Reading and Writing (2022) 35:1889–1910
https://doi.org/10.1007/s11145-022-10272-8

Predicting native Chinese readers' perception of sentence boundaries in written Chinese texts

Kun Sun[1] · Xiaofei Lu[2]

Accepted: 13 December 2021 / Published online: 17 March 2022

Abstract

The notion of sentencehood in Mandarin Chinese is much less well-defined than in many other languages, with a block of clauses often joined by commas without conjunctions and with the period often occurring at the end of a block of clauses to indicate meaning completeness rather than the completeness of a sentential structure. The potential factors that may affect native Chinese speakers' judgment of meaning completeness and perception of sentence boundaries have not yet been systematically examined. In light of this research gap, this study investigates the factors that may play a role in native Chinese speakers' sentence boundary perception. To this end, we conducted text re-punctuation experiments in two separate groups, a training group and a testing group, using different stimuli texts. The stimuli texts were annotated with multiple levels of linguistic information to identify potentially relevant variables that could affect the participants' sentence boundary perception. Logistic regression and the Bayesian statistical methods were applied to test the potential effects of multiple variables on the participants' responses. The logistic regression model trained on the data from the training group achieved a high level of accuracy in predicting the responses by the testing group. The model revealed a more important role of semantic information than syntactic information in the participants' sentence boundary perception. The implications of our findings for understanding the perception of Chinese sentence boundaries are discussed.

Keywords Sentence boundary perception · Re-punctuation · Corpus annotation · Logistic regression models

✉ Kun Sun
kun.sun@uni-tuebingen.de

[1] Department of Linguistics, University of Tübingen, 72074 Tübingen, Germany

[2] Department of Applied Linguistics, The Pennsylvania State University, 234 Sparks Building, University Park, PA 16802, USA

Introduction

Linguists working on Chinese have placed much emphasis on syntactic differences between Chinese and Indo-European languages (Chappell et al., 2007; Wu & He, 2015), with many studies discussing the unique characteristics of certain Chinese syntactic constructions (e.g., the *ba* construction and the serial verb construction) (Paul, 2008; Shi, 2000; Sun, 2018). Somewhat surprisingly, however, a more fundamental issue pertaining to the nature of basic syntactic units, i.e., the notion of sentencehood or sentence boundaries in Chinese, has not been systematically examined in comparison to the notion in other languages. In written texts in English and many other languages, sentence-final punctuation marks such as the period are used to indicate the completeness of a sentential structure, constrained largely by well-established syntactic rules (Huddleston & Pullum, 2002, pp. 1723–1732; Partridege, 1998, pp. 9–13). For example, a simple declarative sentence in English is "a complete unit of meaning which contains a subject and a verb, followed, if necessary, by other words which make up the meaning" (Alexander, 2019, p. 4), marked by a period at the end. However, the concepts of sentence and sentence boundary in Mandarin Chinese are both quite distinct from those in English. A complete sentence in written Chinese as punctuated by sentence-final punctuation represents the writer and reader's judgment of the completeness of the meaning or idea being expressed rather than of the completeness of a sentential structure. Indeed, a sentence is often described as "the completeness of an idea or meaning" (Lu, 2013, p. 21) by Chinese grammarians and linguists (e.g., Huang & Shi, 2016; Li & Thompson, 1989).

As illustrated in Example (1), multiple clauses can be joined using commas without conjunctions in Chinese texts, with the period (i.e., "。" in Chinese) occurring at the end of the block of clauses to indicate the completeness of the meaning or idea therein rather than the completeness of a sentential structure (Huang & Liao, 2007; Lu & Zhu, 2013, p. 322; Xue & Yang, 2011).

[Ex1:]

[i]	(a)	但	到		了	第 二		天,								
		dàn	dào		le	dì	èr	tiān								
		but	arrive		PFV	No	two	day								
	(b)	人	虽		起	了	床,									
		rén	suī		qǐ	le	chuáng									
		he	although		rise	PFV	bed									
	(c)			头	还	沉沉	的。									
		$ø_1$		tóu	hái	chénchén	de									
		(grandpa)		head	still	dazed	PTCP									
[ii]	(d)	祖父	当真	已	病	了。										
		zǔfù	dàngzhēn	yǐ	bìng	le										
		Grandpa	really	already	sick	PFV										
[iii]	(e)	翠翠	显得	懂事	了	些,										
		Cuìcuì	xiǎnde	dǒngshì	le	xiē										
		Cuicui	seem	sensible	PTCP	more										
	(f)		为	祖父	煎	了	一罐	大发药,								
		$ø_2$	wèi	zǔfù	jiān	le	yīguàn	dàfāyào								
		(Cuicui)	for	grandpa	concoct	PFV	a M	medici-nal herbs								
	(g)		逼	着	祖父	喝,										
		$ø_2$	bī	zhe	zǔ fù	hē,										
			force	PTCP	grandpa	drink										
	(h)		又	在	屋后	菜园	地	里	摘取	蒜苗	泡在	米汤	里	作	酸	蒜苗,

[i]	(a)	但	到		了	第二		天,								
		$\o_2$	yòu	zài	wūhòu	càiyuán	dì	lǐ	zhāiqǔ	suàn-miáo	pàozài	mǐtāng	lǐ	zuò	suān	suàn-miáo
			also	at	houseback	vegetable	plot	in	pluck	garlic-shoot	soak	rice-soup	in	make	sour	garlic-shoot
	(i)		一面	照料	船只,											
		$\o_2$	yīmiàn	zhàoliào	chuánzhī											
			while	take-care-of	boat											
	(j)		一面	还	时时刻刻	抽空	赶	回家	里	来	看	祖父,				
		ϕ_2	yīmiàn	hái	shíshíkèkè	chōukòng	gǎn	huíjiā	lǐ	lái	kàn	zǔfù				
			while	also	constantly	find-time	rush	home	in	come	see	grandpa				
	(k)		问	这样	那样。											
		ϕ_2	wèn	zhèyàng	nàyàng											
			ask	this	that											

[Translation] But on the next day, although he (grandpa) was out of bed, his head was still heavy. Grandpa was really sick. Cuicui, rising to the occasion, prepared a cooling concoction and made him take it; she then picked some garlic shoots from the vegetable garden behind the house and soaked them in rice water to make sour garlic shoots; she took care of the boats and found time to rush back home to check on grandpa whenever possible, asking how he was doing.[1]

Example (1) contains three sentences punctuated with the period. The fact that the third sentence [iii] contains seven clauses joined with commas without conjunctions well exemplifies the point that the period in Chinese marks the completeness of a meaning or idea (as judged by the writer) rather than the end of a complete sentential structure. This point can be further illustrated by two additional facts: (1) a comma could have been used to join the first two sentences without adding any conjunction, and (2) a period could have been inserted at the end of several clauses in the third sentence, such as (e), (g) or (h) simply by replacing the empty category $\o_2$ (refer to "Cuicui") with a pronoun. The alternative ways to punctuate the sentences in Example (1) also indicate that the judgment of meaning completeness is a more subjective task than that of the completeness of a sentential structure, given the absence of well-defined rules. All in all, the same rules that govern the use of the period as a sentence-final punctuation mark in English written texts do not fully apply in Chinese written texts.

Previous psycholinguistic research on sentence boundary perception in English has focused on spoken language, as the task in written language is relatively established with the existence of clear syntactic constraints. Such research has reported important effects of prosodic cues (e.g., pauses) and syntactic/semantic contextual variables on native English listeners' sentence segmentation and utterance understanding (e.g., Marslen-Wilson, 1975; Steinhauer & Friederici, 2001). Psycholinguistic research on Chinese sentence boundary perception has so far focused on prosodic and phonological boundaries in spoken Chinese as well (Lai et al., 2016). More broadly, a body of theoretical and experimental studies have explored the prosodic, syntactic, and semantic functions of punctuation in spoken or written texts and the ways in which punctuation may affect sentence processing and comprehension (e.g., Baron, 2001; Heggie & Wade-Woolley, 2018; Hirotani et al., 2006; Liu et al., 2010; Niikuni & Muramoto, 2014; Pynte & Kennedy, 2007; Scholes & Willis, 1990; Schou, 2007). Some researchers have also profiled the frequency distribution of punctuation marks (e.g., Kulig et al., 2017; Sun & Wang, 2019) and developed algorithms for automatic text punctuation (e.g., Christensen et al., 2001; Liu et al., 2006). Despite the issues surrounding Chinese sentence boundaries discussed above, however, the potential factors that affect native Chinese speakers' meaning completeness judgments and sentence boundary perception have not been systematically explored. The importance of this issue is similar to that of the issues investigated in studies of sentence boundary perception in spoken English language

[1] The abbreviations used in the literal translations of this Chinese text are as follows: M–measure unit, PFV–perfective aspect, PTCP–participle. These abbreviations are also applied in the translations of other Chinese examples.

(Marslen-Wilson, 1975; Steinhauer & Friederici, 2001) or word boundary segmentation in Chinese (given the lack of word boundary markers in Chinese) (Li et al., 2009; Ma et al., 2014; Yen et al., 2012).

In light of the research gap, the current study sets out to determine the potential role of various syntactic and semantic factors in native Chinese speakers' sentence boundary perception. To this end, we administered a re-punctuation task to two groups of native Chinese speakers (a training group and a testing group) using two different stimuli texts. We annotated the stimuli texts for a number of syntactic, semantic and textual factors, and employed logistic regression and the Bayesian statistical methods to examine the potential effects of such factors on the participants' responses. The logistic regression model trained on the data from the training group was then used to predict the responses by the testing group, and the performance of the model is compared against that of a machine learning model.

Research questions and hypothesis

Specifically, the current study aims to address the following two research questions:

(1) What syntactic and semantic factors may affect native Chinese readers' meaning completeness judgments and sentence boundary perception?
(2) How well can a model of such factors predict native Chinese readers' sentence boundary perception?

Based on our own observations and informed by the findings from the specific body of studies on sentence boundary perception in spoken language and the broader body of studies of the functions of punctuation marks in spoken and written texts, we hypothesize that native Chinese speakers' period use is affected by a combination of syntactic, semantic, and textual features.

Our first hypothesis is that native Chinese speakers' sentence boundary perception may be influenced by the syntactic structure and length of a clause, particularly with respect to whether a single clause may be a standalone sentence. A single clause with either a full *subject-predicate* structure or a phrasal structure (e.g., a verb phrase or noun phrase) could be a standalone sentence on its own, and a single-clause sentence may be either long or short. However, it remains to be seen whether longer clauses or clauses with a full subject-predicate structure are more likely to be perceived as shorter clauses or clauses with a phrasal structure, given the difference in the amount of information encoded in such clauses.

Our second hypothesis is that the *semantic relations* between clauses may influence native Chinese speakers' judgments of whether the clauses are parts of the same complete meaning or different meanings. In particular, we hypothesize that two clauses with the following five semantic relations, adopted (along with their abbreviations) from The Penn Discourse Treebank (PDTB, Webber et al., 2019) and the Chinese Discourse Treebank (Zhou & Xue, 2015), will likely be judged to be parts of the same complete meaning: (1) temporal te, including succession, precedence, and simultaneity; (2) contingency (ce), including cause-effect, conditional, and purpose; (3) comparison (cm), including contrast and concession; (4) expansion (ex), including conjunction, succession, coordination, progression; and (5) elaboration (el), including further explanations or provisions of additional details in different categories. We also hypothesize that the use of explicit markers to indicate these semantic relations may make it more likely for the two clauses to be judged as parts of the same complete meaning.

Our third hypothesis is that certain types of semantic shifts at the textual level may affect native Chinese speakers' judgments of meaning completeness. The first type of semantic shift hypothesized to affect meaning completeness judgment is "topic shift". In Chinese, a block of clauses may form a "topic chain" when they share the same topic, which is explicitly mentioned in a topic clause but implicitly referred to with an empty category in several comment clauses (Li, 2004; Sun, 2019), as illustrated in the third sentence in Example (1), in which the topic "Cuicui" is explicitly mentioned in clause (e) and stays the same through the end of the topic chain. A "topic shift" occurs if a different topic arises in the next block of clauses. A block of clauses may also be put in the same sentence if their topics are different but related semantically and thematically, as illustrated in the first sentence in Example (2), in which the topics of the clauses all pertain to the natural environment. In this case, a topic shift occurs when the topic of the next block of clauses changes thematically, as illustrated in the second sentence in Example (2), whose topic is "Wukui," a person. The point at which a topic shift occurs may be taken as a point of meaning completeness, indicated by the use of a period, as Example (2) exemplifies.

[Ex 2:]

[i]	(a)	终	一日，	太阳	还	没有	出来，	村口	河岸			
		zhōng	yīrì,	tàiyáng	hái	méiyǒu	chū lái,	cūnkǒu	héàn			
		finally	one-day,	sun	still	NOT	come-out,	village-entrance	river-bank			
		一层	薄雾	闪动	着	蓝光。						
		yīcéng	báowù	shǎndòng	zhe	lánguāng 。						
		a-layer	thin-mist	shine	PTCP	blue-light						
[ii]	(b)	五魁	瞧见	女人	提	着	篮子	到	河边	洗	衣服	了。
		Wǔkuí	qiáojiàn	nǚrén	tí	zhe	lánzi	dào	hébiān	xǐ	yīfu	le
		Wukui	see	woman	carry	PTCP	basket	arrive	river-bank	wash	clothes	PFV

Second, a "character shift" occurs when the character or person of concern shifts from that in one block of clause to a different one in a subsequent block of clause. When the characters are also the topics of the two blocks of clauses, a character shift becomes a subtype of topic shift. In Example (1), the character of concern is "grandpa" in clause (d) and changes to "Cuicui" in next block of clauses (e-k).

Table 1 Information about the passages used in the training and testing groups

Counts	Training group	Testing group
Number of participants	80	50
Number of passages	8	7
Number of characters in all passages	894	882
Number of periods in the original passages	32	30
Number of commas in the original passages	56	57
Number of punctuation marks removed	88	87
Number of temporal relations	6	7
Number of contingency relations	19	10
Number of comparison relations	3	7
Number of expansion relations	32	41
Number of elaboration relations	10	6
Number of explicit markers for the five semantic relations	30	26
Number of other/no semantic relations	20	16
Number of topic shifts	7	12
Number of character shifts	12	4
Number of category shifts	8	5
Number of time shifts	9	6
Number of space shifts	10	2

A character shift may indicate meaning completeness, as the period at the end of clause (d) in Example (1) illustrates.

Third, a "category shift" occurs when the category of activities or behaviors described changes from one block of clauses to another block (e.g., from physical activities to psychological activities), even if these behaviors or activities are performed by the same person. This is illustrated in Example (3), in which clauses (d-f) talk about physical activities of the "Third Master," while clause (f) shifts to describing his psychological activities. Such a category shift may indicate meaning completeness.

[Ex 3:]

[i]	(a)		在	盐坨	里	藏	了	一天一夜,					
			zài	yántuó	lǐ	cáng	le	yī tiān yī yè					
		ø$_1$(he)	at	salt-pile	in	hide	PFV	one-day-and-one-night					
	(b)		饿	了	就	抓	点	盐	末子往		嘴	上	抹抹。
			è	le	jiù	zhuā	diǎn	yán	mòzi	wǎng	zuǐ	shàng	mǒmǒ
		ø$_1$	hungry	PFV	then	grab	some	Salt	bits	toward	mouth	up	rub
[ii]	(c)		第二	天	清早	才	爬	出来,					
			dìèr	tiān	qīngzǎo	cái	pá	chūlái					
		ø$_1$	Next	day	morning	just	clime	out					
	(d)		刚	走到	宫北,								
			gāng	zǒudào	gōngběi								
		ø$_1$	just	walk-arrive	Gōngběi								
	(e)		忽	听	有人	叫	“三爷”。						
			hū	tīng	yǒurén	jiào	sānyé						
		ø$_1$	suddenly	hear	someone	call	“Third Master”						
[iii]	(f)		他	心里	一惊。								
			tā	xīnlǐ	yījīng								
			he	heart	shocked								
	(g)		因为	这	几个	月	没	听	人	叫	他	“三爷”了。	
			yīnwèi	zhè	jǐgè	yuè	méi	tīng	rén	jiào	tā	sānyé	le
			because	these	few	months	NOT	hear	people	call	him	“Third Master”	PFV

Finally, a “time shift” or “space shift” occurs when there is an obvious change in time or space from one block of clauses to another, even if the topic, character of concern, and category of activities remain the same. For example, a block of clauses may describe the physical activities of a person at one time and/or in one place, while the next block may continue talking about the same person’s physical activities at a different time and/or in a different place. A time or space shift, generally marked explicitly by a time or place expression, may indicate completeness. For example, in Example (3), there is a time shift between clauses (a-b) and (c-e), as indicated by “the next morning” at the beginning of clause (c), and the preceding clause is punctuated with a period.

Notably, the factors of character, time, and space have been extensively explored in text processing studies. In particular, the Event Indexing Model (EIM) proposes that people use the general perceptual apparatus to build situation models from narrative texts (Zwaan et al., 1995). In the EIM, events are conceptualized as activated memory nodes, and a story is represented as a set of memory nodes and the connections between them. Each memory node is coded for time, space, characters, objects, and goals (or causes), and a change in these elements activates a new memory node. Our hypothesis that a shift in character, category of activities, time, or space may indicate meaning completeness and therefore prompt the start of a new meaning aligns with the ideas of the EIM.

Methods

Materials

We selected 15 short passages from a number of well-known modern Chinese novels and removed all punctuation marks from those passages. The original passages contained commas and periods only. Eight passages were assigned to the training group and the other seven to the testing group (see the Participants section below).[2] The actual passages assigned to the training and testing groups (henceforth stimuli texts) are presented in Online Supplementary Material A and B and detailed information about the original, modified, and annotated passages is summarized in Table 1.

Participants

Altogether, 130 native Mandarin Chinese speakers (95 female, 35 male) volunteered to participate in the study, and all participants received a small remuneration for their time. Participant age ranged from 21 to 29 years (M = 24.5, SD = 0.75). Among the participants, 56 were undergraduate students majoring in English-Chinese translation, 20 were undergraduate students majoring in computer science, 52 were postgraduate students majoring in Chinese linguistics or English-Chinese translation, and two had a PhD in linguistics. Given their native speaker status and educational background, all participants were highly proficient in reading Chinese and familiar with the use of punctuation marks in written Chinese. The 130 participants were divided into two groups, with 80 in the training group and 50 in the testing group.

[2] The passages for the training group were selected from *fuxi fuxi* (by Liu Heng), *shenbian* (by Feng Jicai), *biancheng* (by Shen Congwen), *shoujie* (by Wang Zeqi), *qiqiechengqun* (by Su Tong), *wukui* (by Jia Pinwa), and *weicheng* (by Qian Zhongshu). The passages for the testing group were selected from *shoujie* (by Wang Zeqi), *qinqiang* (by Jia Pinwa), *weicheng* (by Qian Zhongshu), *nanrende yiban shi nüren* (by Zhang Xianliang), *xizao* (by Yang Jiang), *furongzhen* (by Gu Hua), and *tapu* (by Liu Zhenyun).

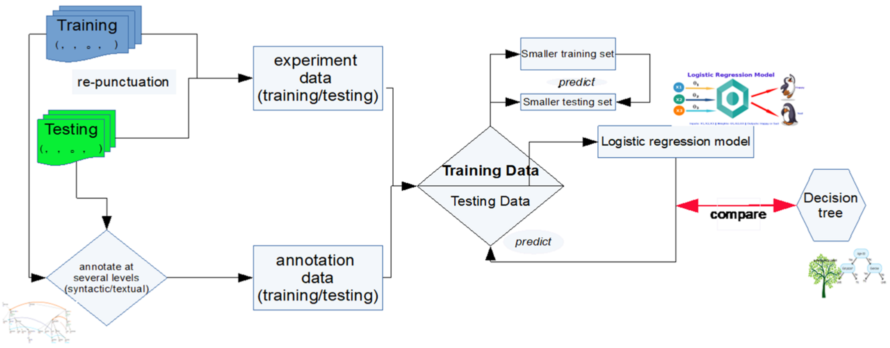
Training
(, , 。,)
re-punctuation
Testing
(, , 。,)
experiment data (training/testing)
annotate at several levels (syntactic/textual)
annotation data (training/testing)
Smaller training set
predict
Smaller testing set
Logistic Regression Model
Training Data
Testing Data
predict
Logistic regression model
compare
Decision tree

Fig. 1 Road map for the current study

The re-punctuation task

The participants were asked to re-punctuate the stimuli texts assigned to their corresponding group. Specifically, each participant received a sheet of paper containing the stimuli texts with all punctuation marks replaced with blanks and was required to fill in those blanks using commas and periods only with a pen. There was no time requirement for the re-punctuation task, and all participants completed the task in 10 to 20 min. The participants also provided information about their age, gender, and educational background at this time. After collecting the sheets from the participants, we recorded their responses and background information in the computer.

Stimuli text annotation

To test the three hypotheses discussed earlier, we analyzed and annotated each stimuli text in a number of ways, including the length and syntactic status of each clause prior to each blank, the semantic relation between each adjacent pair of clauses (i.e., temporal, contingency, comparison, expansion, elaboration, or other/no relation), the use of any explicit semantic markers indicating one of the five semantic relations of interest, and all instances of the types of semantic shift discussed above (i.e., topic, character, category, time, or space). The annotation was done by two L1 Chinese postgraduate students majoring in Chinese linguistics. The annotators received training from the first author and achieved over 80% agreement on a pilot passage. They then independently annotated all passages and met to resolve all discrepancies. This annotation work was carried out before the re-punctuation test was taken. The participants in the experiments were not provided with the annotation information. The annotators did not know how the participants filled in punctuation marks.

Table 2 Significant predictors in the logistic regression models

Independent variables	Model 1 (trained on the smaller training dataset)	Model 2 (trained on the Training dataset)
TopicShift	1.60***	1.62***
	[1.30, 1.89]	[1.37, 1.88]
CategoryShift	1.67***	1.85***
	[1.39, 1.94]	[1.61, 2.09]
CharacterShift	1.68***	1.71***
	[1.40, 1.97]	[1.46, 1.96]
TimeShift	1.31***	1.25***
	[0.97, 1.64]	[0.96, 1.54]
SpaceShift	2.03***	1.99***
	[1.76, 2.31]	[1.76, 2.23]
SemRel(comparison)	– 1.01	– 1.01
	[– 2.20, 0.19]	[– 2.04, 0.02]
SemRel(elaboration)	– 1.17***	– 1.14***
	[– 1.46, – 0.87]	[– 1.40, – 0.88]
SemRel(expansion)	– 1.29***	– 1.27***
	[– 1.52, – 1.07]	[– 1.47, – 1.08]
ExplicitMarker	– 0.34**	– 0.31**
	[– 0.60, – 0.09]	[– 0.53, – 0.09]
Length	0.32***	0.31***
	[0.22, 0.42]	[0.22, 0.40]
N	4091	5466
AIC	3273.06	4305.16
BIC	3348.85	4384.44
Pseudo R^2	0.33	0.33

$^{***}p < 0.001$; $^{**}p < 0.01$

Statistical and machine-learning methods

For both the training group and the testing group, the data gathered from the re-punctuation experiment and the annotated stimuli texts were merged.[3] We refer to the merged data for the two groups the Training dataset and the Testing dataset. The former was used to train statistical and machine learning models, and the latter was used to test those models. As the two datasets used different stimuli texts, this would give us a good sense of the reliability of the trained models on new data. Furthermore, we also randomly divided the Training dataset into two smaller parts, referred to as the smaller training dataset (75% of the Training dataset) and the smaller testing dataset (25% of the Training dataset), and used them to train and test statistical

[3] All experimental and annotation data can be accessed at https://osf.io/4u8cs/

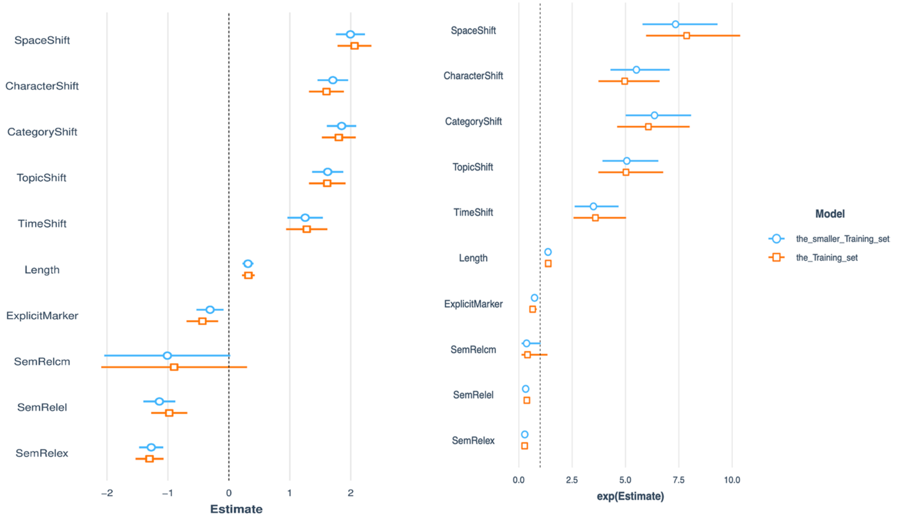

Fig. 2 Coefficient uncertainty as normal distributions. The left panel shows the raw coefficients, and the right panel shows the exponentiation coefficients. SemRelcm = the comparison relation; SemRelel = the elaboration relation; SemRelex = the expansion relation

and machine learning models as well. Given that the smaller training and testing datasets used the same stimuli texts, we could expect the trained model to perform somewhat better in this case, but a comparison of the performance of the models trained and tested in these two configurations would shed useful light on the stability of the models.

Across the Training and Testing datasets and the smaller training and testing datasets, the same 11 independent variables were hypothesized to affect the same response variable. The response variable, named 'Punctuation,' was a binary variable, coded as either "comma" or "period" depending on what a participant provided in each blank in the stimuli texts. Among the 11 independent variables, six were binary variables, namely, topic shift, character shift, category shift, time shift, space shift, and explicit markers, all of which were coded as either "1" (if present) or "0" (if absent). Four were categorical variables, namely, gender (male or female), education (undergraduate, postgraduate, or Ph.D.), syntactic status (subject-predicate, verb phrase, noun phrase, or adjectival phrase), and semantic relation (temporal, contingency, comparison, expansion, or elaboration). The last variable, clause length (i.e., number of characters in the clause) was numeric. The distribution of the semantic shift variables, explicit markers, and different categories of semantic relations in the Training and Testing datasets can be found in Table 1.

Given the binary nature of the response variable and the diversity of the types of independent variables in our dataset, categorical logistic regression modeling appears to be especially appropriate for assessing the effects of the independent variables on the response variable (Palei & Das, 2009; Sperandei, 2014). As such, we used the "glm" function in R to train and test two logistic regression models using the data partitions described above. Several additional analyses were then performed

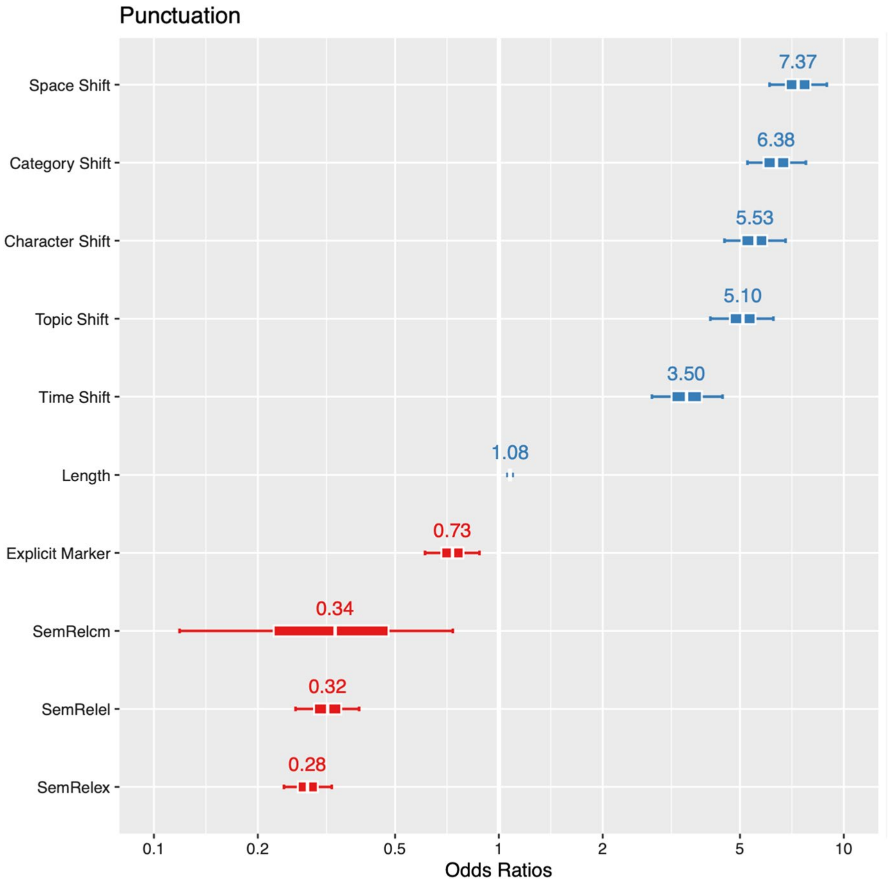

Fig. 3 The posterior distribution of significant predictors in the Bayesian fitting model

to verify the effects observed in the logistic regression models and evaluate the performance of their predictions on the testing datasets. First, we confirmed the significant fixed effects observed in the logistic regression models using Bayesian methods (Gelman et al., 2013), implemented with the R package “brms” (Bürkner, 2017). Second, we used the R package “lme4” (Bates et al., 2014) to train and test two linear mixed-effect models using the same data partitions and compared the performance of these models against that of the logistic regression models. Third, we used receiver operating characteristic (ROC), k-fold cross-validation, and bootstrap cross-validation to test the validity of the statistical models. Finally, we also trained two decision tree-based machine learning models (Hothorn et al., 2006; Song & Ying, 2015) using the same data partitions and compared their performance with that of the logistic regression models. The road map of the current study is summarized in Fig. 1.

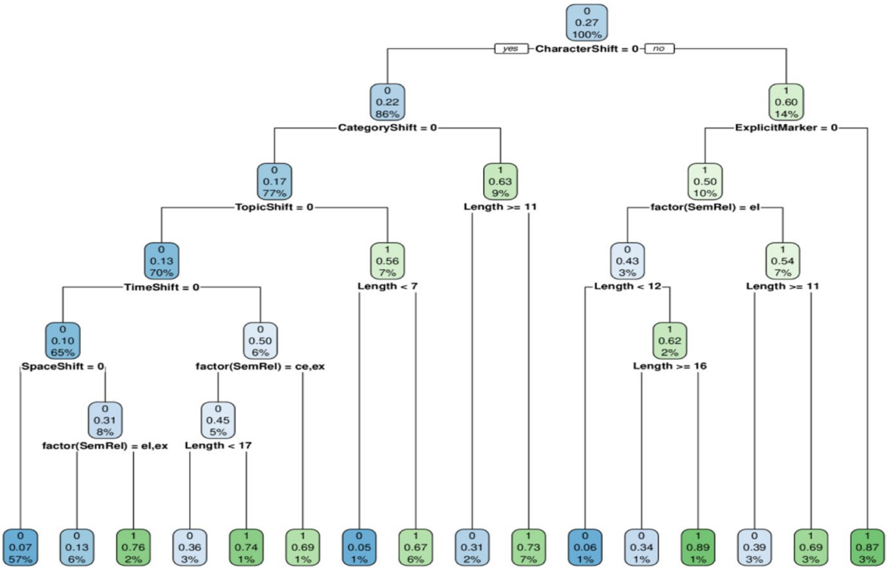

Fig. 4 A decision tree-based machine learning model for sentence boundary detection

Results

Predictors of native Chinese readers' sentence boundary perception

To determine whether significant differences in period use existed among the participants, we ran two ANOVAs separately on the Training set and the Testing set using the R function **aov()**. The tests revealed significant differences in period use among the 80 participants in the Training set (F(2, 87) = 72.48, $p < 2\text{e}{-}16$) as well as among the 50 participants in the Testing set (F(2, 86) = 85.55, $p < 2\text{e}{-}16$). These results confirmed that the native Chinese readers varied substantially in terms of their period use.

We trained a fitting regression model on the smaller training dataset (Model 1) and a second fitting regression model on the entire Training dataset (Model 2) with the same variables discussed above. The following variables: educational level, gender, and syntactic status were not significant predictors in the logistic regression tests. All five types of semantic shifts were significant predictors in both models. Three categories of semantic relations were also found to be significantly predictors in both models, while the contingency and temporal relations were not significant predictors. Explicit semantic relation markers and clause length were also significant predictors in both models. The significant predictors in the models are listed in Table 2 along with their coefficients and the corresponding 95% confidence intervals.

To better compare the strength and directionality of the significant predictors, we plotted their coefficients in Fig. 2. It is clear that the semantic shift variables have stronger effects than all other significant predictors. While the five semantic shift variables and clause length all had positive effects, explicit markers and the three semantic relation categories all had negative effects.

We further used Bayesian methods to confirm the significant fixed effects observed in the logistic regression model with the “glm” function in R. The same significant effects were replicated in the best Bayesian fitting model, as indicated by the “rhat” values of 1.0 of the same significant predictors in the “brms” package (Bürkner, 2017). Figure 3 presents the posterior distribution of the significant predictors in the best Bayesian fitting model. The effects of the significant predictors in this model largely converged with those observed in the logistic regression model. Again, five semantic shift variables and clause length positively affected native Chinese readers’ sentence boundary judgments, while three semantic relation categories and the use of explicit markers negatively affected such judgments.

To evaluate the reliability of the two logistic regression models trained, we used Model 1 and Model 2 to predict native Chinese readers’ period use on the smaller testing data and the Testing data, respectively. Model 1 correctly predicted 76.36% of the periods used in the smaller testing dataset, while Model 2 correctly predicted 75.47% of the periods used in the Testing dataset, indicating good stability of the models trained. The performance of these models was comparable to the performance of linear mixed effect models (implemented with the “lme4” function in R) trained and tested on the same datasets, which predicted 77.01% of the periods used in the smaller testing dataset and 75.01% of the periods used in the Testing dataset. The R scripts for training and testing the logistic regression models and linear mixed-effect models can be found in Online Supplementary Material C. The validity of the logistic regression model was further confirmed through ROC, k-fold cross-validation, and bootstrap cross-validation, as detailed in Online Supplementary Material D.

Results of a decision tree-based machine learning model

The results of the logistic regression model (Model 2) were also compared against those of a decision tree-based machine learning model that considered the same variables. Binary classification with the decision tree-based model was executed using an R package (party) (Hothorn et al., 2006) (see Online Supplementary Material C for the R scripts). The decision tree model trained on the smaller training dataset achieved 84.75% precision on the smaller testing dataset, higher than the 76.36% achieved by the logistic regression model. The decision tree model trained on the Training dataset (shown in Fig. 4) achieved 81.63% precision on the Testing dataset, also higher than the 75.47% achieved by the logistic regression model. While the logistic regression model achieved somewhat lower precision than the decision tree-based machine learning model, it is nevertheless more cognitively interpretable than the latter. Meanwhile, the results of the decision tree-based machine learning model further attested to the effectiveness of the variables considered in the current study for predicting native Chinese readers’ sentence boundary perception.

Discussion

The logistic regression analysis yielded a number of factors that significantly affected native Chinese readers' sentence boundary perception. Several predictors negatively affected the participants' sentence boundary judgments, indicating native Chinese readers were more likely to use a comma instead of a period in a blank when these predictors were present. Specifically, when an explicit marker of semantic relation was present or when the semantic relation between two clauses was that of comparison, expansion, or elaboration, the participants were more likely to use a comma than a period between those two clauses. These results appear to align with the positive effects observed for semantic shifts in topic, character, category, time, and space. In other words, when a clause appears to be a semantic continuation of the previous clause, particularly when that continuation is marked by an explicit semantic marker, native Chinese readers are more likely to see the two clauses as parts of the same complete meaning and thus less likely to insert a period as a sentence boundary marker between them. On the other hand, when a clause exhibits a clear semantic shift from the previous clause, native Chinese readers are more likely to see the current clause as initiating a new meaning or idea and thus more likely to insert a period at the end of the preceding clause to indicate the completeness of a meaning at that point. It can be seen that explicit markers, semantic relations, and semantic shifts function not at the level of syntactic structures but at the level of meaning and discourse flow (e.g., Moder & Martinovic-Zic, 2004). In addition, native Chinese readers were also more likely to use periods after longer clauses than after shorter ones. This may not be surprising, as on average longer clauses provide more room for meaning expression and also impose higher cognitive loads to language users than shorter clauses (Mikk, 2008). In terms of the magnitude of the effects observed for the significant predictors, the semantic shift variables exhibited the largest effect, followed by the three semantic relation categories, while explicit markers and clause length demonstrated smaller effects. These differences shed light on the greater importance of semantic shifts and relations than explicit markers and clause length in native Chinese readers' sentence boundary perception.

Gender, educational level, and the "syntactic status" of the clause (i.e., subject-predicate, verb phrase, noun phrase, or adjectival phrase) did not significantly affect the participants' sentence boundary perception. The insignificant effect of "syntactic status" is particularly worth noting. As mentioned earlier, in many other languages such as English, sentence boundaries in written texts tend to be constrained by well-established syntactic rules. That is, a complete sentence is generally expected to have a complete sentential structure, with some stylistic and/or contextual exceptions. Our results indicate that native Chinese readers' sentence boundary perception in written Chinese texts is affected primarily by semantic factors pertaining to meaning completeness at the discourse level rather than syntactic factors pertaining to structural completeness. In this sense, the concept of sentencehood in Chinese as tacitly understood by native Chinese speakers appears to be different from that defined syntactically in other languages.

The model trained and tested on data from the same stimuli texts and one trained and tested on data from different stimuli texts demonstrated comparably satisfactory performance (76.36% vs. 75.47%) in predicting the participants' period use. The effects observed for the significant predictors in the logistic regression model were also confirmed using Bayesian methods and validated using ROC, k-fold cross-validation, and bootstrap cross-validation. A decision tree-based machine learning model, which achieved somewhat better performance in predicting the participants sentence boundary perception than the logistic regression model, further confirmed the usefulness of the independent variables considered in the current study. Overall, these results suggest that the logistic regression was stable and reliable and may serve as an efficient and generalizable model for sentence boundary detection in written Chinese texts. It can also be treated as a simple cognitive model for understanding how native Chinese speakers judge meaning completeness and determine sentence boundaries. Meanwhile, the model's 24.53% error rate on the Testing dataset may be explained by the subjectivity in meaning completeness judgment discussed in the Introduction and evidenced in the significant differences among the participants revealed by the ANOVA results. It could also be the case that additional factors beyond those explored in the current study may be at play.

A theoretical implication of our findings is that sentences in written Chinese texts may resemble "text-like sentences" often found in spoken or digital English (e.g., Lotherington & Xu, 2004), characterized not by complete sentential structures but by blocks of clauses judged to contain a complete meaning. This is the case as native Chinese speakers primarily draw on semantic and discourse information rather than syntactic structural information to determine meaning completeness and sentence boundaries in a block of clauses. Theoretical explanations of the significant positive effects observed for semantic shifts on native Chinese readers' sentence boundary perception are possible from the lens of the Event Indexing Model (EIM; Zwaan et al., 1995). As mentioned earlier, this model conceptualizes events as activated memory nodes, each coded for time, space, characters, objects, and goals (or causes); a change in these elements activates a new memory node, and the memory nodes and their connections form the representation of the story in a narrative text. Several semantic shifts (i.e., time, space, and character shifts) defined in the current study overlap with the elements in the EIM, and it may be the case that the activation of a new memory node is associated with the initiation of a new meaning or idea in Chinese. Finally, the significant negative effects observed for semantic relations concur with findings from previous efforts in annotating semantic relations between clauses in discourse corpora in Chinese that such relations often reside within sentence boundaries (Webber et al., 2019; Zhou & Xue, 2015).

Conclusion

Using data from a re-punctuation experiment and annotated stimuli texts, this study examined the effects of a set of syntactic and semantic factors on native Chinese readers' sentence boundary perception. Our findings revealed significant positive effects of five types of semantic shifts and clause length on native Chinese readers' sentence boundary perception and significant negative effects of explicit markers and several categories of semantic relations on such perception. No significant effects were observed for the syntactic status of the clauses or the two basic sociolinguistic factors of gender and educational level. These results showed that native Chinese readers relied primarily on semantic factors relevant to meaning completeness instead of syntactic factors pertaining to structural completeness in judging sentence boundaries in written Chinese texts. The findings also suggest that sentences in written Chinese texts may to a large extent resemble "text-like sentences" in spoken or digital English, characterized not by complete sentential structures but by blocks of clauses judged to contain a complete meaning. Building upon the initial findings of the current study, future research can fruitfully employ additional experimental methods to investigate the effects of additional factors and their interactions (e.g., working memory) to better understand the mechanisms underlying native Chinese speakers' meaning completeness judgment and sentence boundary perception, including in particular the subjectivity and individual variation that exists in such judgment and perception. The ways in which native Chinese speakers' linguistic knowledge of the materials being processed may affect their sentence boundary perception is also worth investigating (e.g., by having them punctuate translations of foreign novels, as one reviewer suggested). Finally, future research can also investigate differences in the mechanisms underlying native Chinese speakers' sentence boundary perception in spoken and written Chinese as well as differences in the mechanism underlying Chinese-English bilingual speakers' sentence boundary perception in spoken and/or written Chinese and English.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11145-022-10272-8.

Acknowledgements We are very grateful to Dr. Yi Shan and Dr. Yi Liu for their help in collecting experimental data for the present study.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was also supported by a grant from the European Research Council (No. 742545) to the first author.

Declarations

Conflict of interest The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

Alexander, L. G. (2019). *Longman English grammar practice*. Addison-Wesley.

Baron, N. S. (2001). Comma and canaries: The role of punctuation in speech and writing. *Language Sciences, 23*(1), 15–67. https://doi.org/10.1016/S0388-0001(00)00027-9

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bürkner, P. C. (2017). BRMS: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chappell, H., Ming, L., & Peyraube, A. (2007). Chinese linguistics and typology: The state of the art. *Linguistic Typology, 11*(1), 187–211. https://doi.org/10.1515/LINGTY.2007.014

Christensen, H., Gotoh, Y., & Renals, S. (2001). Punctuation annotation using statistical prosody models. In *Proceedings of the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding* (pp. 35–40). International Speech Communication Association.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.

Heggie, L., & Wade-Woolley, L. (2018). Prosodic awareness and punctuation ability in adult readers. *Reading Psychology, 39*(2), 188–215. https://doi.org/10.1080/02702711.2017.1413021

Hirotani, M., Frazier, L., & Rayner, K. (2006). Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language, 54*(3), 425–443. https://doi.org/10.1016/j.jml.2005.12.001

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Party: A laboratory for recursive part(y)itioning. *R package version* 0.9-11.

Huang, B., & Liao, X. (2007). *Xiandai Hanyu [Modern Chinese]* (4th ed.). Higher Education Press.

Huang, J., & Shi, D. (2016). *A reference grammar of Chinese*. Cambridge University Press.

Huddleston, R., & Pullum, K. G. (2002). *The Cambridge grammar of the English language*. Cambridge University Press.

Kulig, A., Kwapień, J., Stanisz, T., & Drożdż, S. (2017). In narrative texts punctuation marks obey the same statistics as words. *Information Sciences, 375*, 98–113. https://doi.org/10.1016/j.ins.2016.09.051

Lai, W., Yuan, J., Li, Y., Xu, X., & Liberman, M. (2016). The rhythmic constraint on prosodic boundaries in Mandarin Chinese based on corpora of silent reading and speech perception. In *INTERSPEECH* 2016 (pp. 87-91). ISCA.

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. University of California Press.

Li, W. (2004). Topic chains in Chinese discourse. *Discourse Processes, 37*(1), 25–45. https://doi.org/10.1207/s15326950dp3701_2

Li, X., Rayner, K., & Cave, K. R. (2009). On the segmentation of Chinese words during reading. *Cognitive Psychology, 58*(4), 525–552. https://doi.org/10.1016/j.cogpsych.2009.02.003

Liu, B., Wang, Z., & Jin, Z. (2010). The effects of punctuations in Chinese sentence comprehension: An ERP study. *Journal of Neurolinguistics, 23*(1), 66–80. https://doi.org/10.1016/j.jneuroling.2009.08.004

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., & Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(5), 1526–1540. https://doi.org/10.1109/TASL.2006.878255

Lotherington, H., & Xu, Y. (2004). How to chat in English and Chinese: Emerging digital language convention. *ReCALL, 16*(2), 308–329. https://doi.org/10.1017/S0958344004000527

Lu, J. M. (2013). *A Course in Modern Chinese Grammar* (Forth). Peking University Press.

Lu, S., & Zhu, D. (2013). *Yufa xiuci jianghua [Lectures on grammar and rhetoric]*. Commercial Press.

Ma, G., Li, X., & Rayner, K. (2014). Word segmentation of overlapping ambiguous strings during Chinese reading. *Journal of Experimental Psychology: Human Perception and Performance, 40*(3), 1046. https://doi.org/10.1037/a0035389

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science, 189*(4198), 226–228. https://doi.org/10.1126/science.189.4198.226

Mikk, J. (2008). Sentence length for revealing the cognitive load reversal effect in text comprehension. *Educational Studies, 34*(2), 119–127. https://doi.org/10.1080/03055690701811164

Moder, C. L., & Martinovic-Zic, A. (Eds.). (2004). *Discourse across languages and cultures*. John Benjamins.

Niikuni, K., & Muramoto, T. (2014). Effects of punctuation on the processing of temporarily ambiguous sentences in Japanese. *Japanese Psychological Research, 56*(3), 275–287. https://doi.org/10.1111/jpr.12052

Palei, S. K., & Das, S. K. (2009). Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach. *Safety Science, 47*(1), 88–96. https://doi.org/10.1016/j.ssci.2008.01.002

Partridege, E. (1998). *You have a point there: A guide to punctuation and its allies*. Routledge.

Paul, W. (2008). The serial verb construction in Chinese: A tenacious myth and a Gordian knot. *The Linguistic Review, 25*(3–4), 367–411. https://doi.org/10.1515/TLIR.2008.011

Pynte, J., & Kennedy, A. (2007). The influence of punctuation and word class on distributed processing in normal reading. *Vision Research, 47*(9), 1215–1227. https://doi.org/10.1016/j.visres.2006.12.006

Scholes, R. J., & Willis, B. J. (1990). Prosodic and syntactic functions of punctuation: A contribution to the study of orality and literacy. *Interchange, 21*(3), 13–20. https://doi.org/10.1007/BF01809416

Schou, K. (2007). The syntactic status of English punctuation. *English Studies, 88*(2), 195–216.

Shi, D. (2000). Topic and topic-comment constructions in Mandarin Chinese. *Language, 76*(2), 383–408. https://doi.org/10.2307/417661

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry, 27*(2), 130. https://doi.org/10.11919/j.issn.1002-0829.215044

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica, 24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

Steinhauer, K., & Friederici, A. D. (2001). Prosodic boundaries, comma rules, and brain responses: The closure positive shift in ERPs as a universal marker for prosodic phrasing in listeners and readers. *Journal of Psycholinguistic Research, 30*(3), 267–295. https://doi.org/10.1023/A:1010443001646

Sun, K. (2018). Approaching the double-nominal construction in Mandarin Chinese through the semantic-cognitive interaction. *Studia Linguistica, 72*(3), 687–724. https://doi.org/10.1111/stul.12085

Sun, K. (2019). Integration functions of topic chains in Chinese discourse. *Acta Linguistica Asiatica, 9*(1), 29–57. https://doi.org/10.4312/ala.9.1.29-57

Sun, K., & Wang, R. (2019). Frequency distributions of punctuation marks in English: Evidence from large-scale corpora. *English Today, 4*, 23–35. https://doi.org/10.1017/S0266078418000512

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2019). *The Penn Discourse Treebank 3.0 annotation manual*. University of Pennsylvania.

Wu, F., & He, Y. (2015). Some typological characteristics of Mandarin Chinese syntax. In W. S. Wang & C. Sun (Eds.), *The Oxford handbook of Chinese linguistics* (pp. 379–392). Oxford University Press.

Xue, N., & Yang, Y. (2011). Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 631–635). Association for Computational Linguistics.

Yen, M. H., Radach, R., Tzeng, O. J. L., & Tsai, J. L. (2012). Usage of statistical cues for word boundary in reading Chinese sentences. *Reading and Writing, 25*(5), 1007–1029. https://doi.org/10.1080/17470218.2015.1061030

Zhou, Y., & Xue, N. (2015). The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations. *Language Resources and Evaluation, 49*(2), 397–431. https://doi.org/10.1007/s10579-014-9290-3

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6*(5), 292–297. https://doi.org/10.1111/j.1467-9280.1995.tb00513.x

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.