Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Chinese Sentence-Initial Indefinites: What Corpora Reveal

Anna Morbiato
Università Ca' Foscari Venezia, Italia; The University of Sydney, Australia

Abstract While the sentence-initial position in Chinese is generally related to givenness/definiteness, instances of informationally new or indefinite sentence-initial NPs may be found in language in use. This paper systematically explores the phenomenon of sentence-initial indefinites (SIIs), their statistical relevance, and the interaction with features typically connected to linear order, such as animacy or locatability. Results of a quantitative and qualitative analysis conducted on three major big-size, generalised corpora show that SIIs in Chinese are not only possible, but also statistically relevant. Animacy and locatability are found to play a key role in increasing SIIs acceptability. Finally, data reveal a new pattern featuring SIIs with proper nouns.

Keywords Sentence-initial indefinites (SIIs). Chinese. Animacy. Information structure. Corpus study. Quantitative analysis. Qualitative analysis.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-05-11 | Accepted 2020-08-18 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/002

1 Introduction[1]

The sentence-initial position in Chinese is generally associated with, and often defined in terms of, a specific information status, i.e. that of givenness/identifiability and, consequently, definiteness. This association is widely accepted in the literature (Xu 1995) and is supported by the fact that bare nouns in Chinese receive a definite reading when preverbal (1a). Furthermore, it is often maintained that indefinite NPs cannot occur in the sentence-initial position (1b): to be first introduced, indefinites should be preceded by an existential or presentational verb, and then predicated upon, hence the construction in (1c) – all examples from Hole (2012, 61):

1. a. 外国人遇到了张三。
wàiguórén *yùdào-le* *Zhāngsān*
foreigner meet-PFV Zhangsan
'The foreigner met Zhangsan'.

   b. * 一个外国人遇到了张三。
yí *ge* *wàiguórén* *yùdào-le* *Zhāngsān*
one CLF foreigner meet-PFV Zhangsan
'A foreigner met Zhangsan'.

   c. 有一个外国人遇到了张三。
yǒu *yí* *ge* *wàiguórén* *yùdào-le* *Zhāngsān*
exist one CLF foreigner meet-PFV Zhangsan
'A foreigner met Zhangsan'.

In Li and Thompson's grammar, the sentence-initial position is the position for the topic, which "always refers either to something that the hearer already knows about – that is, it is definite – or to a class of entities – that is, it is generic" (1981, 85). Newly-introduced referents cannot be topics, hence they "must follow the main verb of the presentative sentence" (1981, 509), as in (1c). Most subsequent literature on topic-comment structures and word order makes similar observations (Chu 2006; Li 2005; Shyu 2016; Tsao 1977, 1989; Xu 1995; Xu, Liu 2007; Zhu 1982, among others); Ho (1993) holds that the fact that the sentence-initial position should be occupied by a definite el-

1 In this paper, I use the term 'Chinese' to refer to *Pǔtōnghuà*, the standard language of the PRC. Simplified Chinese characters and the *Pinyin* romanisation system have been used throughout the article. The glosses follow the general guidelines of the Leipzig Glossing Rules. Additional glosses include: BEI = 'Chinese 被 *bèi* marker'; COS = 'change of state'; EXP = 'experiential aspect'; MKR = 'marker'; NMLZ = 'nominalizer'; SFP = 'sentence-final particle'; SP = 'structural particle'. I am very grateful to the two anonymous reviewers for their constructive comments and suggestions.

ement "is so strictly adhered to that [...] Chinese has a last resort, which is to prefix a dummy verb 有 *you* [...] to postpone the indefinite NP in the initial position", as in (1c).

However, observations have been raised against the generalisations above. In particular, it has been noted that not all sentence-initial referents are informationally old, i.e. known both to the hearer and to the speaker (Paul 2015): they may be specific – i.e. non identifiable by the hearer – and even indefinite (Bisang 2016; Lu, Pan 2009; Morbiato 2018; Wu 1998). The possibility of indefinites to occur sentence-initially was also stressed by Fan (1985) and subsequent literature by Chinese scholars (Fang 2019; Fu 2013; Liu 2018; Liu, Zhang 2004; Lu, Pan 2009; Tang 2011; Wang 2003; Xu 1997, 1999; Zhang 2007; Zhou, Chen 2013, among others) on so-called 'indefinite-subject sentences' (无定主语句 *wúdìng zhǔyǔ jù*) (see § 2) and is borne out by corpus data:

2. 一位年轻助教谈起了他刚读过一本关于文物保护的著作 [...] (PKUcorpus)
   yí *wèi* *niánqīng* *zhùjiào* *tán-qǐ-le*
   one CLF young teaching.assistant tell-start-PFV
   tā *gāng* *dú-guo* *yì* *běn* *guānyú* *wénwù*
   3SG.M just read-EXP one CLF on cultural.relic
   bǎohù *de* *zhùzuò*
   protection SP work
   'A young teaching assistant started telling he had just read a book on cultural heritage protection'.

This challenges the widely accepted association of the sentence-initial position with topichood, givenness, and definiteness, as well as analyses that postulate a definiteness restriction on the sentence-initial position. However, several aspects of sentence-initial indefinites (henceforth SIIs) in Chinese have not yet been fully explored: how widespread is this phenomenon? How does it interact with other features typically connected to the sentence-initial position (such as animacy and locatability)? Crucially, corpus-based studies on the topic remain the minority and are usually conducted on relatively small, genre-specific corpora.

This paper adopts corpus methodologies and tools to investigate SIIs, with a particular focus on determining (i) the statistical relevance of SIIs of the type of '一 *yī* CLF N' in big-size corpora and (ii) its interaction with the semantic feature of animacy and, secondly, with the referential property of locatability. To this end, it proposes the results of a large-scale, quantitative and qualitative analysis conducted on three major big-size, generalised corpora, namely the PKU CCL corpus (Centre for Chinese Linguistics, Peking University, 470 million characters, henceforth PKU), the BCC corpus of Modern Chinese (Beijing Language and Culture University, 15 billion characters, henceforth BCC), and the Sketch Engine ZHTenTen (Stanford

Tagger) simplified Chinese corpus (13,5 billion characters, henceforth ZHTenTen (ST)). A corpus approach is chosen as it contributes to grounding the analysis on empirical, natural data: corpora allow adhering more to real language in use; moreover, they may help reveal new patterns or phenomena, thus contributing towards deeper and more complete linguistic descriptions even for languages that are over-described, like Chinese.

The rest of the article is organised as follows: § 2 provides an overview of the literature on Chinese SIIs and their characteristics. § 3 presents the study, its research questions, methodology, and linguistic data. §§ 4 and 5 discuss the findings of the quantitative and qualitative analyses, respectively. § 6 draws the conclusions and briefly discusses the implications of such findings on theoretical accounts of the sentence structure of Chinese and onto Chinese as a second/foreign language teaching.

2 (In)definiteness and the Sentence-Initial Position in the Literature

The term 'definiteness' denotes a grammatical category featuring a formal distinction that marks an NP as *identifiable*:[2] this formal distinction may consist of a variety of grammatical means, "including phonological, lexical, morphological, and word order" (Chen 2015, 408). Among the first linguists that associated definiteness with word order in Chinese is Chao, who claims that the encoding of definite/indefinite reference is not much connected to grammatical functions (subject/object): rather, it is the "position in an earlier or later part of the sentence that makes the difference" (1968, 76-7). Crucially, Chao himself proposes a counterexample of SII, of the type of a thetic judgement (3a), commenting that it is a less preferred pattern if compared to the definite>verb>indefinite pattern displayed by (3b):

3. a. 一个卖刷子的在门口呐。

yí	*ge*	*mài*	*shuāzi*	*de*	*zài*	*ménkǒu*	*na*
one	CLF	sell	brush	NMLZ	be.at	door	SFP

b. 门口有一个卖刷子的。

ménkǒu	*yǒu*	*yí*	*ge*	*mài*	*shuāzi*	*de*
door	exist	one	CLF	sell	brush	NMLZ

'A brush peddler is at the door'.

2 Identifiability is an addressee-oriented notion relating to the speaker's assumptions as to whether the addressee "is able to identify the particular entity in question among other entities of the same or different class in the context" (Chen 2015, 408).

Li and Thompson (1981, 167-8) also identify exceptions to their above-mentioned definiteness restriction to the preverbal position, which they illustrate with sentences in (4a)-(4d). All four sentences feature sentence-initial NPs of the type of '一 *yī* CLF N'; however, Li and Thompson hold that such exceptions are only apparent: all sentence-initial NPs in (4) are indeed formally indefinite, but according to them they all receive a definite reading. In (4a), 一 *yī* refers to a specific "absolute quantity" and is therefore definite; in (4b), 一 *yī* in fact means "each", hence, it is not indefinite; in (4c)-(4d), they maintain, 一 *yī* introduces "something that is part of an entity already known by the hearer" (i.e. the leg of a known person, the peasants of a known village) and "can therefore be considered a definite noun phrase":

4. a. 一个人就够了。
yí ***ge*** ***rén*** *jiù* *gòu* *le*
one CLF person then (be).enough PFV/COS
'One person will be enough'.

b. 一个人吃一口。
yí ***ge*** ***rén*** *chī* *yì* *kǒu*
one CLF person eat one mouth
'Each person gets one mouthful'.

c. 一条腿断了。
yì ***tiáo*** ***tuǐ*** *duàn-le*
one CLF leg break-PFV/COS
'One of its legs is broken'.

d. 一个农夫说,"我想出一个办法了"。
yí ***gè*** ***nóngfū*** *shuō* *wǒ* *xiǎng-chū* *yí* *ge* *bànfǎ* *le*
one CLF peasant say 1SG think-exit one CLF way COS
'A peasant said "I've thought of a way"'.

Indeed, the examples above show that not all sentence-initial NPs of the type of '一 *yī* CLF N' are true indefinites. They may emphasise the *quantity* (4a) or receive a *distributive* reading (4b) (see also Lu, Pan 2009). Other readings are possible, e.g. *generic* reference (to a specific class), as in (5) below:

5. 一个年轻人应当有志气。 (Lu, Pan 2009)
yí ***ge*** ***niánqīng*** ***rén*** *yīngdāng* *yǒu* *zhìqì*
one CLF young man should have ambition
'A young man / Young men should be ambitious'.

However, the underlined NPs in (4c)-(4d) can hardly be labelled as definite. In (4c), the implicit body-part (or possession/containment etc.)

relationship might enable the hearer to identify the referent the leg belongs to; however, which specific leg is broken (left/right?) is not identifiable. Similarly, in (4d), 一个农夫 *yí ge nóngfū* 'a peasant' might be assumed to be specific (known by the speaker) but can hardly be considered identifiable by the hearer, especially with no context. On the other hand, the context of these utterances may render the referent *locatable* (Morbiato 2018; Wu 1998), i.e. located within a given/identifiable set (i.e. the two legs) or setting (i.e. the village where the peasant lives; the notion of locatability will be discussed in more depth below). Moreover, none of Li and Thompson's explanations account for Chao's example in (3), a SII *tout court*.

Some scholars put forward a more nuanced view of the definiteness-preverbal position association: Chen (2015, 410), for example, talks about definiteness- and indefiniteness-inclined positions, holding that preverbal NPs are overwhelmingly, but not exclusively, definite. Hole (2012, 61-2), after commenting on (1) that "subject DPs in Chinese must be interpreted as definite", adds that indefinite subjects are barred from the sentence-initial position in *non-thetic* (i.e. all-focus, topicless) sentences, thus implying that SIIs may occur in thetic judgements. However, examples of thetic sentences he includes, such as 一张床睡三个人 *yì zhāng chuáng shuì sān ge rén* 'one bed accommodates three people', do not display an indefinite reading, but rather a distributive one. Lu, Zhang and Bisang (2015) and Bisang (2016) go one step further, arguing that subjects, unlike topics, may be indefinite (they see indefiniteness as a subjecthood test): in thetic sentences, they claim, "preverbal indefinite subjects are acceptable" (Bisang 2016, 356):

6. 一个杯子被我打碎了。[3] (Bisang 2016, 356)
   yí *ge* *bēizi* *bèi* *wǒ* *dǎ-suì-le*
   one CLF cup BEI 1SG hit-break-PFV/COS
   'A cup was broken by me'.

Major contributions to the literature on SIIs come from Chinese scholars. In his influential paper, Fan (1985) notes that SIIs are not only possible, but also rather common in some genres such as news reports: sentences with indefinite subject NPs, he claims, do constitute a sentence pattern in Chinese – they are neither uncommon nor peculiar. Since then, a number of studies have followed (Fang 2019; Fu 2013; Liu 2018; Liu, Zhang 2004; Lu, Pan 2009; Tang 2011; Wang 2003; Xu

3 Note, however, that such a string in Google obtains only 5 results, none of which are thetic sentences (they all have a topic beforehand). A similar string with a third person pronoun 他 *tā* 'he', as in 一个杯子被他打碎了 *yí ge bēizi bèi tā dǎ suì-le* 'a glass was broken by him' gives two occurrences, both of which in grammars that list the sentence as ungrammatical.

1997, 1999; Zhang 2007; Zhou, Chen 2013, among others), mostly focusing on the semantic and syntactic characteristics that license or increase the acceptability of SIIs. Generally, these regard: (i) the type of predicate – highly transitive, dynamic, and stage-level predicates are preferred over low-transitive, stative, and individual-level ones; (ii) the referential characteristics of the SII – the more information is provided that increases the referent's identifiability, the higher the SII's acceptability; and (iii) information structure – thetic sentences may host SIIs, especially when the referent is locatable in clear spatio-temporal frames. In what follows, main contributions will be briefly presented, with particular reference to corpus-based studies.

Several scholars focused on singling out properties and related licensing conditions to SIIs. Tang (2005) holds that SIIs are acceptable only in highly transitive sentences. Zhang (2007) concludes that SIIs occur in topicless (非主题判断 *fēi zhǔtí pànduàn*) – i.e. thetic – judgements, whereby the entire clause is a single unit conveying new information. Lu and Pan (2009) elaborate on this and claim that SIIs occur in (a) thetic sentences, where the whole predicate is projected into the core domain and is constrained by an existence operator, and (b) with stage-level predicates (expressing an event), but not with individual-level predicates (that express some judgement). Chen (2015) also remarks that SIIs are more acceptable with dynamic predicates but hardly occur as subject with stative ones (7):

7. *一个人很聪明。 (Chen 2015, 410)

yí	*ge*	*rén*	*hěn*	*cōngming*
one	CLF	person	very	smart

'One person is very smart'.

With reference to the above considerations, Wang (2003), Huang (2004), Wei and Chu (2007), and Lu and Pan (2009), among others, put forward a number of corollary licensing conditions to SIIs – e.g. SIIs cannot occur with modal verbs, negative adverbs, and tense. However, corpus studies found that most of these conditions are only tendencies, as counterexamples can be found for each parameter. Specifically, Zhou and Chen (2013) measured the descriptional accuracy of such licensing conditions with the method of parameter setting and measurement against a relatively small test corpus (i.e. a 1,000-sentence subcorpus of the PKU). From their analysis, it appears that all factors indeed contribute through a complex interplay to increasing SII's identifiability, and hence acceptability rate, but none constitutes an absolute restriction.

A widely accepted generalisation on SIIs is that the greater the amount of information on the referent (e.g. by means of longer nominal modifiers), the higher its degree of identifiability and, hence, its acceptability (Xu 1999). Wang (2003), for example, talks about degree

of (cognitive) *accessibility* (可及度 *kějídù*) and of *identifiability* (个体化程度 *gètǐhuà chéngdù*). Indeed, the acceptability difference between (8a) and (8b) lies in the long, informationally-rich (complex relative clause plus noun) modifier of the SII:

8. a. *一种方法最近问世。 (Zhou, Chen 2013, 373)
yì zhǒng fāngfǎ zuìjìn wènshì
one CLF method recently come.out
'A method was recently introduced'.

b. 一种取几根头发就可准确断定被检测者是不是吸毒者的检毒方法最近问世。
yì zhǒng qǔ jǐ gēn tóufǎ jiù kě zhǔnquè
one CLF pick some CLF hair then can accurately
duàndìng bèijiǎncèzhě shì-bú-shì xīdúzhě de
determine subject be-NEG-be drug.addict SP
jiǎndú fāngfǎ zuìjìn wènshì
detection method recently come.out
'A hair drug test for accurately determining whether a subject is a drug addict has recently come out'.

A very interesting perspective is provided by Fu's (2013) corpus-based, diachronic study, which reveals that SIIs very likely originated during the Song Dynasty (960-1279) and evolved from earlier constructions whereby an indefinite NP is the subject of the sentence following a perceptual verb, like 见 *jiàn* 'see'. Early instances of 'see' + indefinite NP patterns – e.g. (9) from *Zhuangzi* – also specify the scene witness (the <seer>, in this case King Wen). Later, the construction became impersonal, by means of markers that express the idea of 'seeing', such as 只见 *zhǐjiàn* and 则见 *zéjiàn*: sentences like (10) are interpreted as if the witness were an omniscient narrator. Later, these markers disappeared (11) (all examples are from Fu 2013):

9. 文王观于臧, 见一丈人钓 […] (*Zhuangzi*, *Tianzifang*)
Wén wáng guān yú Zāng ***jiàn*** *yí zhàngrén diào*
Wen king look SP Zang **see** one man fish
'King Wen was (once) looking about him at Zang, when he saw an old man fishing […]'[4]

4 Translation source: the *Chinese Text Project* (https://ctext.org).

10. 两边人犹未散, 只见一个庄客在东边墙角下叫道 […] (*Stories to Awaken the World*, 1627)

liǎng-biān	*rén*	*yóu*	*wèi*	*sàn*		
two-part	people	still	NEG	scatter		
zhǐjiàn	*yí*	*ge*	*zhuāngkè*	*zài*	*dōng-bian*	*qiángjiǎo*
MKR	one	CLF	farm.worker	at	east-side	corner
xià	*jiàodào*					
under	say					

'The people had not yet scattered; a farm worker at the east corner said […]'

11. 正说处, 一个小和尚点了灯来请洗澡。 (*Journey to the West*, § 62)

zhèng	*shuōchù*	*yí*	*ge*	*xiǎo*	*héshang*	*diǎn-le*	*dēng*
right	say.out	one	CLF	little	monk	light-PFV	lamp
lái	*qǐng*	*xǐzǎo*					
come	invite	shower					

'As they were talking, a young monk came in to light the lamp and invite Sanzang to take his bath'.[5]

Locatability. From the data in the literature analysed so far, an important feature of SIIs that scholars, however, never explicitly mention seems to be locatability, intended as identifiability of the referent's setting rather than identifiability of the referent itself. An example of non-identifiable, locatable referent is the sentence-initial NP in *a person in the airplane started shouting*: the hearer (and even the speaker) might not know who this person is, but they are definitely able to locate the referent within the group of people on that specific airplane. In other words, the referent itself is not identifiable: what can be identified is the scene/setting/set/frame where the referent is located. Locatability is typically granted by the presence of a phrase that expresses a temporal or spatial frame for the utterance, which is an inherent characteristic of Chinese topics (Chafe 1976; Her 1991; Morbiato 2018; Paul 2015) and is the property Li and Thompson tried to recall with respect to (4c)-(4d): the referents are not identifiable/definite, but rather locatable within a known set – one of two legs of an individual in (4c) – or a temporal/spatial setting – one of the peasants of a known village in (4d). This also suggests that locatability, rather than givenness and identifiability, is a more accurate restriction to the preverbal position in Chinese (see Morbiato 2018, 2020 for discussion). This is confirmed by Liu and Zhang's (2004) corpus investigation of eight novels and children stories: most (although no statistics are provided) of the SIIs they detected feature a temporal or spatial reference occurring before the indefinite NP. Such tem-

5 Translation from 'Internet archive' (`https://bit.ly/3pu33AZ`).

poral or spatial reference situates the referent within identifiable spatio-temporal coordinates. It may be either a phrase (12) or a sentence/clause (13). Other sentences may feature no explicit temporal reference, but according to Liu and Zhang (2004, 99) "从上下文中, 可以明显看出指的就是'正在此时'的意思" (the context allows the identification of the reference time as 'just now' [Author's translation]). In other words, they have an implicit *stage topic*.[6]

12. 1990年11月, 一份诉状递到了北京市西城区人民法院。
yījiǔjiǔlíng *nián* *shíyī* *yuè* (SPATIO-TEMPORAL FRAME)
1990 year 11 month
yí *fèn* *sùzhuàng* *dìdào-le* *Běijīng* *shì* *Xīchéng*
one CLF complaint submit-PFV Beijing city Xicheng
qū *Rénmín* *Fǎyuàn*
district People Court
'In November 1990, a complaint was submitted to the People's Court of Xicheng District, Beijing'.

13. 正在审问的时候, 一只大老虎跳进公堂 […]
zhèngzài *shěnwèn* *de* *shíhòu* (SPATIO-TEMPORAL FRAME)
PROG interrogate SP time
yì *zhǐ* *dà* *lǎohǔ* *tiáo-jìn* *gōng-táng*
one CLF big tiger jump-enter public-hall
'During the interrogation, a big tiger jumped into the public hall […]'

An account in terms of locatability also explains Xiong's (2008) claim that SIIs admissibility depends on the presence of a specific component that meets the topic's needs: what Xiong actually means is that some contextual element is needed that renders the topic referent locatable; such an element may be a temporal/locative phrase, even an implicit one (stage topic). It also sheds light on Liu's (2003) observation that the role of SIIs within the narration is to create a plot transition: in this case, the new topic also involves a shift of setting (for example, a new scene or a new time reference, with different spatio-temporal coordinates).

All the above studies highlight significant features of SIIs. However, they reveal little about their statistical relevance, as most corpus-based studies are qualitative and/or conducted on small-size corpora. Furthermore, little is said on another rather significant cross-linguis-

6 Given an utterance, stage topics are its implicit spatio-temporal coordinates that allow the assessment of its truth value. This captures the fact that a sentence like *it is snowing!* is true and informative only with reference to the temporal and spatial setting of its discourse. According to Erteschik-Shir, "thetic sentences are viewed as having implicit 'stage' topics indicating the spatio-temporal parameters of the sentence (here-and-now of the discourse). These are contextually defined" (2007, 16).

tic feature of the sentence-initial position, i.e. *animacy*: does this semantic trait interact at all with SIIs in Chinese?

3 The Study. What Corpora Tell on SIIs

As said earlier, this study adopts a corpus approach, with the aim to ground the analysis on empirical, natural data. Specifically, corpora contribute towards: (i) verifiability and reproducibility as monitoring mechanisms for a given analysis, as results can be checked by repeating the same query; and (ii) highlighting facts, data, or details that had not been observed before and have not yet been integrated in linguistic descriptions. Let us now turn to corpus data: a banal query with the string '。一位' (. *yí wèi*) in the PKU corpus gives 5,751 results; the first 5 occurrences are reported in table 1. The same query gives 1,466 results in the BCC corpus and 605,379 in the ZHTenTen (ST) corpus. On the other hand, the string '。一个' (. *yí ge*) occurs 13,399 times in the PKU corpus; the first 5 occurrences are shown in table 2.

Table 1 PKU corpus: first 5 occurrences of the string '。一位' (. *yí wèi*)

[…] 两位具有马克思主义传统的欧洲思想家	。一位	是意大利共产党领导人和理论家安东尼·葛兰西, 另一位是 […]
liǎng wèi jùyǒu Mǎkèsī-zhǔyì chuántǒng de Ōuzhōu sīxiǎngjiā	*. yí wèi*	*shì Yìdàlì Gòngchǎndǎng lǐngdǎorén hé lǐlùnjiā Āndōngní Gélánxī, lìng yí wèi shì*
[…] two European thinkers within the Marxist tradition	. One	is the leader and theoretician of the Italian Communist Party, Antonio Gramsci, the other is […]
当时有两位大史学家 […]	。一位	是黄梨洲, 他著了一部《明夷待访录》[…]
dāngshí yǒu liǎng wèi dà shǐxuéjiā	*. yí wèi*	*shì Huáng Lízhōu, tā zhù le yí bù Míngyí Dàifǎng Lù*
At that time, there were two great historians […]	. One	is Huang Lizhou, who wrote the *Mingyi Daifang Lu* […]
[…] 这就是哲学家康德和他的仆人拉普	。一位	传记家赞叹道: "康德的一生就像是一个最规则的动词 […]
zhè jiù shì zhéxuéjiā Kāngdé hé tā de púrén Lāpǔ	*. yí wèi*	*zhuànjìjiā zàntàn dào: "Kāngdé de yìshēng jiù xiàng shì yí ge zuì guīzé de dòngcí*
[…] these are the philosopher Kant and his manservant Lampe	. A	biographer said admiringly: "Kant's life is like a regular verb […]
这项研究已经成为社会学学术进展的一个很重要的组成部分	。一位	著名的美国社会学家就认为, 这方面的研究已经不是在与主流社会学 […]
zhè xiàng yánjiū yǐjīng chéngwéi shèhuìxué xuéshù jìnzhǎn de yí ge hěn zhòngyào de zǔchéng bùfèn	*. yí wèi*	*zhùmíng de Měiguó shèhuìxuéjiā jiù rènwéi, zhè fāngmiàn de yánjiū yǐjīng búshì zàiyǔ zhǔliú shèhuìxué*
This research has already become a milestone in the field of sociology	. A	well-known American sociologist holds that research in this area no longer lies within mainstream sociology […]

这篇文章讲的是一个动人的故事	。一位	名叫苏珊·斯蒂芬的母亲, 愿为她患肾炎的儿子捐出一个肾。
zhè piān wénzhāng jiǎng de shì yí ge dòngrén de gùshì	*. yí wèi*	*míng jiào Sūshān Sīdìfēn de mǔqīn, yuàn wéi tā huàn shènyán de érzi juānchū yí ge shèn*
This piece of writing tells a moving story	. A	mother named Susan Stephen is willing to donate a kidney to her son who suffers from nephritis.

Table 2 PKU corpus: first 5 occurrences of the string '。一个' (*. yí ge*)

[…] 社会正在进行一场新技术革命	。一个	国家生产力的发展, 国民经济的增长, 越来越依靠科学技术的进步 […]
shèhuì zhèngzài jìnxíng yì chǎng xīn jìshù gémìng	*. yí ge*	*guójiā shēngchǎnlì de fāzhǎn, guómín jīngjì de zēngzhǎng, yuè lái yuè yīkào kēxué jìshù de jìnbù*
[…] society is undergoing a new technological revolution	. A	country's productivity development and the growth of its national economy rely more and more on the progress of science and technology; […]
[…] 就是强调学校教育工作的时效性。(5)持久性	。一个	人所受的从幼儿园开始到大学的教育, 要经历17-18年的时间 […]
jiùshì qiángdiào xuéxiào jiàoyù gōngzuò de shíxiàoxìng (5) chíjiǔxìng	*. yí ge*	*rén suǒ shòu de cóng yòu'éryuán kāishǐ dào dàxué de jiàoyù, yào jīnglì 17-18 nián de shíjiān*
[…] it emphasises the timeliness of school education. (5) *Persistence.*	. A	person's education from kindergarten to university takes 17-18 years[…]
这是不少学者专家的共识	。一个	人, 作为生命个体, 从出生之日起, 就与周围环境 […]
zhè shì bù shǎo xuézhě zhuānjiā de gòngshì	*. yí ge*	*rén zuòwéi shēngmìng gètǐ, cóng chūshēng zhī rì qǐ, jiù yǔ zhōuwéi huánjìng*
[…] it is the internal driving force of individual development. This is the general consensus among several scholars and experts	. A	person, as an individual form of life, from the date of her birth, clashes with the surrounding environment […]
[…] 所谓自由、责任、义务, 都是幻想的名词	。一个	人对于社会的有用与否, 完全看遗传如何。
suǒwèi zìyóu, zérèn, yìwù, dōu shì huànxiǎng de míngcí	*. yí ge*	*rén duìyú shèhuì de yǒuyòng yǔ fǒu, wánquán kàn yíchuán rúhé*
[…] so-called freedom, responsibility, and obligation are all fantasy terms	. (Whether) a	person is useful to society depends entirely on her inheritance.
香港的幼儿教育(又称为学前教育)分为两个系统	。一个	是香港政府教育署管辖的幼稚园, 另一个系统是 […]
Xiānggǎng de yòu'ér jiàoyù (yòu chēng wéi xuéqián jiàoyù) fēn wéi liǎng gè xìtǒng	*. yí ge*	*shì Xiānggǎng zhèngfǔ jiàoyù shǔ guǎnxiá de yòuzhìyuán, lìng yí ge xìtǒng shì*
Early childhood education (also known as preschool education) in Hong Kong is divided into two systems	. One	consists of the kindergartens under the jurisdiction of the Education Department of the Hong Kong Government; the other system is […]

Such very preliminary data have little statistical relevance but open up interesting perspectives. First, SIIs do exist and are not statistically insignificant: results in all corpora are of the order of thousands; moreover, five out of five sentences in table 1 present sentence-initial NPs that receive a true indefinite reading. Second, corpora are tools that must be used *cum grano salis*: in table 2, the first four NPs are in fact generic, while only the fifth is a true indefinite. Hence, quantitative data will need to be filtered through a subsequent qualitative examination, to assess the extent to which sentence-initial NPs of the type of '一 *yī* CLF N' are true indefinites. Third, a striking difference is highlighted between a very common, generic classifier like 个 *ge* 'unit' and the highly specific classifier 位 *wèi*, i.e. the polite classifier for people: although 个 *ge* is much more frequent in absolute terms (its total occurrences as classifier in the ZHTenTen (ST) corpus is 9,265,680, as compared to 1,007,191 for 位 *wèi* – see table 3 below), the former occurs just little above twice as the latter in the '。一 *yī* CLF' pattern. This, together with the different ratio of true SIIs (100% *vs* 20%, respectively), suggests that the semantics of the classifier (e.g. the trait ±animate/±human) might also be relevant with respect to the acceptability degree/statistical relevance of SIIs. This hypothesis is supported by the cross-linguistic tendency of animate NPs to occur sentence-initially, regardless of their semantic role, syntactic function, and information status (non-agent, non-subject, and non-given animates still display this tendency). An experimental study carried out by Verhoeven on a sample of heterogeneous languages (German, Greek, Turkish, and Chinese) shows that "animate-first effects occur across languages" (2014, 129). This, according to Verhoeven, is an expected result under the view that "these effects come from asymmetries in the mental representation of the referents", which are independent from language-specific characteristics (2014, 129) – see also Van Bergen (2011) for a cross-linguistic overview of animacy and word order and Iemmolo and Arcodia (2014) for Chinese.

3.1 Research Questions and Scope

Against the background laid out so far, this study aims at answering the following research questions:

> RQ1 How significant is the phenomenon of SIIs from a quantitative/statistical perspective?
>
> RQ2: Does the trait of animacy play a role in the phenomenon?

The study focuses on indefinite NPs marked through the major indefiniteness encoding means in Chinese (Chen 2015, 409), i.e. a noun

phrase containing the string 一 *yī* 'one' + classifier (CLF),[7] that occurs sentence-initially. In fact, indefiniteness may be conveyed, more in general, by the string numeral + classifier (Li 1997, 18, among many others); however, indefinite NPs with numerals other than 一 *yī* 'one' (e.g. 三/几个学生 *sān/jǐ ge xuéshēng* 'three/some students') are excluded from the study, for two main reasons: the first is that the study itself would be more complex in terms of corpus queries; moreover, it would involve relying more on the accuracy of the tagging, which is not always high (see discussion in § 6) and is different in each corpus (e.g. the PKU is not POS-tagged), thus not allowing a comparison between the three corpora. Finally, numerals other than 'one' often emphasise the *quantity* or receive a *distributive* reading, as discussed by Li and Thompson with reference to (4a)-(4b) above, while the focus here is mainly on true indefinite readings. This implies that this study only accounts for singular indefinite NPs of the type of '一 *yī* CLF (N)' and that the number of SIIs identified in this study is smaller than those actually existing in the corpora.

Possible indefinite NPs may consist of simple patterns of the type of '一 *yī* CLF (N)', where the head noun may be overt or omitted. In some cases, the classifier may also be omitted; however, these cases are comparatively rarer and harder to detect, and thus will not be considered. This also implies that, again, the number of SIIs identified in this study is smaller than those existing in the corpora. Indefinite NPs may also include modifiers (nouns, adjectives, verbs, relative clauses etc.). These generally occur in two positions: between the classifier and the noun (14b) and to the left of the '一 *yī* CLF N' string (14c) – the former suggests a descriptive reading, the latter a restrictive one, see e.g. Chao (1968, 286-7):

14.	a.	[Numeral + CLF]		[Noun]
	b.	[Numeral + CLF]	[Modifier(s)]	[Noun]
	c. [Modifier(s)]	[Numeral + CLF]		[Noun]

Below are examples of SII types above. For pattern (14c), modifiers may include nouns/adjectives (15c), but also verbal elements occurring, for example, within a relative clause (15c'). Finally, other elements, such as time/location phrases, may occur to the left of the NP – see e.g. (12) above:

7 Indefinite NPs in Chinese may take two forms: nouns modified by a number + classifier structure and bare nouns, when postverbal (Li 1997, 18). Since the present article investigates the sentence-initial position, it focuses on the pattern '一 *yī* CLF N'.

15. a. 一位传记家赞叹道 […] (PKU)
yí *wèi* *zhuànjìjiā* *zàntàn-dào*
one CLF biographer admire-say
'A biographer said admiringly […]'

b. 一位著名的美国社会学家就认为 […] (PKU)
yí *wèi* *zhùmíng* *de* *Měiguó* *shèhuìxuéjiā*
one CLF famous SP American sociologist
jiù *rènwéi*
indeed think
'A famous American sociologist thinks that […]'

c. 加油站的一位工作人员说, 从下午三四点钟开始 […] (ZHTenTen (ST))
jiāyóuzhàn *de* *yí* *wèi* *gōngzuòrényuán* *shuō*
gas.station SP one CLF worker say
cóng *xiàwǔ* *sān-sì* *diǎnzhōng* *kāishǐ*
from PM 3-4 o'clock start
'A staff member of the gas station said that from 3-4 PM onwards […]'

c'. 刚来的一位天津大厨 […] (Wangyi News)[8]
gāng *lái* *de* *yí* *wèi* *Tiānjīn* *dàchú*
REL [just come SP] one CLF Tianjin chef
'A newly arrived chef from Tianjin […]'

3.2 Methodology and Data

Quantitative analysis. Identifying SIIs as described above involves examination of complex strings, including punctuation and sentence boundaries. Hence, for the quantitative analysis, three generalised, big-size corpora were chosen that allow such a query: the PKU corpus (470 million characters), the BCC corpus (15 billion characters), and the ZHTenTen simplified Chinese corpus mounted at Sketch Engine (Stanford Tagger subcorpus, 1,73 billion characters). Each corpus involves a different query system, and only the BCC and the ZHTenTen (Stanford Tagger, henceforth ST)[9] are POS-tagged; hence, the results are more or less fine-grained depending on the corpus. Specifically, while the BCC and the ZHTenTen (ST) corpora also allow queries through the POS tag for classifiers (*q* and *M*, respectively), in the

8 https://bit.ly/37wXhFe.

9 The ZHTenTen Stanford Tagger is POS tagged following the Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank. The corpus allows a rather detailed interrogation, lends itself to concordancing, collocation, and term extraction (Xu 2015).

PKU corpus the number of occurrences needs to be collected for each single classifier. To this end, Sketch Engine's wordlist tool was used to obtain a frequency list of the nominal classifiers listed in the 汉语量词词典 *Hanyu liangci cidian* (Chen et al. 1988): a total of 36 classifiers with more than 20 thousand occurrences as classifier in the ZHTenTen (ST) were identified. Units of measure, e.g. 元 *yuán* (RMB), 分 *fēn* (unit of length/area/money/time), 吨 *dūn* (ton), 亩 *mǔ* (unit of area), 公里 *gōnglǐ* (km) were excluded, in that they are mainly used to express specific quantities rather than indefiniteness. To tackle RQ2 (§ 3.1), particular attention was devoted to classifiers denoting animate nouns – marked as +A(nimate) – including 名 *míng*, 位 *wèi*, 只 *zhī* and 头 *tóu* (for animals), and 伙 *huǒ* (collective). Other classifiers used with people but also with inanimate nouns (±A) such as 个 ge, 行 *háng* (row), 家 *jiā* (for families and for shops), and 排 *pái* (line) were treated separately, as it is not possible to verify whether their frequency is connected with the occurrence of animate nouns. The classifier 对 *duì* 'couple', while compatible both with animates and inanimates, was marked as +A, in that a cursory examination of 150 random tokens of sentence-initial '一对 *yí duì*' NPs in all three corpora reveals that 90% of tokens introduce animate nouns. Table 3 shows the resulting list of examined classifiers, along with their frequency:

Table 3 List of classifiers

CLF	Animacy trait	Frequency as classifier in the ZHTenTen (ST) c.	CLF	Animacy trait	Frequency as classifier in the ZHTenTen (ST) c.
个	±A	9,265,680	座	–A	194,739
项	–A	1,458,480	本	–A	182,384
名	+A	1,156,327	系列	–A	174,548
条	±A	1,104,219	台	–A	174,530
位	+A	1,007,191	只	+A	164,721
级	–A	858,424	户	–A	160,875
家	±A	807,627	门	–A	114,744
批	±A	461,612	组	±A	105,680
件	–A	407,054	处	–A	104,857
份	–A	340,977	道	–A	85,349
期	–A	329,997	首	–A	81,823
所	–A	293,366	把	–A	79,768
篇	–A	278,140	对	+A	79,199
套	–A	260,345	班	±A	71,086
句	–A	234,465	间	–A	68,961
部	–A	216,625	头	+A	33,993
张	–A	214,591	排	±A	16,522
块	–A	208,768	伙	+A	6,596

For patterns (a) and (b) in (14), the string '一 *yī* CLF' is at the beginning of the sentence and can be easily detected with the appropriate syntax (i.e. (; |:|。|? |!)$一CLF in the PKU corpus; [。 ; ? !]一q/CLF in the BCC corpus; and <s> [word="一"][tag="M"] and <s> [word="一"][word="CLF"] in Sketch Engine). On the other hand, detection of pattern (c), where the modifier(s) occur(s) between the punctuation mark and the '一 *yī* CLF' string, is more complex and, in some cases, problematic. Specifically, modifiers such as relative clauses cannot be detected, as queries including verbs before the '一 *yī* CLF' string may both identify SIIs, as (15c'), but also postverbal indefinites, as in the following example:

16. 刚来了一位天津大厨

gāng	*lái-le*	*yí*	*wèi*	*Tiānjīn*	*dàchú*
just	arrive-PFV	one	CLF	Tianjin	cook

'A cook from Tianjin has just arrived'

To avoid that, the queries exclude verbal elements, but include adjectival and nominal modifiers (e.g. <s>[tag="JJ"][tag="N.*"]{1,7}[word="一"][word="CLF"&tag="M"], in the ZHTenTen (ST)). Finally, SIIs with leftmost time/location phrases separated by commas, as in (12), are hard to identify quantitatively and are not considered either. Again, this implies that the number of SIIs identified in the quantitative analysis does not include all possible patterns.

Qualitative analysis. As noted in § 2, while the string '一 *yī* CLF' is the most common formal marker for Chinese indefinite NPs, it does not always involve a true indefinite meaning, as the NP may display a quantitative (4a), distributive (4b), or generic (5) reading. The quantitative analysis as described above necessarily identifies all types, as they are formally identical. To determine the average ratio of true indefinites, as well as of NPs receiving a quantitative, distributive, or generic reading, a qualitative analysis was conducted on a random sample of sentences from the ZHTenTen (ST) corpus, collected[10] with the following query: <s>[tag="JJ|N.*"]{0,7}[word="一"][word="CLF1| CLF2|… "]. Each sample consists of 100 sentences for each subtype of classifiers (+A, ±A, –A), for a total of 300 sentences, a number that preserves the representativeness of the sample.

10 With the Sketch Engine function 'get a random sample', the same number of lines generated from a given concordance produces the same concordance lines: thus, the search can be easily repeated and reproduced.

4 Quantitative Results

The tables below show results for each corpus. In the paper, 'CLF' denotes each specific classifier, while 'CLF' indicates the word class. S.I. stands for 'sentence-initial', while *de* corresponds to the Chinese noun modifier marker 的 *de*, which may but need not be present. Orange, blue, and green mark +A, ±A, and –A classifiers, respectively (see § 3.2). Columns for pattern (c) as shown in (14) report figures of different modifiers patterns; the type and number of detectable patterns depend on the tools and CQL queries each corpus offers. The last column (ratio) shows the percentage of sentence-initial occurrences of each classifier in the pattern '一 *yī* CLF' over all occurrences of the pattern in any position in the sentence; in other words, it captures how often an indefinite noun phrase with a specific classifier occurs sentence-initially.

Table 4 ZHTenTen (ST) corpus

CLF	Any position	Patterns (a) – (b)	Pattern (c): S.I."一 *yī* CLF" occurrences with						All patterns			Ratio
	'*yī* CLF'	S.I. '*yī* CLF'	leftmost noun mod.	leftmost noun mod. +*de*	leftmost adj. mod.	leftmost adj. mod. +*de*	leftmost adj./ noun mod.	leftmost adj./noun mod.+*de*	Total detected without *de*	Total detected with *de*	Total detected	S.I. '*yī* CLF'/ '*yī* CLF'
名	207,535	6,035	878	190	58	2	52	13	8,005	619	8,624	3.48%
位	300,812	27,182	2,351	907	142	4	103	0	33,425	2,732	36,157	10.20%
只	64,460	1,887	205	22	35	3	15	1	2,228	56	2,284	3.36%
头	15,569	424	112	20	4	1	12	2	681	36	717	3.69%
伙	3,633	83	17	1	1	0	4	0	192	1	193	2.92%
对	40,725	1,065	151	26	32	2	13	1	1,427	57	1,484	3.17%
个	3,923,883	98,525	5,101	2,432	1,957	189	321	509	110,524	5,497	116,021	2.78%
条	204,214	2,575	437	99	119	35	24	6	3,397	209	3,606	1.61%
家	197,900	3,938	714	63	54	8	83	9	7,893	361	8,254	2.46%
批	253,206	2,841	342	24	96	6	48	6	3,578	90	3,668	1.33%
组	32,120	710	184	29	102	6	17	3	1,112	83	1,195	3.27%
班	11,040	150	152	0	51	0	6	0	419	3	422	3.25%
排	6,649	113	36	13	13	0	2	2	171	18	189	2.69%
项	236,816	3,059	431	313	208	17	17	46	4,401	1,272	5,673	1.73%
级	116,584	2,231	1,733	37	74	4	102	8	4,856	77	4,933	3.59%
件	115,770	1,360	142	37	98	10	7	4	1,668	61	1,729	1.43%
份	164,759	2,487	225	113	46	6	9	7	2,989	523	3,512	1.76%
期	52,922	2,259	1,053	14	274	0	159	6	5,324	47	5,371	7.11%
所	61,758	1,583	165	5	43	0	7	1	2,166	44	2,210	2.92%
篇	64,164	1,224	248	27	64	2	7	2	1,626	126	1,752	2.45%
套	105,632	956	0	18	41	7	20	7	1,221	50	1,271	0.99%
句	113,840	3,728	252	136	251	38	12	17	4,458	407	4,865	3.89%
部	73,383	2,252	195	11	37	2	18	2	2,651	60	2,711	3.43%
张	89,515	1,737	202	23	54	8	4	0	2,088	50	2,138	2.27%
块	74,084	816	141	26	38	7	6	3	1,060	58	1,118	1.40%
座	71,757	1,324	106	36	23	3	12	2	1,579	76	1,655	2.10%

本	68,435	1,536	158	10	65	4	9	1	1,890	43	1,933	2.61%
系列	163,248	1,750	50	104	61	7	3	15	1,940	250	2,190	1.22%
台	42,980	828	95	7	32	2	7	1	1,074	21	1,095	2.26%
户	13,017	135	46	1	1	0	4	1	215	7	222	1.44%
门	47,792	456	22	5	24	1	4	0	587	9	596	1.07%
处	37,630	190	199	8	12	0	9	2	518	30	548	1.12%
道	48,975	493	63	15	53	0	6	1	648	33	681	1.29%
首	36,766	1,256	122	25	42	3	15	5	1,525	139	1,664	3.99%
把	63,835	742	386	22	29	1	25	4	1,365	39	1,404	1.89%
间	21,696	385	100	24	7	2	10	2	547	50	597	2.44%

Thanks to the Corpus Query Language (CQL) option, the ZHTenTen (ST) is the corpus that allowed extraction of the most detailed data. Table 4 presents the number of occurrences for each classifier for patterns (14a)-(14b) (column 3) and some possible patterns for (14c), distinguishing different modifier types (adjective, noun, or both, and with or without 的 *de*); modifiers are up to 7 characters long. Columns 10 and 11 show the total amount of detected S.I. '一 *yī* CLF' patterns that occur without and with 的 *de*, respectively,[11] while column 12 (total detected) provides the sum of these two. The classifier with the highest total occurrences in the three patterns identified in (14) is 个 *ge* (116,021), followed by 位 *wèi* (36,157 – about one third). However, an inverse tendency is observable in the last column, which again captures how often an indefinite noun phrase with a specific classifier occurs sentence-initially: the classifier where this ratio is by far the highest is 位 *wèi* (more than 10%); other +A classifiers are all around 3%, followed by 个 *ge* that drops to 2.78%.

Table 5 BCC corpus

CLF	Any position	Patterns (a) – (b)	Pattern (c): S.I. '*yī* CLF' occurrences with						All patterns	Ratio
	'*yī* CLF'	S.I. '*yī* CLF'	leftmost noun mod.	leftmost noun mod. *+de*	leftmost adj. mod.	leftmost adj. mod. *+de*	leftmost adj./ noun mod.	leftmost adj./noun mod.*+de*	Total detected	S.I. '*yī* CLF'/ '*yī* CLF'
名	5,252	236	4	3	0	0	0	0	243	4.49%
位	29,484	1,673	26	12	4	2	0	0	1,717	5.67%
只	34,460	1,209	34	21	1	3	4	0	1,272	3.51%
头	8,676	161	20	26	2	1	2	0	212	1.86%
伙	1,540	83	2	0	0	0	0	0	85	5.39%
对	6,019	223	12	3	4	2	0	0	244	3.70%
个	351,862	14,327	275	53	37	27	13	2	14,734	4.07%

11 Used queries include: <s>[tag="JJ|N.*"]{0,7}[word="一"][word="CLF"] and <s>[tag="JJ|N.*"]{0,7}[word="的"] [word="一"][word="CLF"], respectively.

条	30,059	673	18	6	1	3	1	0	702	2.24%
家	14,639	329	44	4	1	1	3	0	382	2.25%
批	2,602	26	11	0	0	1	0	0	38	1.00%
组	690	15	2	0	0	0	0	0	17	2.17%
班	690	150	3	0	1	0	0	0	154	2.17%
排	3,131	91	5	2	0	0	0	0	98	2.91%
项	2,323	16	1	0	0	0	0	0	17	0.69%
级	782	6	2	0	1	0	0	0	9	0.77%
件	25,072	267	18	1	0	1	0	0	287	1.06%
份	6,353	42	2	0	0	0	0	0	44	0.66%
期	277	0	1	0	1	0	0	0	2	0.00%
所	2,613	16	2	0	0	0	0	0	18	0.61%
篇	3,916	41	0	1	0	0	0	0	42	1.05%
套	5,195	33	1	1	2	1	0	0	38	0.64%
句	24,806	376	25	1	4	2	0	0	408	1.52%
部	9,793	206	7	18	0	0	0	0	231	2.10%
张	23,339	519	16	5	6	0	1	0	547	2.22%
块	23,430	268	13	1	4	3	0	0	289	1.14%
座	10,229	222	2	1	4	1	1	0	231	2.17%
本	9,240	108	5	0	0	2	1	0	116	1.17%
系列	874	15	0	0	0	0	0	0	15	1.72%
台	988	31	1	0	0	0	0	0	32	3.14%
户	405	5	0	0	0	0	0	0	5	1.23%
门	1,195	9	2	0	0	0	0	0	11	0.75%
处	4,522	30	0	0	1	0	0	0	31	0.66%
道	10,229	275	10	5	0	0	0	0	290	2.69%
首	2,855	37	0	0	0	0	0	0	37	1.30%
把	13,777	106	1	0	0	0	1	0	108	0.77%
间	6,605	90	2	2	1	3	1	0	597	1.36%

In the BCC corpus **[tab. 5]**, it is more difficult to elaborate the query to include longer leftmost nominal or adjectival modifiers. Hence, detected modifiers are up to 2 characters long;[12] furthermore, composite queries to detect multiple patters (as in columns 9-10 of table 4) are not possible. This implies that the number of undetected tokens is higher than that in the ZHTenTen (ST) corpus. This is reflected in the figures, that are sensibly lower. The classifier with the highest ratio in the last column is still 位 *wèi*, although the ratio is lower (5.67%), about half the ratio in the ZHTenTen (ST) corpus.

12 Queries are of the type of [。; ? !](a/n/a n) (的)一CLF.

Table 6 PKU corpus

CLF	Any position	Patterns (a) – (b)	Pattern (c)	All patterns	Ratio	CLF	Any position	Patterns (a) – (b)	Pattern (c)	All patterns	Ratio
	'*yī* CLF'	S.I. '*yī* CLF'	Leftmost 2-character mod. +*de*	Total detected	S.I. '*yī* CLF'/ 'yī CLF'*yī* CLF'		'*yī* CLF'	S.I. '*yī* CLF'	Leftmost 2-character mod. +*de*	Total detected	S.I. '*yī* CLF / '*yī* CLF'
名	46,340	1,202	45	1,247	2.59%	所	8,724	194	7	201	2.22%
位	90,775	6,062	350	6,412	6.68%	篇	13,131	140	26	166	1.07%
只	26,904	597	51	648	2.22%	套	18,512	116	20	136	0.63%
头	11,513	154	25	179	1.34%	句	32,656	497	46	543	1.52%
伙	3,187	40	2	42	1.26%	部	54,013	781	53	834	1.45%
对	11,513	300	31	331	2.61%	张	27,310	401	27	428	1.47%
个	674,846	15,941	670	16,611	2.36%	块	27,310	286	23	309	1.05%
条	69,434	711	68	779	1.02%	座	26,148	272	13	285	1.04%
家	53,862	818	65	883	1.52%	本	18,022	337	20	357	1.87%
批	47,638	757	25	782	1.59%	系列	34,350	115	26	141	0.33%
组	6,981	62	3	65	0.89%	台	9,075	133	3	136	1.47%
班	3,709	29	3	32	0.78%	户	2,495	26	2	28	1.04%
排	4,069	92	3	95	2.26%	门	4,914	33	2	35	0.67%
项	53,679	354	68	422	0.66%	处	9,243	60	6	66	0.65%
级	11,528	79	6	85	0.69%	道	20,764	170	6	176	0.82%
件	34,619	283	11	294	0.82%	首	7,675	100	15	115	1.30%
份	27,932	194	25	219	0.69%	把	18,846	127	6	133	0.67%
期	9,047	194	4	198	2.14%	间	7938	94	11	105	1.18%

Since the PKU corpus is not tagged, complex queries involving nominal or adjectival modifiers highlighted in the previous corpora (pattern in (14c) are not possible **[tab. 4]**; however, the query (。| ? | ; | !) $2的一CLF was used to single out one/two-character modifiers (columns 4, 9). Such a query singles out, for example, modifiers such as the one in (17).

17. 我的一个好朋友他是浙江人 (PKU)
wǒ de yí ge hǎo péngyou tā shì Zhèjiāng-rén
1SG SP one CLF good friend 3SG be Zhejiang-man
'A good friend of mine (, he) comes from Zhejiang'.

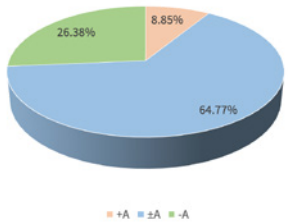
Such a limited interval minimises statistical possibilities of including verbal items and, hence, postverbal indefinites (see discussion in § 3.2). However, this involves that SIIs with longer modifiers – as in (15c') – are missing from the total count, hence the remarkably lower figures in table 4.

Discussion. Overall, results show that all examined classifiers occur with 一 *yī* in the sentence-initial position. Figures for pattern (14c) are higher in the ZHTenTen (ST) corpus, but this does not come as

a surprise, as leftmost modifiers detected in the ZhTenTen (ST) are up to 7 characters, while in the other two corpora they are up to two characters (see § 3.2). Let us focus on the two classifier 位 *wèi* and 个 *ge*: the former's total occurrences in the (14a-b-c) patterns are 36,157 in the ZHTenTen (ST), 1,717 in the BCC, and 6,412 in the PKU; the latter's are 116,021 in the ZHTenTen (ST), 14,734 in the BCC, and 16,611 in the PKU. Crucially, ratio-wise 位 *wèi* significantly outranks 个 *ge* (10.2% over 2.78% in the ZHTenTen (ST)): in other words, while the string '一位' *yí wèi* overall occurs far less than '一个' *yí ge*, in the sentence-initial position the former occurs much more frequently than the latter. Other classifiers with a relatively high ratio (last column), especially in the ZHTenTen (ST) corpus, include +A classifiers in general and ±A classifiers like 组 *zǔ* 'group' and 班 *bān* 'class' (highly compatible with +A nouns) – almost all show a ratio above 3% in the ZHTenTen (ST). Relatively high ratios are also displayed by some –A classifiers, such as 级 *jí* 'level' (3.59%), 期 *qī* 'period' (7.11%), 部 *bù* 'part' (3.43%), 句 *jù* 'line' (3.89%), and 首 *shǒu* 'piece (e.g. of poetry/lyric', 3.99%). Indefinite noun phrases with the first three classifiers (级 *jí* 'level', 期 *qī* 'period', 部 *bù* 'part') display an interesting common semantic trait related to partitivity: the referent may denote a part of a given whole, a level of a given multi-layered structure, a step of a given path, or else a phase of a given plan or project (see examples in sections below). The relatively high frequency of such NPs in the sentence-initial position might then be connected to the fact that the referent, although not identifiable, is at least *locatable* in a given set/whole/container that is comprehensible thanks to the semantics of each classifier (e.g. one level of a specific hierarchy, one step of a specific procedure etc.); it may also be specified in the previous context or, otherwise, be implicit (stage topics,[13] see discussion for sentence (4c)). This point will be examined in the qualitative analysis below. Conversely, 句 *jù* and 首 *shǒu* (classifiers for lines/quotes, and for songs/poems, respectively) come rather unexpected. We will look further into these classifiers through the qualitative analysis.

Let us now have a closer look at aggregated data with respect to the animacy trait (+A, ±A, and –A) in the ZHTenTen (ST) corpus **[tab. 7]**.

13 This is, in turn, related to the frame-containment property of topics (Chafe 1976; Her 1991; Morbiato 2020): topics express a frame of validity for the rest of the predication and are often a semantic container/whole/setting for what comes next.



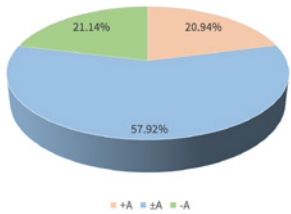


Chart 1 All '一 *yī* CLF' occurrences per animacy trait

Chart 2 Sentence-initial '一 *yī* CLF' occurrences per animacy trait

Table 7 Distribution of '一 *yī* CLF' patterns in the ZHTenTen (ST) corpus

	Sentence-initial position				**Any position**	**Ratio**
	(a) + (b)	(c) without *de*	(c) with *de*	all patterns	all patterns	
+A	35,666	10,292	3,501	49,459	632,734	7.82%
±A	108,852	18,242	6,261	133,355	4,629,012	2.88%
–A	32,787	13,609	3,472	49,868	1,885,358	2.65%
Total				**232,682**	**7,147,104**	**3.26%**

A total of 232,682 sentence-initial NPs introduced by '*yī* CLF' were detected in the corpus. As discussed, such a total includes neither NPs modified by relative clauses nor NPs preceded by modifiers longer than 7 characters and separated by commas (e.g. temporal/locative frame topics). Interestingly, almost 8% of animate NPs introduced by '一 *yī* CLF' are sentence-initial, while the ratio drops to 2.88% for ±A classifiers, and to 2.65% for –A classifiers. Charts below represent the percentage of '一 *yī* CLF' tokens over the total amount of tokens in all positions **[chart 1]** and in the sentence-initial position **[chart 2]**, divided per animacy trait: as can be seen, the percentage of +A tokens is significantly higher (more than double) in the sentence-initial position (8.8% *vs* 20.9%).

5 Qualitative Results

As discussed in § 3.2, a random sample of 300 '一 *yī* CLF' tokens was extracted from the ZHTenTen (ST) corpus, 100 for each type of classifiers: solely +A, (名 *míng*, 位 *wèi*, 只 *zhǐ*, 头 *tóu*, 伙 *huǒ*), ±A (个 *ge*, 条 *tiáo*, 家 *jiā*, 批 *pī*, 组 *zǔ*, 排 *pái*, 班 *bān*), and –A (项 *xiàng*, 级 *jí*, 件 *jiàn*, 份 *fèn*, 期 *qī*, 所 *suǒ*, 篇 *piān*, 套 *tào*, 句 *jù*, 部 *bù*, 张 *zhāng*, 块 *kuài*, 座 *zuò*, 本 *běn*, 系列 *xìliè*, 台 *tái*, 户 *hù*, 门 *mén*, 处 *chù*, 道 *dào*, 首 *shǒu*, 把

bǎ, 㠯 *jiān*). The referential properties of each NP introduced by '一 *yī* CLF' were analysed in all three subcorpora; results are in table 8.

Table 8 Referential properties of '一 *yī* CLF' tokens for each subcorpus of the ZHTenTen

	+A	±A	–A
SIIs	94	34	28
Generic	3	43	27
Referential	2	9	4
Referential SIIs	0	0	11
Numeral	0	9	25
Distributive	0	0	1
Wrong (postverbal)	1	5	4
Total	**100**	**100**	**100**

Let us first focus on SIIs: strikingly, 94% of +A tokens display an indefinite reading and hence are true SIIs. In other categories, conversely, the percentage of true SIIs drops to 34% for ±A and 28% for –A tokens. If we assume that the above figures are statistically relevant (although this would benefit from more tests conducted on different samples), we could consider these three percentages as coefficients that enable determining the true amount of SIIs from quantitative data presented in § 4. For data from the ZHTenTen (ST) corpus, results would be as follows:

Table 9 Percentage of true SIIs per +A, ±A, and–A animacy traits, ZHTenTen (ST)

	Total detected '*yī* CLF'	**Percentage of '*yī* CLF'**	**Samples' SII coefficient**	**Number of true SIIs**	**Percentage of true SIIs**
+A	49,459	21%	94%	46,491	44%
±A	133,355	58%	34%	45,341	43%
–A	49,868	21%	28%	13,963	13%

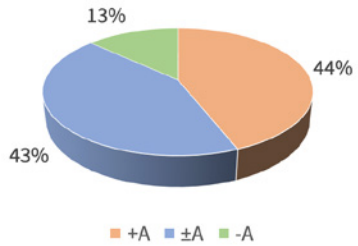


Chart 3 Percentage of true SIIs per animacy traits in the ZHTenTen (ST)

Figures in table 9 also show that animate SIIs in fact constitute a much higher percentage in the corpus, i.e. about 44% (see chart 3).

Let us now look more closely at the ±A subcorpus. First, the 100 tokens were analysed and differentiated according to the animacy trait of their head noun: 35 tokens consisted of +A NPs, 60 were –A NPs, while 5 were invalid tokens. Then, SIIs were identified in each group; figures are in table 10.

Table 10 Animate *vs* inanimate SIIs in the ±A subcorpus

±A	+A	–A
SII	12	22
Other	23	38
Total	35	60

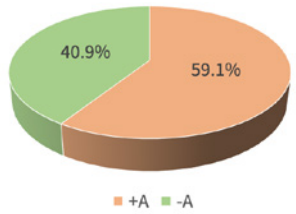


Chart 4 Percentage of true SIIs per +A and –A animacy traits, ZHTenTen (ST)

Interestingly, a reverse tendency can be observed with respect to +A tokens within the ±A subcorpus: only 12 (34%) are true SIIs (as compared to 94% in the +A subcorpus). Moreover, getting back to the comparison between 个 *ge* and 位 *wèi*, in the qualitative analysis, +animate (and +human) tokens introduced by 位 *wèi* tend to be referential/specific SIIs; conversely, for those introduced by 个 *ge*, generic NPs are twice as much as specific SIIs. This is very likely connected to their semantics: 位 *wèi* implies respect or courtesy and likely involves that the speaker knows the referent (specific indefi-

nite); 个 *ge*, on the other hand, means 'unit' and is more suitable to talk about a generic class, e.g. the NP 一个四川人 *yí ge Sìchuān-rén* 'A Sichuanese' in (18) from the ±A subcorpus:

18. 一个四川人可能很真诚的为"扬州十日"而 垂泪 […]

yí	***ge***	***Sìchuān-rén***	*kěnéng*	*hěn*	*zhēnchéng*	*de*
one	CLF	Sichuan-person	maybe	very	sincerely	SP
wèi	*Yángzhōu*	*Shí*	*Rì*	*ér*	*chuí-lèi*	
for	Yangzhou	10	days	SP	shed-tear	

'A Sichuanese may sincerely shed tears for the "Ten Days of Yangzhou" […]'

If we further split ±A SIIs into A+ and –A and add this data to percentages indicated in table 11, we obtain the following figures:

Table 11 Percentage of true SIIs per +A and –A animacy traits, ZHTenTen (ST)

	Number of true SIIs	Percentage of true SIIs
+A	62,494	59%
–A	43,301	41%

Such a projection suggests that, in the ZHTenTen (ST) corpus, a total of 105,795 SIIs can be detected. If compared to the total amount of '一 *yī* CLF' occurrences in the corpus, SIIs are 1.48%. Moreover, it suggests that, roughly, 6 SIIs out of 10 are animate. This proves that animacy is indeed a very significant trait for sentence-initial indefinite NPs. Again, this is in line with other cross-linguistic studies on the sentence-initial position and animacy.

Some examples. Let us now look at some of the most relevant examples of SIIs. As said, most are +animate (in fact, +human) and specific (known to the speaker but not to the hearer). A significant amount of examples involving +human SIIs introduce reported speech, either indirect (19) or direct (20). Verbs occurring in these sentences include: 提出 *tíchū* 'mention', 说 *shuō* 'say', 说明 *shuōmíng* 'explain', 坦言 *tǎnyán* 'say frankly', 告诉 *gàosù* 'tell', 表示 *biǎoshì* 'express'. Crucially, these verbs imply that the utterance is contextually situated in specific spatio-temporal coordinates, i.e. where and when the sentence is uttered (hence, it is locatable):

19. 一位人类学家曾经提出, 正常男女生交往的空间距离是 […]

yí	*wèi*	*rénlèixuéjiā*	*céngjīng*	*tíchū*	*zhèngcháng*
one	CLF	anthropologist	once	suggest	normal
nánnǚshēng	*jiāowǎng*	*de*	*kōngjiān*	*jùlí*	*shì*
male.female	interact	SP	spatial	distance	be

'An anthropologist once suggested that the normal spatial distance between boys and girls is […]'

20. 一名姓程的出租车司机说: "上下班时间是最多人打车的 […]
yì *míng* *xìng* *Chéng* *de* *chūzūchē* *sījī* *shuō*
one CLF surname Cheng SP taxi driver say
shàngxiàbān *shíjiān* *shì* *zuìduō* *rén* *dǎchē* *de*
commute time be most people take.taxi SP
'A taxi driver surnamed Cheng said: "Most people take taxis during commuting hours […]"'.

Reported speech SIIs are also found with inanimates, although such cases are much rarer:

21. 一项令人振奋的新研究表明 […]
yí *xiàng* *lìng* *rén* *zhènfèn* *de* *xīn*
one CLF cause people excite SP new
yánjiū *biǎomíng*
research show
'An exciting new study shows that […]'

Some +A SIIs are not specific; however, the context makes them at least *locatable* (see discussion in § 2). This is the case of (22): the referent of 一位父亲 *yí wèi fùqīn* 'a father' is not identifiable, but rather locatable within the temporal and spatial settings previously specified in the article, namely a dancing event at the Huazhong Agricultural University (cf. context). Similarly, in (23) the context makes it clear that the referent of 一位坐在最后一排的演 *yí wèi zuò zài zuìhòu yì pái de yǎnyuán* 'an actor sitting in the last row' cannot be identified, but rather located, within the given venue/group of 160 meeting participants:

22. [Context: article on a dancing event at the Huazhong Agricultural University; the previous two sentences contain no mentions of any event participant]
一位父亲领着自己刚及膝盖的女儿在场内跳着华尔兹 […]
yí *wèi* *fùqīn* *lǐng-zhe* *zìjǐ* *gāng* *jí* *xīgài* *de*
one CLF father lead-DUR REFL just reach knee SP
nǚ'ér *zài* *chǎng-nèi* *tiào-zhe* *huá'ěrzī*
daughter at field-in jump-DUR waltz
'A father with his daughter, who barely reaches his knees, dances waltz on the dancefloor […]'

23. [Context: meeting between a party committee and 160 employees in a huge venue]
一位坐在最后一排的演员站起来, 向市委宣传部副部长王立光提问 […]
yí *wèi* *zuò* *zài* *zuìhòu* *yì* *pái* *de* *yǎnyuán*
one CLF sit (be).at last one row SP actor
zhàn-qǐlái *xiàng* *Shìwěi*
stand-up towards Municipal.Party.Committee
Xuānchuán-bù *fùbùzhǎng* *Wáng* *Lìguāng* *tíwèn*
Propaganda-dept. vice.minister Wang Liguang ask
'An actor sitting in the last row stood up and asked Wang Liguang, Deputy Minister of the Municipal Party Committee Propaganda Department […]'

Other 'locatable' SIIs bear a partitive or whole-part relationship with previous sentences, as in (24). A partitive relationship is particularly frequent in occurrences of inanimate classifiers with an inherent partitive meaning (as hypothesised in § 4), e.g. 级 *jí* 'level' and 期 *qī* 'period, phase'.[14] In most cases, these receive a definite/numeral reading, e.g. 'the first phase' in (25).

24. [Context: story. The previous two sentences describe the protagonist looking at his own feet, and moving one to the wall's corner "一只移向墙角。" *yì zhǐ yí xiàng qiángjiǎo*]
一只移向门外 […]
yì *zhǐ* *yí* *xiàng* *mén-wài*
one CLF move towards door-out
'(I move) the other outside the door […]'

25. [Context: Text presenting an energy production plant]
一期装置拟建年产180万吨甲醇、68万吨烯烃。
yì *qī* *zhuāngzhì* *nǐ* *jiàn* *nián* *chǎn* *yībǎibāshí*
one CLF plant plan build year output 180
wàn *dūn* *jiǎchún* *liùshíbā* *wàn*
ten.thousand ton methanol 68 ten.thousand
dūn *xītīng*
ton olefin
'In the first phase, the plant is planned to produce 1.8 million tons of methanol and 680,000 tons of olefins per year'.

14 Qualitative data also reveal that the high frequency of patterns like '一级' *yì jí* is also connected to frequency in tables (tabs are also counted as sentence boundaries (<s>) in the ZHTenTen (ST) and are hard to rule out from the search).

A very interesting subtype found in –A tokens are referential SIIs, which come in three types: the first type (26) features a modifier that renders the referent uniquely identifiable, such as 最后 *zuìhòu* 'the last' or 最初 *zuìchū* 'the first'. The second type (27), also common in other languages (including English), is a sort of cross-clausal apposition linked to a referent mentioned in the previous context:

26. 最后一篇则包括了七个冥想练习 […]
zuìhòu *yì* *piān* *zé* *bāokuò-le* *qī* *ge*
last one CLF conversely include-PFV seven CLF
míngxiǎng *liànxí*
meditation exercise
'The last, on the other hand, includes seven meditation exercises […]'

27. [Context: the protagonist has just recalled a sentence pronounced by her grandmother]
一句看似无心的话, 却准确的[地]预测了我的未来
yí *jù* *kànsì* *wúxīn* *de* *huà* *què*
one CLF look.as unintentional SP word but
zhǔnquè *de* *yùcè-le* *wǒ* *de* *wèilái*
correctly SP predict-PFV 1SG SP future
'A seemingly unintentional sentence had in fact accurately predicted my future'.

The third type (28)-(29) interestingly features a proper name rather than a common name introduced by '一 *yī* CLF'. Classifiers occurring in this (not rare) pattern include 句 *jù* and 首 *shǒu*, thus explaining these classifiers' high sentence-initial ratios observed in table 4. This pattern had not been identified in our preliminary discussion, which confirms that corpora may help singling out new phenomena or patterns in a given language:

28. 一首《春天的故事》记录了1979年的那段往事 […]
yì *shǒu* *Chūntiān* *de* *Gùshì* *jìlù-le*
one CLF spring SP story record-PFV
yījiǔqījiǔ *nián* *de* *nà* *duàn* *wǎng-shì*
1979 year SP that CLF past-event
'A (the) (song) "The Story of Spring" recorded the events that happened in 1979 […]'

29. 一本《明朝那些事儿》可能就会让很多从来不看历史的人, 从此变成历史书的读者。
yì *běn* *Míng Cháo* *nà* *xiē* *shìr* *kěnéng*
one CLF Ming Dynasty that CLF(some) thing maybe
jiù *huì* *ràng* *hěn-duō* *cónglái* *bú* *kàn* *lìshǐ*
then will make very-many ever NEG read history

de rén cóngcǐ biànchéng lìshǐ shū de dúzhě
SP people from.now.on become history book SP reader
'A (the) book "Those Things Happened in the Ming Dynasty" may make many people who never read about history become readers of history books'.

We had found an example of such a pattern in table 1 above, reported in (30) below. In this case, the pattern occurs postverbally, but still features a proper noun (here, a title) introduced by the indefinite marker '一 *yī* CLF'.

30. 当时有两位大史学家 […]。一位是黄梨洲, 他著了一部《明夷待访录》 […]
dāngshí yǒu liǎng wèi dà shǐxuéjiā
that.time there.be two CLF great historian
yí ***wèi*** *shì Huáng Lízhōu tā zhù-le*
one CLF be Huang Lizhou 3SG.M write-PFV
yí bù Míngyí Dàifǎng Lù
one CLF Mingyi Daifang Lu
'At that time, there were two great historians […]. One is Huang Lizhou, who wrote a (the) *Mingyi Daifang Lu* […]'

If we look at this pattern from the perspective of its meaning, it seems to introduce unique referents, that are generally referred to with a proper name (such as book titles or pieces of poetry): in particular, while the speaker knows about that referent, (s)he might be not sure whether the interlocutor has some knowledge of it. Nonetheless, this would benefit from further research.

Generic readings are present in the +A subcorpus, as in (18), but are very rare (3%), while they are much more frequent with inanimates (43%), e.g. (31). Numeral (32) and distributive readings were found only in inanimate NPs:

31. 一篇短短的千字文, 往往凝结了作者十年的心血
yì piān duǎn-duǎn de qiān-zì wén
one CLF short-short SP thousand-character text
wǎngwǎng níngjié-le zuòzhě shí nián de xīnxuè
often condense-PFV author ten year SP blood
'A short thousand-word essay often condenses the author's ten years of hard work'.

32. 一套设备, 多种功能, 一本万利。
yí tào shèbèi duō zhǒng gōngnéng yì běn wànlì
one CLF device many CLF function one CLF profit
'One device, multiple functions, great profits'.

6 Conclusions and Limitations

The present study was designed to determine the statistical significance of SIIs in Chinese as well as the interconnections with features such as animacy and locatability. The quantitative and qualitative analyses discussed so far support our initial hypotheses.

Specifically, with reference to our initial research questions, this study shows that: (RQ 1) first, SIIs do exist in Chinese; statistically, their number is not unimportant. Statistical data and the analysis laid out so far suggest that, in the ZHTenTen (ST) corpus, a total of more than 100 thousands of true SIIs (i.e. sentence-initial '一 *yī* CLF' forms with a true indefinite reading) can be detected. If compared to the total amount of '一 *yī* CLF' occurrences in the ZHTenTen (ST) corpus, SIIs are 1.48%. Crucially, this analysis was not able to detect all SIIs (e.g. those introduced by numbers other than 一 *yī*, those with longer modifiers, or those modified by restrictive relative clauses as in (15c)): hence, the true amount of SIIs in the corpus is very likely to be higher. This has important implications: a theoretically sound account of the Chinese language and its word order should consider and discuss the existence and characteristics of this pattern. Similarly, SIIs should be introduced in Chinese grammars and teaching materials as well, explaining their peculiarities, tendencies, and restrictions. Of course, specific (cross-sectional or longitudinal) studies should be conducted to determine at what stage/proficiency level SIIs should be taught.

(RQ2) Animacy is indeed a factor that has significant impact on SIIs: the study shows that almost 8% of animate NPs introduced by '一 *yī* CLF' are sentence-initial, percentage that drops to 2.6 for non-animate NPs. Furthermore, roughly, 6 SIIs out of 10 are animate. Again, this is in line with other cross-linguistic studies on animacy and the sentence-initial position. Animacy was found to be a relevant factor in determining the order of event participants cross-linguistically. Studies conducted on different languages, including Spanish, Italian, Greek, Japanese, German, Dutch, Odawa (North America), and Yucatec, reveal that animate referents tend to occur before inanimate ones, regardless of their role in the event (see Van Bergen 2011 for an overview). When animate participants play the role of patients, speakers tend to produce passive sentences or to place the animate patient at the beginning of the sentence as a topic.

Finally, the above results confirm that corpora indeed contribute towards a better understanding of languages, even on topics with an established scholarship such as Chinese word order and referentiality, and allow finding new previously unobserved or underdescribed patterns in the language: the study has revealed a new reading for seemingly indefinite patterns of the type of '一 *yī* CLF N', i.e. those featuring a proper noun, as in (28) and (29).

On the other hand, the study has also highlighted some limitations of corpus tools. First, in this case a qualitative, sentence-by-sentence check was essential to refine, interpret, and validate quantitative results. Second, corpus design and POS tagging do not have a 100% reliability. For example the query "[。; ? !]n一对" in the BCC, corpus which should reveal only nominal modifiers, also identified the following (postverbal) token:

33. 若不是一对夫妇 […]
ruò bú shì ***yí duì fū-fù***
if NEG be one CLF husband-wife
'If they weren't a married couple […]'

All in all, the study clearly shows that SIIs are not only possible, but also do not constitute isolated exceptions, and that animacy and locatability indeed play a crucial role in increasing the acceptability of SIIs.

Bibliography

Bisang, W. (2016). "Chinese Syntax". Chan, S.-W.; Minett, J.; Li, W.Y.F. (eds), *The Routledge Encyclopedia of the Chinese Language*. Routledge Handbooks Online, 354-77. https://10.4324/9781315675541.ch20.

Chafe, W.L. (1976). "Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Points of View". Li, C. (ed.), *Subject and Topic*. New York: Academic Press, 25-55.

Chao Y. (1968). *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Chen B. 陈保存 et al. (eds) (1988). *Hanyu liangci cidian* 汉语量词词典 (Chinese Classifiers Dictionary). Fuzhou: Fujian renmin chubanshe.

Chen P. (2015). "Referentiality and Definiteness in Chinese". Wang W.S.Y.; Sun C. (eds), *The Oxford Handbook of Chinese Linguistics*. New York: Oxford University Press, 404-13.

Chu C. 屈承熹 (2006). *Hanyu pianzhang yufa: Lilun yu fangfa* 汉语篇章语法: 理论与方法 (Mandarin Chinese Discourse Grammar. Theory and Practice). *Russian Language and Literature Studies*, 3(13), 1-15.

Erteschik-Shir, N. (2007). *Information structure. The Syntax-Discourse Interface*. Oxford: Oxford University Press.

Fan J. 范继淹. (1985). "Wuding NP zhuyu ju" 无定NP主语句 (Indefinite Subjects Sentences). *Zhongguo yuwen*, 5, 321-8.

Fang M. 方梅. (2019). "Cong huayu gongneng kan suowei 'wuding NP zhuyu ju'" 从话语功能看所谓"无定NP主语句" (So-Called "Indefinite-Subject Sentences" from a Discourse Perspective). *Shijie Hanyu jiaoxue*, 33(2), 189-200.

Fu Y. 付义琴 (2013). "Lun Hanyu 'wuding zhuyu ju' de jushiyi" 论汉语"无定主语句"的句式义 (A Syntactic Analysis of the Chinese Sentence with an Indefinite Subject). *Yunnan shifan daxue xuebao*, 11(5), 41-6. https://doi.org/10.16802/j.cnki.ynsddw.2013.05.008.

Her, O.-S. (1991). "Topic as a Grammatical Function in Chinese". *Lingua*, 84(1), 1-23. https://doi.org/10.1016/0024-3841(91)90011-S.

Ho, Y. (1993). *Aspects of Discourse Structure in Mandarin Chinese*. Lewiston; Queenston; Lampeter: Mellen University Press.

Hole, D.P. (2012). "The Information Structure of Chinese". Krifka, M.; Musan, R. (eds), *The Expression of Information Structure*. Berlin; Boston: De Gruyter Mouton, 45-70.

Huang S. 黄师哲 (2004). "Wuding mingci zhuyu tong shijian lunyuan de guanxi" 无定名词主语同事件论元的关系 (The Relationship Between Indefinite Subjects and Event Argument). Huang Z. 黄正德 (ed.), *Zhongguo yuyan xuelun cong* 中国语言学论丛 (Essays on Chinese Linguistics). Beijing: Shangwu yinshuguan, 93-100.

Iemmolo, G.; Arcodia, G.F. (2014). "Differential Object Marking and Identifiability of the Referent. A Study of Mandarin Chinese". *Linguistics*, 52(2), 315-34. https://doi.org/10.1515/ling-2013-0064.

Li, C.N.; Thompson, S.A. (1976). "Subject and Topic. A New Typology of Language". Li, C.N. (ed.), *Subject and Topic*. New York: Academic Press, 457-89.

Li, C.N.; Thompson, S.A. (1981). *Mandarin Chinese. A Functional Reference Grammar*. Berkeley; Los Angeles: University of California Press.

Li, W. (2005). *Topic Chains in Chinese. A Discourse Analysis and Application in Language Teaching*. München: Lincom Europa.

Li, Y.A. (1997). *Structures and Interpretations of Nominal Expressions*. Los Angeles: University of Southern California.

Liu A. 刘安春 (2003). "'Yi ge' de yongfa yanjiu" "一个"的用法研究 (Research on the Usage of 'yi ge'). *Zhongguo shehui kexue yanjiushengyuan boshi xuewei lunwen*. Dissertations published by the Graduate School, Chinese Academy of Social Sciences, Beijing.

Liu A. 刘安春; Zhang B. 张伯江 (2004). "Pianzhang zhong de wuding mingci zhuyu ju ji xiangguan jushi" 篇章中的无定名词主语句及相关句式 (The Discourse Function of the Sentence with an Indefinite NP Subject). *Journal of Chinese Language and Computing*, 14(2), 97-105.

Liu X. 刘晓亚 (2018). "Wuding NP zhuyu ju de shiyong tiaojian" 无定NP主语句的使用条件 (Conditions for the Use of Sentences with Indefinite Subject NPs). *Qingnian wenxuejia*, 20.

Lu, B.; Zhang, G.; Bisang, W. (2015). "Valency Classes in Mandarin". Malchukov, A.; Comrie, B. (eds), *Valency Classes in the World's Languages*. Berlin: Mouton De Gruyter, 709-64.

Lu S. 陆烁; Pan H. 潘海华 (2009). "Hanyu wuding zhuyu de yuyi yunzhun fenxi" 汉语无定主语的语义允准分析 (The Semantic Licensing Conditions of Indefinite Subjects in Mandarin Chinese). *Zhongguo yuwen*, 6, 528-37.

Morbiato, A. (2018). *Word Order and Sentence Structure in Mandarin Chinese. New Perspectives* [PhD dissertation]. Venice; Sydney: Ca' Foscari University of Venice; The University of Sydney.

Morbiato, A. (2020). "Cognitive and Functional Principles Shaping Chinese Linear Order. The Containment Schema". *Cognitive Linguistic Studies*, 7(2), 307-33.

Paul, W. (2015). *New Perspectives on Chinese Syntax*. Berlin, Boston: De Gruyter.

Shyu, S. (2016). "Information Structure". Huang, C.; Shi, D. (eds), *A Reference Grammar of Chinese*. Cambridge: Cambridge University Press, 518-76.

Tang C. 唐翠菊 (2005). "Cong jiwuxing yongdu kan Hanyu wuding zhuyu ju" 从及物性角度看汉语无定主语句 (Transitivity and Sentences with an Indefinite NP as Subject). *Yuyan jiaoxue yu yanjiu*, 3, 9-16.

Tang H. 唐或 (2011). "'Shu (liang) ming' wuding zhuyu ju" "数(量)名"无定主语句的使用特点分析 (Analysis of the Characteristics of the Use of "Num-

ber (Classifier) Noun" Indefinite Subject Sentences). *Xinan daxue xuebao*, 37(S1), 204-6.

Tsao, F.-F. (1977). *A Functional Study of Topic in Chinese. The First Step toward Discourse Analysis*. Los Angeles: University of Southern California.

Tsao, F.-F. (1989). "Comparison in Chinese. A Topic Comment Approach". *The Tsing Hua Journal of Chinese Studies*, New Series, 19(1), 151-89.

Van Bergen, G. (2011). *Who's First and What's Next. Animacy and Word Order Variation in Dutch Language Production* [PhD dissertation]. Nijmegen: Radboud University.

Verhoeven, E. (2014). "Thematic Prominence and Animacy Asymmetries. Evidence from a Cross-Linguistic Production Study". *Lingua*, 143, 129-61. https://doi.org/10.1016/j.lingua.2014.02.002.

Wang C. 王灿龙 (2003). "Zhiyue wuding zhuyu ju shiyong de ruogan yinsu" 制约无定主语句使用的若干因素 (Constraints of Indefinite-Subject Sentences). *Yufa yanjiu he tansuo* 语法研究和探索 (Research and Explorations into Grammar). Beijing: Shangwu yishuguan, 224-39.

Wei H. 魏红; Chu Z. 储泽祥 (2007). "'You dingju hou' yu xianshixing de wuding NP zhuyu ju" '有定居后'与现实性的无定NP主语句 (On the 'Postposition of Definite Subjects' and the Actual Sentences with Indefinite Subject NPs). *Shijie Hanyu jiaoxue*, 3, 38-51.

Wu, G. (1998). *Information Structure in Chinese*. Beijing: Peking University Press.

Xiong Z. 熊仲儒 (2008). "Hanyu zhong wuding zhuyu de yunzhun tiaojian" 汉语中无定主语的允准条件 (Licensing Conditions of Indefinite Subjects in Mandarin Chinese). *Anhui shifan daxue xuebao* (*Renwen shehui kexue ban*), 36(5), 541-8.

Xu, J. (2015). "Corpus-Based Chinese Studies". *Chinese Language and Discourse. An International and Interdisciplinary Journal*, 6(2), 218-44. https://doi.org/10.1075/cld.6.2.06xu.

Xu, L. (1995). "Definiteness Effects on Chinese Word Order". *Cahiers de Linguistique – Asie Orientale*, 24(1), 29-48. https://doi.org/10.3406/clao.1995.1465.

Xu, L. (1997). "Limitation on Subjecthood of Numerically Quantified Noun Phrases. A Pragmatic Approach". Xu, L. (ed.), *The Referential Properties of Chinese Noun Phrases*. Paris: Ecole des Hautes Etudes en Sciences Sociales, 25-44.

Xu L. 徐烈炯 (1999). "Mingcixing chengfen de zhicheng yongfa" 名词性成分的指称用法 (On the Semantic Content of Noun Phrases). Xu L. 徐烈炯 (ed.), *Gongxing yu gexing – Hanyu yuyanxue zhong de zhengyi* 共性与个性—汉语语言学中的争议 (Generality and Individuality. Controversies in Chinese Linguistics). Beijing: Beijing yuyan wenhua daxue chubanshe, 176-90.

Xu L. 徐烈炯; Liu D. 刘丹青 (2007). *Huati de jiegou yu gongneng* 话题的结构与功能 (Topic. Structural and Functional Analysis). Shanghai: Jiaoyu Chubanshe.

Zhang X. 张新华 (2007). "Yu wuding mingci zhuyu ju xiangguan de lilun wenti" 与无定名词主语句相关的理论问题 (On Indefinite-Subject Sentences). *Beijing daxue xuebao* (*Zhexue shehui kexue ban*), 44(6), 103-11.

Zhou S. 周思佳; Chen Z. 陈振宇 (2013). "'Yi liang ming' buding zhi mingci zhuyu ju yunzhun tiaojian jiliang yanjiu" "一量名" 不定指名词主语句允准条件计量研究 (A Quantitative Study of the Licensing Conditions of Indefinite Subjects Marked by "Yi(一) + Quantifier + Noun"). *Yuyan kexue*, 12(4), 371-82.

Zhu D. 朱德熙 (1982). *Yufa jiangyi* 语法讲义 (Lecture Notes on Grammar). Beijing: Shangwu yishuguan.