfaction.'

(2) *Ni zhan qilai,pao qilai,guan ta you xie mei xie.*

you stand up run up no matter it have shoe not have shoe

'Just stand up and run, no matter whether you wear shoes or not.'

In the above examples, *ta* does not refer to any entity in the real world; in other words, expletive *ta* is realized phonetically, while it is semantically empty. This form-meaning mismatch has inevitably posed significant challenges to lingustic studies and natural language processing.

While a great deal of research has been conducted on expletive *ta* from a qualitative perspective, which has raised controversial questions about its syntactic and semantic properties, it has been studied far less from a data-driven approach. One of the central shifts in linguistic research is that it calls for a quantitative study methodology as Liu [10] points out. He remarks that data-driven research is consistent with modern artificial intelligence based on deep learning. Zhang et al. [11] stress that understanding languages' nature and organizational rules provides essential insights for natural language processing. Therefore, this paper provides a thorough and fine-grained quantitative survey of Chinese expletive *ta* and relative constructions

based on authentic data extracted from the Chinese corpus at the Center for Chinese Linguistics of Peking University (CCL). The contributions of this study are twofold: first, it is hoped to provide a detailed and objective analysis of the characteristics of expletive *ta* constructions; second, some insights into dependency parsing for the constructions at issue will be offered.

This paper is organized as follows. Section 2 presents a critical review of previous studies on Chinese expletive *ta* and relative constructions. Section 3 introduces the data and methods adopted in the current study. Section 4 presents the research results. Section 5 discusses our findings. Finally, Section 6 offers a conclusion.

2. Previous Studies on Mandarin Expletive *ta*

We review two areas of research relevant to the design and implementation of this study: research on the syntactic properties and semantic constraints of expletive *ta* from the perspective of traditional linguistics and research on pronoun resolution from the perspective of computational linguistics.

2.1. Research from the Perspective of Traditional Linguistics. The existing research on expletive *ta* is mainly concerned with constructions such as *shui ta san-tian* ('to sleep for three days'), as illustrated in example (1), where a considerable amount of ink has been spilled on the syntactic properties of *ta*. Some researchers argue that the construction in question is a double-object construction or pseudo-dual-object construction, and *ta* serves as the object or pseudo-object ([1–3, 12] among many others), while Yuan [13] states that the construction is a single-object construction, and *ta* is combined with the preceding verb, and together, they form a disyllable word. Similarly, Ōta [14], and Lin and Zhang [8] argue that *ta* is encliticised in the verb root, which has grammatically evolved into a clitic. In contrast, Lei and Hu [15] argue that *ta* is closer to the word that follows it. For example, they believe that *ta* ('he') and *ge* (CL) have developed into a compound word in *he ta ge gou* ('drink to one's satisfaction').

The semantic constraints of the verb before *ta* have also attracted much attention. Some researchers observe that the verbs in constructions such as *shui ta san-tian* ('to sleep for three days') are volitional verbs and cannot adopt perfective aspect markers [3, 16]. Other researchers argue that the constructions in question can adopt perfective aspect markers based on their intuition (e.g., Wang and Wang [17]).

As we can see, much research has been done on expletive *ta* constructions similar to the constructions in example (1), but the above issues are still unclear, mainly because the studies discussed so far have mostly been qualitative and based on the researchers' personal linguistic knowledge and linguistic intuition. Moreover, little is known about constructions such as *guan ta youxie meixie* ('no matter whether one wears shoes or not') as illustrated in example (2), where *guan* is understood as a conjunction meaning 'no matter what' or 'regardless of' and *ta* occupies the subject position of the conditional clause (Lü [18]), which have different characteristics from constructions such as example (1).

2.2. ta in the Perspective of Computational Linguistics. Pronoun processing has long been a key issue in natural language processing, focusing on anaphora resolution, that is, identifying the entities with which pronouns can match (see, e.g., [19]). Taking Chinese pronoun resolution as an example, many researchers have investigated various methods to increase the reference resolution effect, including a deep learning method ([20]; Chen et al. [21]), a tree kernel method [22], and a combinatorial method of semantics and rules [19].

Although the previous research has contributed significantly to the information processing and mining of Chinese texts, the existing methods focus on the processing of referential pronouns and aim to filter out the correct referents. A natural consequence is that nonreferential (expletive) pronouns have been relatively neglected. Nonreferential pronouns, we assume, are treated similarly to referential pronouns, and noun phrases or other types of elements are incorrectly assigned to them as referents.

In this study, we take automatic dependency parsing as an example to explore the challenges that nonreferential pronoun *ta* poses to Chinese information processing, hoping that a more complete and fine-grained quantitative exploration of expletive *ta* and relative constructions in a latter section may shed some light on those questions.

3. Data and Methods

3.1. Data Sources. The authentic language data of expletive *ta* are taken from the Center for Chinese Linguistics corpus of Peking University (CCL). CCL (http://ccl.pku.edu.cn:8080/ccl_corpu) is a comprehensive corpus designed to represent both Classical Chinese and Modern Mandarin, which contains about 700 million Chinese characters with samples of both spoken and written Chinese. It consists of various types of texts such as newspapers, magazines, literature, practical texts, television scripts, movie scripts, natural dialogues, TV interviews, and translated texts. Many types of language genres together represent a good diversity of usages in Modern Mandarin. Consequently, the study of expletive *ta* based on the CCL corpus is assumed to be convincing.

3.2. Data Collection. We collected 2,758,189 instances of 他('he'), 423,213 of 它 ('it'), and 779,885 of 她 ('she') using open-source statistical software PowerGREP (Joyce [23]) since the pronoun *ta* has three orthographical forms. In order to obtain only potential expletive *ta* tokens, examples with an obvious referential interpretation were excluded from the results.

For example, cases where *ta* was used to compose words such as the plural pronoun *tamen* ('they') and possessive

pronoun *tade* ('his/her/its') were excluded. Subsequently, the remaining data were manually examined and rechecked, ultimately extracting 1,346 instances in which *ta* had a nonreferential reading. As shown in Table 1, there are 617 instances of expletive *ta* ('he') and 729 of *ta* ('it'), while no expletive *ta* ('she') is found.

3.3. Dependency Parsing. Dependency parsing of expletive *ta* and relative constructions is grounded in dependency grammar [24–28]. Dependency parsing is currently the mainstream method for Chinese syntactic analysis, especially in the field of natural language processing, due to the simplicity of its form, adherence to human language intuition, ease of application, and ability to deal with word-order flexibility (Liu and Zhang [29]). According to dependency grammar, the syntactic structure of a sentence consists of nothing but the dependencies between individual words. A syntactic dependency relation has a number of properties:

(a) It is a binary relation between two linguistic units.

(b) It is usually asymmetrical and directed, with one of the two units acting as the head and the other as the dependent.

(c) It is labeled, and the type of the dependency relation is usually indicated as using a label on top of the arc linking the two units ([30]: 1568).

Figure 1 provides an example. The directed arcs in the graph from the head to dependent show the asymmetric relationship between two units. The label on the arc stands for the dependency type or syntactic function and shows the dependency relation between the head and dependent. For example, the label on the arc from "has" to "apple" is "obj," which means "apple" serves as the object of the verb "has." Dependency analysis can thus be viewed as the set of all dependencies in a sentence.

This study is mainly concerned with the dependencies related to expletive *ta*, including the head and dependent asymmetric relationship and the corresponding dependencies, which are automatically processed by Stanford Parser (Chen and Manning [31]).

3.4. Automatic Parsing. Word segmentation and part-of-speech tagging were first implemented using software ICTCLA (developed by the Chinese Academy of Sciences) before manual correction. Subsequently, the tagged data were processed by the Stanford Parser, an automatic parsing tool, to conduct dependency parsing. According to an empirical study by Liu and Zhang [29], the accuracy of the Stanford Parser for automatic syntactic parsing of Chinese is the highest among current parsers. They report that the precision rates of dependencies for literary and nonliterary Chinese texts done by Stanford Parser are as high as 82.16% and 85.82%, respectively.

3.5. Quantitative Analysis. After the automatic parsing, we manually checked the results, focusing on the dependencies related to expletive *ta*. Then, we took a closer look at the extracted original data further to explore the features of the constructions in question and provided a complete and fine-grained quantitative analysis of the constructions, hoping to shed some light on the syntactic parsing of expletive *ta* constructions.

Table 1: Expletive *ta* in the CCL corpus.

Expletive	Instances
Expletive *ta* ('he')	617 (45.8%)
Expletive *ta* ('it')	729 (54.2%)
Total	1346

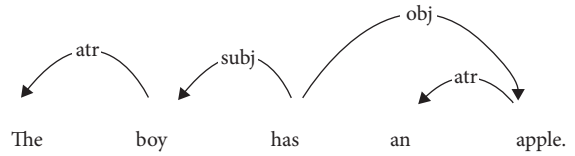

Figure 1: Syntactic dependency structure of *The boy has an apple.*

4. Findings and Results

4.1. Automatic Parsing Results

4.1.1. Low Automatic Parsing Accuracy. We first examined the accuracy rate of dependencies with respect to expletive *ta*, including head-dependent relation and dependencies. The results show that the dependency accuracy of parsing special constructions is much lower than the mean accuracy of 82.16% for literary or 85.82% for nonliterary texts (Liu and Zhang [29]). In the current study of expletive *ta* constructions, only 29.3% of the head-dependent relations were correctly identified, and the dependencies were all incorrectly annotated, as illustrated in Table 2.

4.1.2. Various Spurious Dependencies. Next, we scrutinized the dependencies related to expletive *ta* and found that several dependencies had been falsely recognized, as shown in Table 3.

As shown in Table 3, dobj (direct object), nn (noun compound modifier), and nsubj (nominal subject) are the most common false dependencies, indicating that expletive *ta* was falsely construed as referential in these instances. As we can see, automatic syntactic parsing of special constructions in Chinese, such as expletive *ta* constructions, is far from satisfactory.

As demonstrated by many empirical studies, linguistic knowledge can effectively improve the performance of dependency parsing ([11] and others). In the next section, we present an objective quantitative analysis of expletive *ta* constructions, attempting to provide linguistic knowledge for syntactic parsing of these constructions.

4.2. Characteristics of Expletive ta and Relative Constructions

4.2.1. Dependencies: Indirect Object (iobj) and Expletive Subject (expl). Based on the CCL data, we manually examined the distributional properties and semantic

Table 2: Accuracy of dependencies of expletive *ta*-related nodes.

	Head-dependent relation (%)	Dependency (%)
Expletive *ta*-related nodes	29.3	0

Table 3: False dependencies of expletive *ta*-related nodes.

Dependency of *ta*-related nodes	Ratio (%)
dobj (direct object)	35.8
nn (noun compound modifier)	30.7
nsubj (nominal subject)	22.5
dep (dependent)	7.7
det (determiner)	1.9
pobj (propositional object)	1.5

constraints of expletive *ta* constructions to clarify these constructions' characteristics further. Firstly, there are mainly two types of constructions that allow expletive *ta*: V + *ta* + NP constructions and conditional clauses, which are usually introduced by the conjunction *guan* ('no matter what'/'regardless of'), as shown in Table 4.

It is worth noting that though *ta* occurs in the postverbal position, that is, the object position in a *V* + *ta* + *NP* construction, it does not serve as the direct object of the verb as automatically parsed by the Stanford Parser. In generative grammar, it is generally assumed that objects are usually subcategorized by verbs and therefore theta-marked. Therefore, expletive pronouns can never appear in a theta-marked position since they are semantically empty. In traditional grammar, as mentioned earlier, such constructions are considered as pseudo-double object constructions, and expletive *ta* is considered a pseudo-object or an indirect object [2, 3, 12].

Moreover, in a conditional clause such as *guan ta youxie meixie* ('no matter whether one wears shoes or not'), expletive *ta* occupies the subject position. Here, *ta* does not refer to an entity. And the conditional clause is usually introduced by *(bu) guan* ('no matter what'), which is a conjunction (see also Lü [18]: 523). Consequently, *ta* cannot be parsed as the direct object of *(bu) guan*, as falsely parsed by the Stanford Parser.

4.2.2. Characteristics of the *V* + *ta* + *NP* Constructions

(1) Verbs in V + ta + NP Are Volitional Verbs. The verbs in *V* + *ta* + *NP* constructions are generally volitional [3]. We collected 974 instances of the *V* + *ta* + *NP* construction, and there are 280 verbs, of which only twenty-one are typical nonvolitional verbs out of context, as listed in Table 5.

It is worth mentioning that nonvolitional verbs as listed above can be used as volitional verbs when they precede expletive *ta* (as is noted in Ma [32], some nonvolitional verbs can be used as volitional ones when occurring in an irrealis context.), as in example (3), where the verb *si* ('to die') is different from that in *Ta si-le* ('He has been dead') since the former means 'to go to die,' which is volitional. In other words, most of the verbs in *V* + *ta* + *NP* constructions are nonstative (since stative verbs are nonvolitional; see [32, 33]).

(3) *Si ta ge guangrong.*
die it CL glory
'To die with glory.'

Moreover, the verbs tend to be monosyllabic. Only a minority are disyllable words, as shown in Table 6.

(2) High-Frequency Verbs in V + ta + NP. In general, both transitive and intransitive verbs can take expletive *ta* as an indirect (pseudo) object in Mandarin, as shown in Table 7, which lists the top 10 high-frequency words.

As can be seen from Table 7, these are primarily regular transitive verbs (nine out of ten of the most frequently used verbs are regular transitive verbs), such as *da* ('to hit'), *gan* ('to do'), and *sha* ('to kill'). However, intransitive verbs can also occur, as shown in the following examples:

(4) a. *Wo xiang xia xianggang qu zhu ta ji-tian.*
I want go Hong Kong go live it several day
'I want to go to Hong Kong and stay several days'
b. *Ku ta ge tian-hun-di-an.*
cry it CL sky murky earth dark
'Have a really good cry.'

Out of context, verbs such as *zhu* ('to live') and *ku* ('to cry') are subcategorized as intransitive because they do not identify an object. Nevertheless, in the above cases, the verb obligatorily takes an NP as its complement, such as the temporal expression *ji-tian* 'several days' in example (4a) and the resultative nominal *ge tian-hun-di-an* 'a murky sky over a dark earth' in example (4b). For this reason, the NPs are treated as 'quasi-object NPs' in the relevant literature [2, 3, 13]. Temporal expressions or resultative expressions as such, albeit usually treated as adjuncts, seem to function more as complements than as verb modifiers in the sense that their occurrence is obligatory, thus blurring the distinction between the argument and adjunct to some extent.

(3) NP in V + ta + NP Constructions: Modified by Classifier ge or a Numeral. Importantly, NP in *V* + *ta* + *NP* constructions appears to be an indefinite nonspecific expression as has also been observed in Ma [3], Lin [4, 34], and Yuan [13]. This explains to some extent an nn (noun compound modifier) relation is often incorrectly detected during automatic parsing by the Stanford Parser.

Specifically, the nonspecific indefinite NP consists solely of the classifier *ge* or a numeral (the numerals here include the word *ji* 'several' and *ban* 'half'; both indicate the number of entities).

As can be seen from Table 8, forty percent of the NPs consist of the classifier *ge*, whose occurrence nominalizes the post-*ta* elements. Sixty percent of the NPs begin with numerals such as *yi* ('one'), *san* ('three'), *shi* ('ten'), and *qian* ('thousand'). Bare NPs or definite NPs do not occur. On the question of why

Table 4: Dependencies of expletive *ta*-related nodes.

Distribution	Dependency	Ratio of instances (%)
V + *ta* + *NP*	Indirect object	72.4
Conditional clause	Expletive subject	27.6

Table 5: Verb forms in *V* + *ta* + *NP* constructions.

	Type	Token	Example
Volitional verb	259	934	*da* 'to hit,' *gan* 'to do,' *sha* 'to kill'
Nonvolitional verb	21	40	*bing* 'to fall ill,' *zao* 'to suffer,' *si* 'to die'

Table 6: Monosyllabic and diphthong verbs in *V* + *ta* + *NP* constructions.

	Monosyllable words	Disyllable words
Verbs	91.1%	8.9%
Example	*da* ('hit'), *shui* ('sleep')	*bianlun* ('debate'), *chengzhi* ('punish')

Table 7: Top 10 high-frequency verbs.

Verb	Tokens
da ('to hit')	72
gan ('to do')	49
sha ('to kill')	36
lai ('to do')	33
he ('to drink')	30
gao ('to do')	26
gei ('to give')	22
shui ('to sleep')	22
chi ('to eat')	21
zhuan ('to earn')	14

Table 8: NPs in *V* + *ta* + *NP*.

NP	Ratio of instances (%)	Example
Modified by *ge* (CL)	40	*ge gou* 'CL enough' *ge xibalan* 'CL pieces'
Modified by a numeral	60	*shi nian* 'ten years' *liangbei* 'two glasses'

only nonspecific indefinite NPs are allowed, Lü [35] made an educated guess: when the NP is specific or definite, the pronominal form is interpreted as referential, as in *Ta na-ge ren ting dou* ('He, that guy, is very funny'), where *ta* serves as an attribute of the definite NP *na-ge ren* ('that guy').

(4) Irrealis Contexts of V + ta + NP Constructions. It is generally assumed that expletive *ta* can only occur in an irrealis context (see, e.g., [35–37]), and the verb preceding *ta* cannot take an aspect marker such as *le* or *guo* (see [3, 8, 38]). We found that the above description is essentially correct, as shown in Table 9. However, we also noticed that there are several exceptions, which allow the occurrence of aspect markers.

4.2.3. Characteristics of the Conditional Clause

(1) (bu) Guan as the Conjunction. In conditional clauses, expletive *ta* comes after a conjunction with the meaning 'no matter what,' and the most commonly used conjunction is *(bu)guan*, as shown in Table 10. In Classic Chinese, the conjunction *(bu)guan* functioned as a verb. According to Wu [44], the conjunctive use of *bu-guan* ('no matter what') developed during the Tang and the Five Dynasties. Wang (2008) identified fifteen examples of conjunctive use of *guan* before the expletive subject *ta* in *Zhuzi Yulei* '*Classified Conversations of Zhu Xi*' (a Song Dynasty classic). He proposed that *guan* must have gone through a grammaticalization process to allow the co-occurrence of the expletive subject *ta*. *Guan* was semantically bleached. Moreover, *guan* was located at a sentence-initial position, which is "a preferred locus for processes of grammaticalization" (Auer [45], p. 297).

(2) Three Categories of Predicates in the Conditional Clause. As shown in Table 11, the predicates in the conditional clause containing expletive *ta* can be statistically divided into three categories: constituent containing an interrogative word, coordinate constituent (A-and-B or A-not-A), and a null form.

Interestingly, sentences consisting of interrogative words such as shenme ('what'), shui ('who'), zenme ('how'), sha ('what'), and duoshao ('how many/much') do not express interrogative meanings. Instead, they have a non-interrogative indefinite interpretation or, more precisely, a universal quantification interpretation as shown in example (5) (Li [46] points out that conditionals allow an indefinite *Wh* freely). In (5), *sai shenme* 'compete what' (this is the word-by-word translation) means "any contest," instead of questioning which contest it is.

(5) *Guan ta sai shenme, nimen de huoshi cha buliao.*
No matter what it compete what you DE meal poor not
'No matter what contest you take part in, your meal would not be poor.'

In a coordinated constituent, the coordinated elements provide alternative conditions to choose from. If we take as an example an A-and-B structure, as shown in example (6), A and B are coordinate in form while generic in meaning. In (6), the element A *shi youpai ye hao* ('be it the Right') and the element B *zuopai ye hao* ('be it the Left') provide two opposite conditions to choose from, while no matter under

Table 9: Contexts of expletive *ta*.

Context	Ratio of instances (%)	Example
Irrealis context	98.9	*Wo yao haohao shui ta san-tian* ('I want to have a sleep for three days')
Realis context	1.1	*Women lian tiao-* ***le*** *ta san-tian sanye-de wu* ('We kept dancing for three days') *Piaozhe ta niaoniao chuiyan* ('The smoke is curling up')

Table 10: Conjunctions introducing a conditional clause.

Conjunction	Token
(bu)guan ('no matter what')	368
ren ('no matter what')	2
napa ('no matter how')	2

Table 11: Predicates in the conditional clause of *(bu)guan ta*.

Category	Ratio of tokens (%)
Interrogative form	31
Coordinate form (A-not-A or A-and-B)	35
Null form	27.6
Others	6.4

which condition, the result in the main clause would remain the same.

(6) *Ruguo zhishi guoguo danchun de richang shenghuo de-hua,*
if only live live simple DE everyday life if
guan ta shi youpai ye-hao,zuopai yehao,dui
no matter it be the Right as well the Left as well to
wo laishuo genben meiyou shenme chabie
me for completely not have what difference
'As for everyday life, no matter what I am, the Right or the Left, it does not make any difference to me.'

Moreover, in addition to the above categories, there are some predicates in the null form (about 27.6%), as shown in example (7), some idiomatic expressions such as *san-qi-ershiyi* ('three times seven equals twenty-one') (about 2.3%), and a small number of clausal predicates (as has been pointed out in Chao [1], clausal predicate is prevalent in Chinese).

(7) *Zan dezui ren duo la,guan ta ne.*
We offend people many particle (PAR) no matter it PAR
'We have offended many people, but who cares.'

5. Discussion

As mentioned earlier, the syntactic properties and semantic constraints of expletive *ta* remain controversial in the existing literature, mainly because the scholars conduct a qualitative study and focus only on certain aspects of expletive *ta* and relative constructions. The current data-driven study shows that some observations made by traditional grammarians are fundamentally correct, while others are not. Section 4 shows that the verbs in $V + ta + NP$ constructions are mostly volitional verbs and monosyllabic words and that expletive *ta* tends to occur in irrealis contexts, which is consistent with Ma's [3] and many others' observations. However, we also show that there are a few exceptions such as disyllable words and realis contexts that allow the aspect marker *le* or *zhe,* contrary to the viewpoint of Lin and Zhang [8] and others. Moreover, we found that the NP in the $V + ta + NP$ construction can be an indefinite nominal phrase modified by the classifier *ge*, or by a numeral, rendering untenable Lei and Hu's [15] claim that *ta ge* is a compound word. However, the properties described in the above section are tendencies rather than rigid constraints.

Expletive elements have raised significant issues about the interface between syntax and semantics, particularly the relationship between syntactic subcategorization and theta-role assignment, and the nature of the subject-predicate relation (see [39–41]). Given the subcategorization properties and theta-role assignment abilities of verbs, it is assumed that object positions are always theta-marked, which precludes the occurrence of the non-theta-marked semantically vacuous expletives. This assumption led Rothstein [5–7] to claim that "there can be no such thing as an object pleonastic" based on her structural theory of predication. The central claim of the structural view of predication is that "predication is a syntactic relation which is independent of theta-role assignment holding between a predicate and a nonpredicate (its 'subject')" [6]. She points out that syntactic constituents such as non-theta-marked maximal projections (e.g., VPs, APs, PPs, and other XPs, including some NPs) are predicates. The subjects of syntactic predicates are projected syntactically according to the predication condition (all syntactic predicates must have a subject) and not thematically. In Section 4, we show that Rothstein's structural theory of predication is problematic when considering Mandarin data since 72.4% of the instances of expletive *ta* occur in postverbal positions, namely, the indirect object position.

Theoretically, the occurrence of expletives in subject positions is to satisfy the extended projection principle, which states that every sentence has a subject or the predication condition, which requires that all syntactic predicates must have subjects. Therefore, in pro-drop languages such as Mandarin Chinese, in which subject NPs can be easily dropped, expletive subjects are reasonably unexpected, as hypothesized by Wu and Cao [37]. The fact that subject expletives exist in pro-drop languages such as Mandarin suggests that their occurrences are not only due to

syntactic requirements but also that pragmatic factors may sometimes play a significant role. Chao [1] already noted that the occurrence of expletive *ta* makes the sentence more vivid.

As shown in Section 4.1, expletive *ta* has created significant challenges to automatic syntactic parsing as all instances of *ta* were falsely construed by the Stanford Parser as referential and annotated with false dependencies. After a detailed analysis of the characteristics of the constructions at issue, as shown in Section 4.2, further work is expected to improve the accuracy of automatic syntactic parsing by implementing rule-based, probability-based, or combined methods. Several possible solutions can be tested in future research to identify expletive *ta*, such as formalizing rules according to the linguistic knowledge of Vs and NPs in $V + ta + NP$ constructions and enriching the knowledge base of generalized function words of modern Chinese for information processing (Zan et.al. [42] and Zhang et.al. [43]), considering the conjunction *bu(guan)* in the conditional clauses.

6. Conclusion

We first extracted 1346 instances of Mandarin expletive *ta* constructions from CCL and investigated the automatic dependency parsing results of these constructions, showing that all instances containing expletive *ta* are taken as referential and misrecognized with various dependencies.

We then conducted a detailed study of these constructions, hoping to shed some light on automatic syntactic analyses. Based on a quantitative exploration, we delineated the characteristics of $V + ta + NP$ constructions and *(bu) guan ta* conditional clauses. We showed that verbs in $V + ta + NP$ constructions are mostly volitional and monosyllabic, and NPs are nonspecific indefinite NPs modified either by the classifier *ge* (CL) or a counting word. Semantically, expletive *ta* is basically restricted to an irrealis context. Moreover, in conditional clauses introduced by (*bu) guan* ('no matter what'), expletive *ta* occupies the subject position, and the predicates usually consist of coordinate constituents or interrogative elements or a null form. It is worth noting that these features are tendencies rather than rigid constraints since there are some exceptions in the database. An experiment on testing the automatic syntactic parsing performance after employing the above-described linguistic knowledge of expletive *ta* constructions is expected in our future work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research was funded by the National Social Science Foundation of China (Grant no. 19BYY107).

References

[1] Y. R. Chao, *A Grammar of Spoken Chinese*, University of California Press, Berkeley, CA, USA, 1968.
[2] D. X. Zhu, *Lectures on Grammar,* The Commercial Press, Beijing, China, 1982/2007.
[3] Q. Z. Ma, "Constructions with double objects in Modern Chinese," *Essays on Linguistics*, vol. 10, pp. 166–196, 1983, in Chinese.
[4] J. W. Lin, "Object non-referentials, definiteness effect and scope interpretation," in *Proceedings of North Eastern Linguistic Society 24*, G. Merce, Ed., pp. 287–301, University of Massachusetts, Amherst, MA, USA, 1994.
[5] S. Rothstein, "Pleonastics and the interpretation of pronouns," *Linguistic Inquiry*, vol. 26, pp. 499–529, 1995.
[6] S. Rothstein, *Predicates and Their Subjects*, Kluwer, Alphen aan den Rijn, Netherlands, 2001.
[7] S. Rothstein, "Predication," in *Encyclopedia of Language and Linguistics*, K. Brown, Ed., Elsevier, Amsterdam, Netherlands, pp. 73–76, 2006.
[8] J.-W. Lin and N. Zhang, "The syntax of the non–referential TA 'it' in Mandarin Chinese," *Chinese Language and Linguistics*, vol. 7, pp. 799–824, 2006.
[9] Y. Wu and S. Matthews, "How different are expletive and referential pronouns? A parsing perspective," *Lingua*, vol. 120, no. 7, pp. 1805–1820, 2010.
[10] H. T. Liu, "Data-driven explorations in applied linguistics," *Modern Foreign Languages*, vol. 44, no. 4, 2021, in Chinese.
[11] D. Zhang, Q. L. Zhou, and G. P. Zhang, "Research on completion of verb-object relation in dependency parsing based on linguistic knowledge," *Application Research of Computers*, vol. 25, no. 4, 2018, in Chinese.
[12] J. Xu, "The semantic relationship of co-reference and the syntactic construction of double objects," *Studies of the Chinese Language*, vol. 4, pp. 302–313, 2004, in Chinese.
[13] Y. L. Yuan, "The syntactic and semantic functions of the non-referential pronoun ta," *Grammar Research and Exploration*, vol. 12, pp. 44–64, 2003, in Chinese.
[14] T. Ōta, *A Historical Grammar of Modern Chinese*, Peking University Press, Beijing, China, 1958/2003.
[15] D. P. Lei and L. Z. Hu, "The study on the formation, nature and function of tage," *Linguistic Sciences*, vol. 4, pp. 29–35, 2006, in Chinese.
[16] P. Y. Wang, *The Bleaching of "Guan" and its Relevant Grammar Research,* Hubei University, Wuhan, China, 2008.
[17] Y. Wang and T. Wang, "New interpretation of "chi ta sange pingguo": inheritance and integration approach," *Foreign Languages in China*, vol. 6, no. 4, pp. 22–30, 2009, in Chinese.
[18] S. X. Lü, *800 Words in Chinese,* The Commercial Press, Beijing, China, 1980/2013.
[19] W. Y. Zhang, C. H. Li, Z. M. Zhong, Y. Wang, and L. Li, "Coreference resolution of Chinese personal pronouns with combination of semantics and rules," *Journal of Data Acquisition & Processing*, vol. 32, no. 1, pp. 14–156, 2017.
[20] X. F. Xi and G. D. Zhou, "Pronoun resolution based on deep learning," *Acta Scientiarum Nauralium Universitaties Pekinensis*, vol. 40, no. 1, pp. 100–110, 2014, in Chinese.

[21] S. Chen, D. B. Wang, and S. Q. Huang, "Research on the resolution of personal pronoun in classical books under the digital humanism," *Information Studies: Theory & Application*, vol. 44, no. 10, pp. 165–172, 2021, in Chinese.
[22] F. Kong and G. D. Zhou, "Pronoun resolution in English and Chinese languages based on tree Kernel," *Journal of Software*, vol. 23, no. 5, pp. 1085–1099, 2012, in Chinese.
[23] J. Joyce, "P.G. R. E. P," *Scientific Computing and Instrumentation*, vol. 20, no. 7, pp. 13–51, 2003.
[24] L. Tesnière, *Éléments de Syntaxe Structurale*, John Benjamins Publishing Company, Amsterdam, The Netherlands, 1959.
[25] I. Mel'čuk, "Levels of dependency in linguistic description: concepts and problems," in *Dependency and Valency. An International Handbook Of Contemporary Research*, V. Agel, L. Eichinger, H.-W. Eroms, P. Hellwig, H. J. Herringer, and H. Lobin, Eds., vol. 1, pp. 188–229, W. de Gruyter, Berlin, Germany, 2003.
[26] A. Ninio, *Language and the Learning Curve: A New Theory of Syntactic Development*, Oxford University Press, Oxford, UK, 2006.
[27] R. Hudson, *Language Networks: The New Word Grammar*, Oxford University Press, Oxford, USA, 2007.
[28] H. T. Liu, *Dependency Grammar from Theory to Practice,* China Science Publishing & Media Ltd, Beijing, China, 2009.
[29] D. J. Liu and Z. Y. Zhang, "An investigation into the reliability of automatic dependency parsing tools in Chinese-English cross language studies," *Foreign Language Research*, vol. 6, pp. 9–16, 2021, in Chinese.
[30] H. Liu, "Dependency direction as a means of word-order typology: a method based on dependency treebanks," *Lingua*, vol. 120, no. 6, pp. 1567–1578, 2010, in Chinese.
[31] D. Chen and C. D. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 Conference On Empirical Methods In Natural Language Processing (EMNLP)*, January, 2014.
[32] Q. Z. Ma, "Volitional verb and non-volitional verb," *Journal of Chinese Linguistics*, vol. 3, pp. 157–180, 1992, in Chinese.
[33] M. J. Yuan, "Complements on classification of non-volitional verb," *Studies of the Chinese Language*, vol. 4, pp. 262–268, 1998, in Chinese.
[34] J.-W. Lin, "Object expletives in Chinese," in *Proceedings of the North American Conference On Chinese Linguistics*, University of Delaware, Newark, Delaware, 1993.
[35] S. X. Lü, *Pronouns in Modern Chinese,* Xuelin Publishing House, Shanghai, China, 1985.
[36] M. Kojima, "The Non-referential Uses of Personal Pronouns," Academia Sinica, Taipei, Taiwan, IACL-14 & IsCLL-10, 2006.
[37] Y. Wu and D. Cao, "Object expletives in Chinese and the structural theory of predication," *International Journal of Chinese Linguistics*, vol. 3, no. 2, pp. 179–200, 2016.
[38] R. Iljic, "Empty ta in Mandarin Chinese," *Computational Analyses of Asian and African Languages*, vol. 26, pp. 15–34, 1987.
[39] P. M. Postal and G. K. Pullum, "Expletive noun phrases in subcategorized positions," *Linguistic Inquiry*, vol. 19, pp. 635–670, 1988.
[40] J. M. Authier, "V-governed expletives, case theory and the projection principle," *Linguistic Inquiry*, vol. 22, pp. 721–740, 1991.
[41] I. Hazout, "On the relation between expletive there and its associate: a reply to williams," *Linguistic Inquiry*, vol. 39, no. 1, pp. 117–128, 2008.
[42] H. Y. Zan, K. L. Zhang, Y. M. Chai, and S. W. Yu, "Studies on the functional word knowledge base of Modern Chinese," *Journal of Chinese Information Processing*, vol. 21, no. 5, pp. 107–111, 2007, in Chinese.
[43] K. L. Zhang, H. Y. Zan, Y. M. Chai, Y. J. Han, and D. Zhao, "Survey of the Chinese function word usage knowledge base," *Journal of Chinese Information Processing*, vol. 29, no. 3, pp. 1–8, 2015, in Chinese.
[44] F. X. Wu, *Study on the Grammar of DunHuang Bianwen,* Yue Lu Press, Changsha, Hunan, China, 1996.
[45] P. Auer, "The pre-front field in spoken German and its relevance as a grammaticalization position," *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, vol. 6, no. 3, pp. 295–322, 1996.
[46] Y. H. A. Li, "Indefinite Wh in Mandarin Chinese," *Journal of East Asian Linguistics*, vol. 1, no. 2, pp. 125–155, 1992.