Applied Linguistics 2021: 42/1: 60–92
doi:10.1093/applin/amz064 Advance Access published on 3 February 2020

An Organic Syntactic Complexity Measure for the Chinese Language: The TC-Unit

QIAONA YU

Department of East Asian Languages and Cultures, Wake Forest University, Winston-Salem, NC, USA
E-mail: yuq@wfu.edu

INTRODUCTION

The interdisciplinary nature of applied linguistics entails the research of various multimodal and dynamic constructs. When measuring multilayered, multifaceted, and dynamic theoretical constructs of language learning and use, inaccurate conceptualization and metric of the construct may result in a skewed observation of the phenomenon. What is operationalized and measured for interpretations may not adequately reflect what the theoretical construct targets, potentially rendering the construct to be underrepresented and thus threatening the validity of studies. This study addresses construct underrepresentation in the case of Mandarin Chinese syntactic complexity by proposing and validating a new syntactic complexity measure for topic-prominent languages: the TC-unit.

The tripartite dimensions of complexity, accuracy, and fluency (CAF) have been widely used for assessing second language performance and development (Wolfe-Quintero *et al.* 1998; Housen *et al.* 2012; Michel 2017). Accuracy indicates the ability to produce target-like and error-free language. Fluency shows the degree of automatization in accessing second language capability and the ability to produce the L2 with native-like rapidity, pausing, hesitation, and reformulation. Distinct from accuracy and fluency, complexity reveals the scope of expanding or restructuring second language knowledge and the ability to use a wide and varied range of sophisticated structures and vocabulary (Lennon 1990; Skehan 1998; Wolfe-Quintero *et al.* 1998; Ellis 2003, 2008; Ellis and Barkhuizen 2005). Complexity assessments have adopted various measures, but many studies have been critiqued for taking 'a rather narrow, reductionist, perhaps even simplistic view on and approach to what constitutes L2 complexity' (Bulté and Housen 2012: 34). Such a reductionist approach was due to focusing solely on selected syntactic levels and not considering complexity comprehensively (De Clercq and Housen 2017). To caution against partial and inconsistent assessment of multidimensional constructs, Norris and Ortega (2009) advocated an *organic* approach that treats complexity as a dynamic and interrelated set of constantly changing subsystems. The organic approach investigates complexity via varied syntactic levels: global, clausal, subclausal, and specific forms. To advance this approach, the present study argues complexity measures at each syntactic level should be treated with language-specific calibration rather than imposing one-size-fits-all measures. Though many complexity measures such as the T-unit and the AS-unit operate around clause subordination (Foster *et al.* 2000; Bulté and Housen 2012), it is not uniform across languages. As such, using the mean length of T-unit, native Chinese speakers were assessed speaking less complex Chinese language than L2 Chinese learners (Jin 2006; Yuan 2009). This counterintuitive assessment suggested that the T-unit is an invalid measure for syntactic complexity analysis of Chinese. One potential cause is that clause subordination does not necessarily apply to all languages in terms of complexity composition.

Complexity is a multidimensional construct, and thus of which varied interpretations are possible. This study focuses on syntactic complexity as distinct from lexical, morphological, and phonological complexity. Bulté and Housen (2012: 22) conceptualized complexity as 'a property or quality of a phenomenon or entity in terms of (1) the number and the nature of the discrete components that the entity consists of, and (2) the number and the nature of the relationships between the constituent components'. Their conceptualization of complexity allows the number and nature of the discrete components, the relationships thereof, and the entity to take different forms in typologically different languages. In order to explore organic measures of complexity, the present study begins by rejecting the assumption that complexity is chiefly determined by subordination; and it emphasizes the local features of the target languages at varied syntactic levels. Next, this study examines how complexity is composed in a language where syntactic complexity is not primarily achieved through clause subordination. Namely, this study proposes the TC-units that are independent from clause subordination to assess Chinese syntactic complexity. Based on the general hypothesis underlying complexity development that the higher one's proficiency level is, the more complex language one is able to produce, this study then collects both spoken and written output from L1 and L2 Chinese speakers performing the same set of tasks. Based on the data collected, this study finally validates four proposed TC-unit-based complexity measures based on the degree to which the measure-based results distinguished proficiency levels in speakers of the target language.

CLAUSAL COMPLEXITY COMPOSITION IN VARYING FORMS

This study questions the notion that clause subordination, independent of the typological features of the language it is applied to, is a valid measure of syntactic complexity. According to the prominence of the notions of topic and subject in the construction of sentences, Li and Thompson (1976) proposed a typology of sentence formation based on the grammatical relations subject-predicate and topic-comment. They classified four categories after surveying thirty languages: (i) topic-prominent, for example, Chinese, Lahu, and Lisu; (ii) subject-prominent, for example, English and most Indo-European languages, Dyirbal, and Indonesian; (iii) neither topic- nor subject-prominent, for example, Tagalog and Ilocano; and (iv) both subject- and topic-prominent, for example, Japanese and Korean. In typologically different languages, the number and nature of the discrete components, the relationships thereof, and the entity that compose complexity by Bulté and Housen's definition (2012) may take different forms. Chinese syntactic complexity analysis, as explained below, showed invalid result when measured in subordination-based complexity measures.

Subordination-based complexity measures

Subordination is the underlying unit for many syntactic complexity measures. Hunt (1965: 20) defined a T-unit as the 'shortest grammatically allowable sentences into which (writing can be split) or minimally terminable unit'. Based on the categorization of clause coordination and subordination, Hunt operationalized the T-unit as 'one main clause with all subordinate clauses attached to it'. Two coordinated clauses are counted as two T-units. As illustrated below, sentence (1) counts as one T-unit of four words in length. Sentence (2) consists of two coordinated clauses conjoined by '*and*', and it counts as two T-units. Sentence (3) counts as one T-unit, which consists of a main clause '*I like the car*' and a subordinate clause '*that was given to me by my father*'. Measuring by the mean length of T-unit, (3) is the longest and thus considered the most complex.

(1) I like the car. (1 T-unit, 4 words/T-unit)
(2) I like the car, and it was given to me by my father. (2 T-units, 6.5 words/T-unit)
(3) I like the car that was given to me by my father. (1 T-unit, 12 words/T-unit)

Since spoken data are not nearly as tidy or clear-cut as written language, Foster *et al.* (2000: 365) proposed the AS-unit: 'a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either'. Though the AS-unit took care of the elliptical nature of spoken language and dealt with incomplete sentences, it still is mainly based on clause coordination and subordination. Aiming to apply the T-unit in Chinese without distorting Hunt's definition of the T-unit in English, Jiang (2013: 5) proposed a working definition of the T-unit for Chinese: 'A single main clause that contains one independent predicate plus whatever other subordinate clauses or non-clauses are attached to, or embedded within, that one main clause.' One example provided by Jiang was sentence (4), analyzed as three T-units (Jiang 2013: 9). Each self-standing clause in (4) is assumed to be one T-unit. However, a potential problem is that the ratio of clauses per T-unit for any Chinese written output may be equal to a constant 1. While there are multiple levels of complexity (i.e. global, clausal, subclausal, and specific forms) along different dimensions (i.e. length, inner-structure, frequency), the mean T-unit length was at most only able to detect clausal complexity along the dimension of length. Essentially, this practice was reductive, because 'universal' indices based on subordination are not indiscriminately applicable in topic-prominent languages like Chinese. Therefore, the T-unit based on clause subordination ultimately failed to capture the essential typological structural feature of Chinese when used for Chinese complexity analysis.

(4) *Wǒ jiějie jiào Mǎlì, jīnnián èrshí suì, zài Běijīng shàng dàxué.*
我姐姐叫玛丽, 今年二十岁, 在北京上大学。
[I elder sister call Mary, this year twenty year, at Beijing go to college.]
My elder sister is called Mary. She is twenty this year. She goes to college in Beijing.

Adopting subordination-based measures in Chinese complexity analysis encounters challenges due to the three features of Chinese language discussed below: topic-prominence, covert connectives, and the interpretive variability of sentence boundary.

Challenges in adopting subordination-based measures

Topic-prominence Different languages feature different principles and strategies to express meaning. For Chinese, 'the subject is literally the subject matter to talk about, and the predicate is what the speaker comments on when a subject is presented to be talked about' (Chao 1968: 70). In contrast to subject-prominent languages like English, topic-comment structure arguably better reflects the sentence structure of Chinese rather than subject-predicate

(Chao 1968; Li and Thompson 1976; Lü 1979; Huang 1984; Chu 1998). The relationship of the topic and comment is 'aboutness'. Consider the following example sentence:

(5) *Jīntiān xīngqīwŭ.*
今天星期五。
[Today Friday.]
Today is Friday.

Sentence (5) represents a frequently used structure in Chinese, which is composed by only two nouns, *jīntiān* (今天, today) and *xīngqīwŭ* (星期五, Friday), and no verb. *Xīngqīwŭ* (星期五, Friday) is a comment about the topic *jīntiān* (今天, today). If use a subject-predicate framework to analyze such a sentence, we have to assume that there is an omitted verb *shì* (是, is). However, the usage of *shì* (是, is) as in the assumed sentence *Jīntiān shì xīngqī wŭ* (今天是星期五。Today is Friday) only occurs in a contrasting sentence when clarifying the day of the week, as in *Jīntiān shì xīngqī wŭ, búshì xīngqī liù* (今天是星期五, 不是星期六。 It is Friday today but not Saturday). Assuming an omitted verb between the topic and comment may inaccurately transfer syntactic analysis of subject-prominent languages to a topic-prominent language.

When successive comment structures are about the same topic, a topic chain is formed. Sentence (6) below was an example in Xu (1991: 264) that illustrated a typical topic chain:

(6) *Nàgŏu huáng máo, hēi yănquān, cháng shēncái, xì gāo tuĭ, tèbié de xiōngmĕng, yào yăo zhù rén, bújiàn diănr xuèxīng wèir, jué bù piēzuĭ.*
那狗$_i$黄毛, ∅$_i$黑眼圈, ∅$_i$长身材, ∅$_i$细高腿, ∅$_i$特别地凶猛, ∅$_i$要咬住人, ∅$_i$不见点儿血腥味儿, ∅$_i$绝不撇嘴。
[That dog$_i$ yellow haired, ∅$_i$ black eye socket, ∅$_i$ long body, ∅$_i$ thin tall leg, ∅$_i$ particularly DE ferocious, ∅$_i$ once bite-achieved people, ∅$_i$ no see a little smell of blood, ∅$_i$ never opens mouth.]
That yellow haired dog with black eye sockets is tall and has long legs. It is particularly ferocious. Once it bites someone, it will never let go until it draws blood.

In topic chain (6), the topic *nàgŏu* (那狗, that dog) in the first clause was shared as the semantic subjects of all the subsequent clauses in (6), but no phonological form was repeated. These empty phonological shells connect the individual topic-comment structures into a topic chain via coreference to the same topic, realized not in audible phonological form but in a semantic stream. The order of these clauses follows the Principle of Temporal Sequence (Tai 1985), which mirrors a sequential order of events, like a stream of consciousness. As such, we can understand a complex sentence in Chinese to be defined as a topic chain, where a sequence of clauses shares a single topic (Tsao 1979, 1990; Li and Thompson 1981; Chu 1998; Liu 2004).

Given that topic chains describe Chinese supra-clausal relationships better than subordinations do, the application of the T-unit to Chinese complexity analysis is problematic. While sentence (5) may be analyzed as one T-unit, it is debatable how many (eight or one?) T-units there are in sentence (6). With no verbs in the first four parts/clauses of sentence (6), it's not clear whether they should be considered T-units referring to Jiang (2013)'s working definition. Without an appropriate unit that is based on topic chains instead of subordination, Chinese syntactic complexity may be measured inadequately or incorrectly and therefore underrepresented.

Covert connectives In addition to topic-prominence, a second feature that challenges the adoption of subordination-based complexity measures is the use of covert connectives in Chinese. Though previous scholarship has attempted to adapt coordination and subordination for the interpretation of Chinese clause combining, the classification of coordination and subordination in Chinese is based on the semantic relation between clauses. Commonly used conjunctions are only listed in Chinese textbooks to give typical examples of each type but are not mandatory. Chinese clauses can exist in juxtaposition without explicitly indicating their structural relationships. Moreover, native Chinese speakers prefer to use covert instead of overt clause subordination as long as the context provides sufficient semantic information (Li 2005). For example, the following sentences (7) and (8), illustrated in Jin (2006: 122), share the same meaning. Sentences (7) and (8) are identical, except the conjunctions *rúguǒ* (如果, if) and *yīncǐ* (因此, therefore) are overt in (7) but covert in (8). Clauses in juxtaposition, as in (8), imply logical connections instead of explicitly marking the coordination or subordination. In fact, the clauses in (8) sound more close-knit than those in (7) separated by the overt conjunctions and therefore made (8) more complex and sophisticated.

As *rúguǒ* (如果, if) indicates subordination of condition and *yīncǐ* (因此, therefore/so) indicates coordination, (7) could be analyzed as two T-units. The question is then whether (8) should be counted the same number of T-units as if the relationship among its clauses are as overtly marked in (7). Referring to Jiang (2013)'s working definition, (7) and (8) are considered three T-units each, and (7) is assessed more complex than (8) because of its longer mean T-unit length. This contradicts the judgment of Chinese native speakers (Li 2005). Essentially, different measures reveal and affect our understanding and interpretations about *what* complex Chinese is and *how* to compose it.

(7) *Rúguǒ chūle wèntí, dānwèi jiāng quánmiàn fùzé jiějué, yīncǐ, gèrén búbì cāoxīn.*
如果出了问题$_i$, 单位将全面负责解决$\emptyset_i$, 因此, 个人不必操心$\emptyset_i$。
[If comes-out PRT problem$_i$, the work unit will completely take in charge solve $\emptyset_i$, therefore, individual no need worry $\emptyset_i$.]

(8) *Chūle wèntí, dānwèi jiāng quánmiàn fùzé jiějué, gèrén búbì cāoxīn.*
出了问题$_i$, 单位全面负责解决$\emptyset_i$, 个人不必操心$\emptyset_i$。
[Come-out PRT problem$_i$, the work unit will completely take in charge solve $\emptyset_i$, individual no need worry $\emptyset_i$.]
If there is a problem, the work unit will be completely responsible for it. Therefore individuals do not have to worry about it.

Furthermore, without overt conjunction, the speaker's use of either coordination or subordination could be subject to individual interpretation. As Lian (1993) suggested, English sentences are more precise and Chinese sentences are more concise. In (9), the first clause *tīngxìn le tā de huà* (听信了他的话, *believed his word*) is semantically the cause of the subsequent clause *Xǔ Lì chéngle Liú Xiǎoxióng de "qiānyuē yǎnyuán"* (许丽成了刘小雄的"签约演员", Li Xu became Xiaoxiong Liu's contracted actress). Sentence (9) can be marked with the cause and effect conjunction pair *yīnwèi…suǒyǐ…* (因为……所以……, because…so…), and thus exhibit causal subordination *(Yīnwèi) Tīngxìn le tāde huà, (suǒyǐ) Xǔ Lì chéngle Liú Xiǎoxióng de "qiānyuē yǎnyuán"* ((因为) 听信了他的话, (所以) 许丽成了刘小雄的"签约演员"。Because she believed his word, Li Xu became Xiaoxiong Liu's contracted actress). Therefore, (9) could be analyzed as one T-unit. However, without overtly marked connectives, the two clauses in (9) can also be taken as chronologically continuous situations. A different conjunction *yúshì* (于是, thereupon) can be used to exhibit a coordination of continuity between the two actions: *Tīngxìn le tāde huà, (yúshì) Xǔ Lì chéngle Liú Xiǎoxióng de "qiānyuē yǎnyuán"* (听信了他的话, (于是) 许丽成了刘小雄的"签约演员"。Believed his word, and then Li Xu became Xiaoxiong Liu's contracted actress). In this case, sentence (9) would then count two T-units.

(9) *Tīngxìn le tāde huà, Xǔ Lì chéngle Liú Xiǎoxióng de "qiānyuē yǎnyuán".*
听信了他的话, 许丽成了刘小雄的"签约演员"。
[Listen believe ASP he DE word, Li Xu became ASP Xiaoxiong Liu DE 'contracted actress'.]
Deceived by his word, Li Xu became Xiaoxiong Liu's contracted actress.

Since both options for supplied connectives in (9) are grammatical and meaningful, whichever semantic relation was implied or expressed thus depended on the context. Therefore, the two clauses in (9) can be analyzed as two T-units as they exhibit a sequential coordination; Or alternatively, one T-unit as they exhibit a cause-and-effect subordination. Judgment of whether these two clauses are coordinated or subordinated could be subjective rather than explicit. As such, it may not always be accurate to add overt connectives to make explicit the syntagmatic relation between clauses that are connected paratactically. Therefore, it is invalid and unreliable to base complexity analysis for Chinese on the number and length of T-units.

Sentence boundary As topic-prominence and covert connectives in Chinese problematize identifying clause coordination and subordination, the interpretive variability of sentence boundary in Chinese further challenges the use of subordination-based complexity measures. Sentence boundary affects the measurement of complexity, because clause coordination and subordination are typically identified within the unit of the sentence. In addition, the length and numbers of sentences are used for complexity analysis, as in the measure of mean length of sentence (Jin 2006), T-units per sentence (Hunt 1965; Monroe 1975), and sentence complexity ratio (clauses per sentence) (Ishikawa 1995). However, the conventional conception of 'sentence' is likely variable and thus unreliable when applied as the sentential unit for complexity analysis in Chinese. In a study conducted by Tsao (1990), 18 Chinese ESL college students were asked to apply punctuation marks in two Chinese written passages and two English passages, where the original punctuation marks had been removed. Interestingly, the results showed that the students, who are all native speakers of Chinese, disagreed considerably both among themselves and with the original author as to the numbers of sentences contained in the Chinese paragraphs. Contrastively, in their punctuation of the two English texts, these Chinese native speakers, who were far from having a native command of English, showed considerably more agreement among themselves and with the original author about sentencehood. A paragraph-based punctuation study by Chu (1998) confirmed Tsao's findings. When it comes to complexity assessment, while coding of written data may follow the original sentence segmentation, the subjective marking of sentence numbers causes problems for transcribing and coding spoken Chinese reliably.

Topic chain in Chinese complexity analysis

Given that subordination does not reflect the Chinese complexity composition due to the aforementioned typological differences, one alternative is to consider how topic chain form Chinese syntactic complexity. With their findings that the mean T-unit length of native Chinese speakers was shorter than that of L2 Chinese speakers, Jin (2006) and Yuan (2009) questioned the validity of the T-unit for Chinese complexity analysis and called for complexity measures tailored to the topic prominence of Chinese. Capturing the topic-prominence feature of Chinese language, Jin's (2006) pioneering work adopted Li's (2005) definition of topic chain and introduced the Terminal Topic-Comment Unit for Chinese syntactic complexity analysis. Though Li's (2005: 25) ten topic chain patterns provide great insights on how topic chains are composed, her categorization, in line with many other proposals and amendments on topic chain (LaPolla 1995; Xu and Liu 1998; Jiang 2011), have come generally from a descriptive rather than an operational perspective. But a descriptive definition of topic chains is not tailored for systemic complexity analysis.

Applying a descriptive definition may fail to comprehensively clarify the number and nature of the constituent components, the relationships thereof, and the entity in complexity composition by Bulté and Housen's definition (2012). For instance, in Jin's (2006) attempt to apply topic chain in complexity analysis, essential questions remained unanswered regarding the conceptualization and operationalization of her proposed Terminal Topic-Comment Unit. First, since topic is a thematic role or information unit, it lacks formal cues to identify the topic. Second, it was unclear how to identify the beginning as well as the end of a topic chain. The boundary of a topic chain was confused with a sentence boundary in Li's (2005) practice. Jin's proposed Terminal Topic-Comment Unit ran into the same problem with the subjective sentence boundary as the T-unit does. Third, no answer was provided regarding the composition of a topic chain and its relationship to other, non-topic-chain output. Without answering these questions, the application of topic chains to Chinese complexity analysis is limited and debatable. Therefore, clarification on the conceptualization and operationalization of topic chains is crucial for a valid complexity analysis.

AN ORGANIC COMPLEXITY MEASURE

When investigating complexity via varied syntactic levels, the underlying unit for complexity composition should be organic. That is to say, complexity analysis should avoid indiscriminately applying one unit for all languages, but should design measures to account for the nature of the constituent components and the relationships thereof at global, clausal, and subclausal levels. This study therefore expands on the initial framework for organically developing complexity measures outlined by Norris and Ortega (2009), and proposes an operationalizable unit of topic chain, in addition to clause coordination and subordination, as one of the complexity forms. It should be noted that there is also the possibility of other form(s) of clause combining according to typological differences in other languages.

TC-unit

This study proposes a taxonomy of TC-units, as illustrated in Figure 1, to capture the supra-clausal and clausal level structures in Chinese. A *terminable TC-*

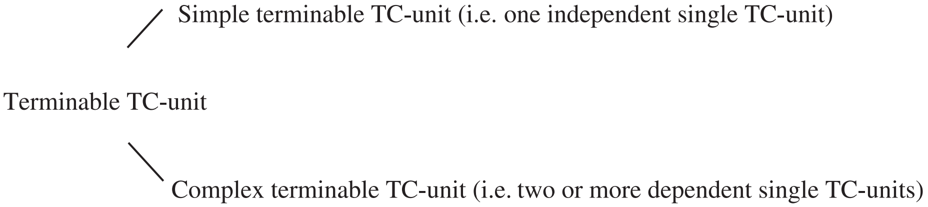
Simple terminable TC-unit (i.e. one independent single TC-unit)
Terminable TC-unit
Complex terminable TC-unit (i.e. two or more dependent single TC-units)

Figure 1: Taxonomy of TC-units

unit is a supra-clausal-level unit that takes the form of a topic chain. A *single TC-unit* refers to one clausal or subclausal level unit in Chinese. Depending on the number of single TC-units it consists of, a terminable TC-unit is categorized into a *simple terminable TC-unit* or a *complex terminable TC-unit*. Whether a complex terminable TC-unit is formed depends on *coreferential zero*. *Coreferential zero* is an element that does not have any phonological form and is unpronounced but corefers to the full-form topic mentioned in the previous or subsequent single TC-unit. A *simple terminable TC-unit* consists of one independent single TC-unit. The topic of a simple terminable TC-unit is not repeated in the form of coreferential zeros in the preceding or subsequent topic-comment structures. A *complex terminable TC-unit* consists of two or more successive dependent single TC-units. The topic of these dependent single TC-units only appears once in its full form and is repeated in the form of coreferential zero in the rest of dependent single TC-units. Coreferential zero is the requisite part, integrating single TC-units into a complex terminable TC-unit (in the form of a topic chain). In a complex terminable TC-unit, the topic is not identified according to its thematic prominence, which could be subject to individual interpretation and is thus unreliable in assessment. Instead, what is repeated in the form of coreferential zero is identified as the topic of a complex terminable TC-unit. In addition, coreferential zero is the key to identifying the starting and end points of a complex terminable TC-unit and therefore avoids the subjective determination of a conventional sentence boundary. The beginning and end points of a terminable TC-unit do not necessarily take the form of a conventional sentence with punctuation marks. Upon the introduction of a new topic, or the repetition of a topic in its full form, in a pronoun, or in a demonstrative (instead of coreferential zero), a new terminable TC-unit is activated. With this taxonomy of TC-units, complexity composition can be interpreted as single TC-units composing a terminable TC-unit, in addition to clause coordination or subordination.

Operationalizing the TC-unit

The following two examples in Figure 1 illustrate how to operationalize the taxonomy of TC-units. Example (10) consists of two single TC-units, each a simple terminable TC-unit. Though the two successive single TC-units share the same topic *tā* (她, she), (10) is not a complex terminable TC-unit, since the two topics are both expressed in the form of pronouns, not repeated in coreferential zero. The two single TC-units in (10) are thus less closely tied together and more equal in isolation. In contrast, in (11), after its first mention, the topic *nǚde* (女的, female) is repeated unpronounced seven times in the subsequent single TC-units. A total of eight single TC-units are dependent and compose one complex terminable TC-unit. Notice that the complex terminable TC-unit in (11) goes beyond the boundary of one sentence marked by a period.

(10) *Tā xǐ gǒu. Tā cā gǒu de tóu.*
‖她$_i$洗狗。‖ 她$_j$擦狗的头。‖ (2 simple terminable TC-units)
[She$_i$ wash dog. She$_j$ wipe dog DE head.]
She washed the dog. She wiped the dog's head.

(11) *Nǚde guì xiàlái, kāishǐ gěi xiǎo gǒu xǐzǎo, yòng shuāzi bǎ tā shuā de hěn xìxīn. Ránhòu, xǐ wán zǎo hòu, yòng máojīn bǎ xiǎo gǒu cā gān, ránhòu hái gěi tā shū máo. Wánle zhīhòu, yòu hǎoxiàng ránhòu duì zìjǐ hěn mǎnyì, gōngzuò zuò de hěn hǎo.*
‖女的$_i$跪下来，| ∅$_i$开始给小狗洗澡，| ∅$_i$用刷子把它刷得很细心。| 然后，∅$_i$洗完澡后，| ∅$_i$用毛巾把小狗擦干，| ∅$_i$然后还给它梳毛。| 完了之后，∅$_i$又好像然后对自己很满意，| ∅$_i$工作做得很好。‖ (1 complex terminable TC-unit consists of 8 dependent single TC-units)
[Woman$_i$ kneel down-come, ∅$_i$ start for dog shower, ∅$_i$ use brush PREP-it brush DE very careful. Then, ∅$_i$ wash-completed shower after, ∅$_i$ use towel PREP–dog wipe dry, ∅$_i$ then also PREP-it comb hair. Finish-PRT later, ∅$_i$ also seems then PREP-self very satisfied, ∅$_i$ job done DE very well.]
The woman kneeled to start washing the dog by carefully brushing it with a brush. Then, after the shower, she dried the dog with towel. Further, she even combed the dog's hair. After all this, she seemed very satisfied with her own work. She thought she did a very good job.

The previously mentioned Example (4) was analyzed as three T-units in Jiang (2013) and the sentence counts as three single TC-units. '*Wǒ jiějie*' (我姐姐, my sister) is the topic and is repeated in the form of coreferential zero twice. Whether these three single TC-units form one terminable complex TC-unit depends on the topic of their previous and successive TC-unit(s). In (4a) below, '*wǒ*' (我, I) is introduced as a different topic. The first three dependent single TC-units sharing the same topic '*wǒ jiějie*' (我姐姐, my sister) thus form one terminable complex TC-unit. The different topic '*wǒ*' (我, I) starts another simple terminable TC-unit, which consists of one independent single TC-unit without continued text. Both the length and the inner-structure of the TC-units are indices of complexity.

(4a) *Wǒ jiějie jiào Mǎlì, jīnnián èrshí suì, zài Běijīng shàng dàxué. Wǒ hěn xǐhuan tā.*

‖我姐姐$_i$叫玛丽，| $\emptyset_i$今年二十岁，| $\emptyset_i$在北京上大学。‖ 我$_j$很喜欢她。‖

(2 terminable TC-units; 4 single TC-units)

[I elder sister$_i$ call Mary, $\emptyset_i$ this year twenty year, $\emptyset_i$ at Beijing go to college. I$_j$ very like her.]

My elder sister is called Mary. She is twenty this year. She goes to college in Beijing. I like her very much.

TC-unit-based measures

In order to investigate Chinese syntactic complexity via global, clausal, sub-clausal, and specific form levels (see Norris and Ortega (2009) for a review), the present study suggests at least seven measures, as listed in Table 1. The measures here proposed can be grouped into three categories: length, frequency, and ratio. Length measures are mean length of terminable TC-unit and mean length of single TC-unit (both independent and dependent). Ratio measures are the ratio of complex terminable TC-unit to all terminable TC-units (both simple and complex), the ratio of different types of terminable TC-unit, single TC-units (independent or dependent) per terminable TC-unit (both simple and complex), and dependents per head. A frequency measure is the frequency of a specific form. The number of characters was used to measure the length of the unit for analysis, because length measures coded in Chinese characters were found more reliable than those coded in Chinese words (Jiang 2013).

RESEARCH QUESTION

The triad of CAF has been widely used to assess language performance and development. The general hypothesis underlying complexity development is that the higher one's proficiency level is, the more complex language one is able to produce. One way of validating complexity measures is to see to what extent the measure-based results distinguish the proficiency level of the target language speakers. Therefore, the research question is: to what extent do combinations of the four measures (Table 1) of Chinese syntactic complexity distinguish between low, high, and native proficiency speakers of Chinese?

Table 1: Inventory of Chinese syntactic complexity measures

Complexity	Measures
Global	• *Mean length of terminable TC-unit • *Complex terminable TC-unit/all terminable TC-units (both simple and complex) • Ratio of different types of terminable TC-unit
Clausal	• *Mean length of single TC-unit (both independent and dependent) • *Single TC-units (independent or dependent) per terminable TC-unit (both simple and complex)
Subclausal/phrasal	• Dependents per head
Specific form	• Frequency of a specific form

Note: This study tested the four measures marked with*.

METHOD

Participants

The participants included both English speaking Chinese L2 speakers and Chinese L1 speakers, to cover the full spectrum of Chinese language proficiency. All the L2 participants had taken a minimum of one year of Chinese language courses at colleges in USA or China. They were undergraduates, graduate students, as well as Chinese language teachers, faculty, and staff who had learned and used Chinese in work for up to 40 years. All the L1 participants were Chinese native speakers who had received or were working on their bachelor's or master's degrees at top national universities in Beijing and had no problem comprehending the English video.

A total of 115 complete spoken data sets and 116 complete written data sets produced by 118 participants were included for analyses, excluding data sets with missing files and/or mismatched dominant language (other than English). All the L2 participants ($n = 86$) were divided into two groups based on their Chinese proficiency as assessed by a Mandarin Elicited Imitation (EI) test (see more in the section *Instruments and procedure* below) developed by Zhou and Wu (2009). As shown in Figure 2, the EI scores of all the L2 participants fell clearly in a bimodal distribution, with a lower proficiency group and a higher proficiency group. The cut-point between the two proficiency groups fell right in the middle of the test, at a score of 60 out of 120. Therefore, L2 participants were divided into two groups based on their Chinese proficiency: (a) *Group Low*: EI score < 60 ($n = 38$, $M = 35.16$, $SD = 12.17$); (b) *Group High*: EI score $\geq$ 60 ($n = 48$, $M = 87.96$, $SD = 15.13$). There was a substantial and meaningful difference between Group Low and Group High in terms of Chinese language proficiency. A *t*-test between the two groups' mean EI

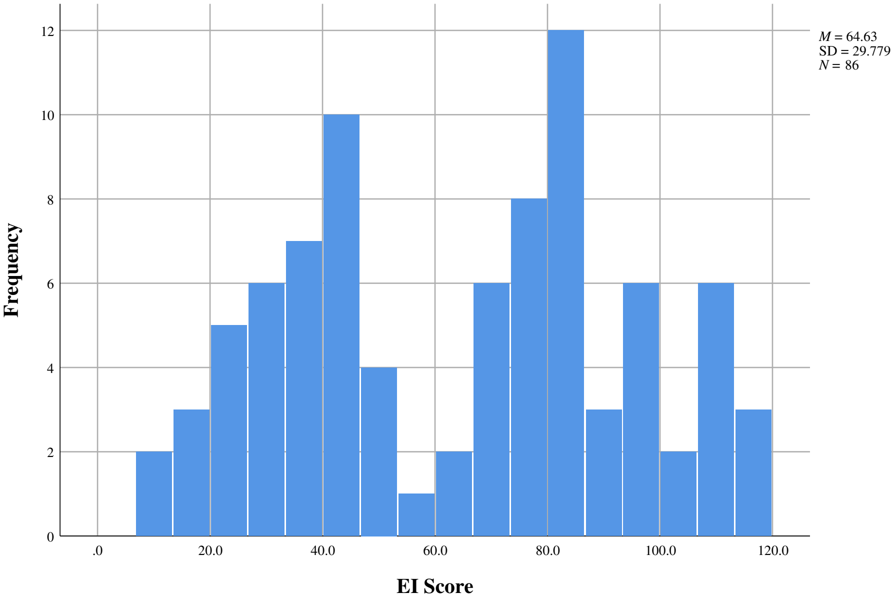

Figure 2: Mandarin EI score distribution of L2 Chinese speaker participants

scores showed a statistically significant difference, with t (84) = −17.48, $p = 0.000$ ($p < 0.05$). All the Chinese L1 speakers ($n = 32$) formed (c) *Group Native*. Table 2 lists the participant distribution in each proficiency group for spoken and written output respectively. All the data in this study were analyzed using IBM SPSS Statistics for Macintosh, Version 26.0.

Instruments and procedure

All the participants completed a designed *Chinese Timed Writing and Speaking Test* (TW&ST) (Pronounced TWIST) online in a computer laboratory setting. TW&ST provided two versions: (i) TW&ST (English instructions, 45 minutes) for L2 participants; and (ii) TW&ST (Chinese instructions, 35 minutes) for L1 participants with no EI test. For this study, participants completed two speaking tasks, two writing tasks, and one proficiency test: (i) a comic strip descriptions (CS) task; (ii) a video story retelling (VR) task; (iii) a free writing task (FW); (iv) a guided re-writing task (GR); and (v) a Mandarin EI test. A short preparation was provided prior to the tasks, to familiarize participants with the interface of TW&ST. TW&ST strictly regulated and standardized the planning and performance time for each task across individuals. Except for the instruction screens, the display time of each screen of the TW&ST was pre-set. The test was designed such that pause, forward, and backward functions were all disabled. Participants were notified of the time limit for each task before they proceeded. The remaining time of display for each screen was also

Table 2: Total cases for speaking and writing tasks by proficiency group

Proficiency group	EI score	Spoken ($N = 115$)	Written ($N = 116$)
Group Low	<60	37 (17 males, 20 females)	36 (17 males, 19 females)
Group High	60–120	47 (21 males, 26 females)	48 (22 males, 26 females)
Group Native	Native speakers	31 (11 males, 20 females)	32 (12 males, 20 females)

available to the participants. Upon reaching the time limit of each screen, the computer automatically proceeded to the next slide. Throughout the whole test, no note-taking was allowed.

Speaking tasks For the CS task, participants had 30 seconds to prepare and up to 90 seconds to tell a story based on a comic strip on a daily life topic. When time was up, the computer automatically proceeded to the next speaking task, the VR task. To complete the VR task, participants first watched a short video on a story about Chinese *Nian* narrated in English. After 30 seconds to prepare, participants then retold the story in Chinese for up to three minutes. Upon reaching the time limit of each speaking session, the recording of the participants' speaking was automatically converted into mp3 files and immediately uploaded to the database online.

Writing tasks For the FW task, participants used up to seven minutes to write a well-organized composition in seven minutes on their relationship with their *father/mother/brother/sister/friend (choose any one of them)*. The GR task provided seven semantically coherent but formally incohesive sentences (lexically controlled at an elementary level) for the participants to connect into one cohesive paragraph in five minutes. Participants could manipulate the sentences by combining them, adding or omitting words, and changing the order of words, but were required not to leave out any of the given information. The provided sentence skeleton allowed for different supra-clausal- and clausal-level combinations and reformatting, as well as the potential for drawing out more variability of syntactic complexity from Chinese speakers of various proficiency levels. The writing tasks allowed typing online and accepted the Chinese romanization system *pinyin*, which excluded the variable of character scripting from Chinese writing proficiency.

Mandarin EI test The Mandarin EI test (see transcript in Zhou 2012: 188) was included at the end of TW&ST (English instructions) to assess the global Mandarin competence of L2 participants. The EI test has been used in second-language proficiency assessment and has proved to be an effective language test with high correlations with the *Oral Proficiency Interview* (Ortega *et al.*

2002; Erlam 2006). The Mandarin EI test has been found to be a valid and reliable tool for Mandarin proficiency assessment (Zhou and Wu 2009; Zhou 2012; Wu and Ortega 2013).

Completion of the EI test required 10 minutes and 40 seconds. The participants were instructed to listen to 30 Chinese sentences of varied length, vocabulary, and grammar structures in sequence, and to repeat each sentence as exactly as possible in the time provided after hearing each sentence. Two raters each rated all 86 test files containing 2,580 responses (86 participants * 30 sentences) referring to a five-point rubric (developed by Ortega *et al.* (2002), with Mandarin examples provided in Zhou (2012: 190)). Before rating all the test files, the two test raters piloted 10 participants separately, compared their rating for each item in the EI test, and made sure they were using the rubric in a consistent way. The inter-rater reliability was good (agreement rate = 75.1 per cent, Cohen's weighted kappa = 0.83) (Altman 1991). The mean score of the two ratings given by the two raters were assigned to all the participants as their final EI score.

Scoring and analysis

Transcribing The researcher transcribed all the collected spoken output twice to maximize the accuracy of the transcription. The researcher segmented the spoken data with '/' rather than adding subjective punctuation marks. The study focuses on complexity and excludes accuracy and fluency factors such as false starts, fillers, and back channel cues and fillers. If there was self-repair, the corrected language form was saved without the part before self-correction. All the written output collected via TW&ST was saved without the need for any transcription.

Measures Four of the proposed syntactic complexity measures in Table 1 were applied to analyze the transcribed data: mean length of terminable TC-unit (MLTTCU), complex terminable TC-unit/all terminable TC-units (both simple and complex) (CTTCU/ATTCU), mean length of single TC-unit (both independent and dependent) (MLSTCU), and single TC-units per terminable TC-unit (STCU/TTCU).

Coding All the data were listed by participant in separate Excel files by task. The coding involved three steps. First, the researcher *identified* and marked on files the boundary of each terminable TC-unit as well as each single TC-unit. Second, in the Excel file, where each single TC-unit was saved in one cell, a formula '=LEN(A1)' was applied to *count* the length of each terminable TC-unit and single TC-unit (in characters). The total number of terminable TC-units, complex terminable TC-units, and single TC-units in each participant's output of each task were also respectively counted. Last, scores on each of the four measures were *calculated*.

Table 3: A coding sample (12)

Coding (12)	STCU	STCU length (in characters)	TTCU	TTCU length (in characters)	CTTCU
在这个漫画里面妈妈$_{1}$好像刚准备好晚餐	1	17	1	24	1
$\emptyset_{1}$把它带到桌子上	1	7			
爸爸$_{2}$已经坐下准备吃饭了	1	11	1	11	0
那可是孩子$_{3}$还没来	1	8	1	8	0
所以呢妈妈$_{4}$叫爸爸去找在隔壁房间的孩子	1	18	1	18	0
爸爸$_{5}$发现孩子在那儿趴着看书	1	13	1	13	0
他$_{6}$把孩子叫去吃饭去了	1	10	1	10	0
可是呢书$_{7}$也引起了爸爸的注意	1	13	1	13	0
后来我们$_{8}$看到了妈妈和孩子坐在桌子旁边等着不在的爸爸	1	25	1	25	0
然后妈妈$_{9}$叫孩子去找爸爸	1	11	1	11	0
孩子$_{10}$回到原处看书的地方	1	11	1	29	1
$\emptyset_{10}$发现爸爸正趴在地上看着孩子刚在看的书	1	18			
Total	(12)	(162)	(10)	(162)	(2)

A coding sample is provided in Table 3 to illustrate the three steps. As the first coding step, all terminable TC-units (TTCU) and single TC-units (STCU) were segmented based on the topic and its repetition(s) in the form of coreferential zero(s). Next, each individual STCU was listed in the first column of Table 3. In the column of *STCU*, in each cell a constant 1 was assigned to mark a STCU listed. The length of each STCU was then counted in characters and recorded in the column of *STCU length (in characters)*. Each terminable TC-unit was then assigned a constant 1 in the column of *TTCU*. For example, the topic of the first STCU, 妈妈 (*māma*, mother), was repeated in the form of coreferential zero in the second STCU; therefore, these two consecutive STCU formed one complex terminable TC-unit (CTTCU) as marked in the column of *CTTCU*. The number of characters used in each TTCU was then counted, and the value recorded in the column of *TTCU length (in characters).* Topic marks were not included for length counting. The last step of score calculation for each measure was conducted by using formulas in the Excel file. Examples of these calculations are listed in Table 4. For all data, the researcher coded each segment twice. The intra-rater reliability indices between the two coding for each task of each participant on each measure showed high correlation coefficients ranging between 0.85 and 0.99.

Table 4: Calculated scores of coding (12) on four measures

Task	ID	MLTTCU	CTTCU/ATTCU	MLSTCU	STCU/TTCU
CS	1	16.20 (=162/10)	0.20 (=2/10)	13.50 (=162/12)	1.20 (=12/10)

Inter-rater reliability

Two trained raters independently analyzed 20 per cent of the data, amounting to 96 output cells (8 participants * 3 proficiency groups * 4 tasks), randomly selected. The raters then met to check the identical demarcation of TC-units and resolve the differences through discussion, as shown in Tables 5 and 6. No significant disputes arose after discussion. Following Jiang (2013), the inter-rater reliability was calculated by dividing the number of identical TC-units by the agreed number of TC-units after discussion. Inter-rater reliability was high both for the number of single TC-units (98.5 per cent = 1317/1337) and for the number of complex terminable TC-units (92.5 per cent = 223/241).

RESULTS

Data screening

Before performing discriminant analysis, data in the 48 cells (4 tasks * 3 proficiency groups * 4 measures) of this design were checked for outliers, normality, homogeneity of variance, homogeneity of variance-covariance matrices, linearity, and multicollinearity. No violation of the respective statistic assumptions was found. Out of 48 groups of scores, a total of eight individual univariate and multivariate outliers were detected, modified, or deleted. Since discriminant analysis is quite robust and minimally affected by nonnormal distributions, especially when the sample sizes in this study were large, analyses on the unadjusted data were kept in this study. It was also confirmed that the results of discriminant analyses on the data with or without adjustment were similar.

Descriptive statistics

The mean values for each measure on four-task production by proficiency level are listed in Table 7. As displayed below in Figure 3, each of the four measures was able to, with varied power, distinguish the varied complexity level among proficiency groups Low, High, and Native.

MLTTCU Note especially on the measure *Mean length of terminable TC-unit* (MLTTCU), the distributions of the scores for each proficiency group in all the four tasks were quite evenly spread, which indicates a great power of measure MLTTCU at distinguishing the syntactic complexity levels of language output

Table 5: Number of single TC-units in 20 per cent of data

Group	Data	Rater 1	Rater 2	Identical	Agreement (%)	Agreed after discussion
Low	Spoken	173	177	170	97.1	177
	Written	110	109	109	99.5	110
High	Spoken	320	322	316	98.4	321
	Written	125	122	121	98.0	124
Native	Spoken	398	404	393	98.0	401
	Written	215	212	208	97.4	204
Total		1,341	1,346	1,317	98.0	1,337

Note: Agreement = the number of identical single TC-units/the mean number of single TC-units counted by Rater 1 and Rater 2.

Table 6: Number of complex terminable TC-units in 20 per cent of data

Group	Data	Rater 1	Rater 2	Identical	Agreement (%)	Agreed after discussion
Low	Spoken	17	17	14	82.4	18
	Written	15	13	13	92.9	13
High	Spoken	62	61	55	89.4	63
	Written	18	16	15	88.2	16
Native	Spoken	86	82	78	92.9	83
	Written	48	49	48	99.0	48
Total		246	238	223	92.1	241

Note: Agreement = the number of identical complex terminable TC-units/the mean number of complex terminable TC-units counted by Rater 1 and Rater 2.

across proficiency groups. MLTTCU seemed to be a strong indicator of the Chinese syntactic complexity itself. The mean score distribution on measure MLTTCU, 8.88–13.27 characters for Group Low, 12.91–15.45 characters for Group High, and 17.83–19.94 characters for Group Native, indicated that on average longer terminable TC-units were produced by participants of higher proficiency while shorter terminable TC-units were produced by those of lower proficiency.

MLSTCU The mean scores on measure *Mean length of single TC-unit* (MLSTCU) were mostly evenly spread out. For the speaking tasks, the mean score of the Group Low on measure MLSTCU ranged from 8.21–8.96 characters, 10.38–10.82 characters for the Group High, and 12.27–13.14 characters for the Group Native. It generally indicated that for spoken Chinese on average longer

Table 7: The mean values for each measure on four-task production by proficiency level

Task	Group	MLTTCU	CTTCU/ATTCU	MLSTCU	STCU/TTCU
CS	Low (37)	8.88	0.08	8.21	1.09
	High (47)	12.91	0.21	10.38	1.25
	Native (31)	17.83	0.37	12.27	1.48
VR	Low (37)	9.82	0.09	8.96	1.10
	High (47)	14.65	0.26	10.82	1.38
	Native (31)	19.74	0.35	13.14	1.52
FW	Low (36)	10.73	0.09	9.77	1.08
	High (48)	14.23	0.17	11.69	1.22
	Native (32)	19.12	0.45	10.74	1.82
GR	Low (36)	13.27	0.09	12.23	1.09
	High (48)	15.45	0.25	12.16	1.31
	Native (32)	19.94	0.32	15.00	1.39

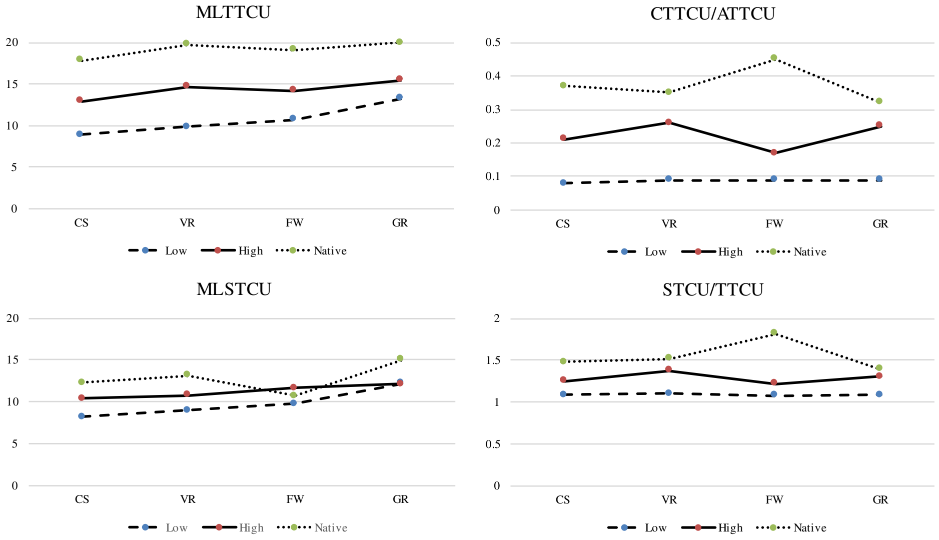

Figure 3: The mean values for each measure on four-task production by proficiency level

single TC-units were produced by participants of higher proficiency while shorter single TC-units were produced by those of lower proficiency.

As for each of the two writing tasks however, the distribution showed some crossover between different groups. In the FW task, the mean score of 11.69 characters for the Group High participants surpassed the mean score of 10.74

characters for Group Native participants on measure MLSTCU. In the GR task, the mean score of 12.23 characters for Group Low was a little higher than the mean score of 12.16 characters for Group High participants on measure MLSTCU. This suggested that for the writing tasks, there might be a nonlinear clausal complexity development in terms of the length of single TC-units as proficiency increases. Another possible interpretation for such score crossover is that for writing tasks, a longer single TC-unit does not necessarily equal more sophisticated or more native-like Chinese syntactic complexity.

CTTCU/ATTCU On measure *Complex terminable TC-unit/all terminable TC-units* (CTTCU/ATTCU), for all the four tasks, participants of higher proficiency group scored higher than those of lower proficiency groups. Generally, it suggested that on average more complex terminable TC-units were produced by higher proficiency participants than lower proficiency participants.

The FW task showed a somewhat different score distribution on this measure as shown in Figure 3 where the gap between the mean scores of Group High and Group Native was stretched when compared with the other tasks. In the FW task, the score of Group High on measure CTTCU/ATTCU was relatively lower than it was for the other three tasks performed while conversely, for Group Native the score was higher.

STCU/TTCU Same on the other ratio measure *Single TC-units per terminable TC-unit* (STCU/TTCU), for each task, participants of higher proficiency scored higher than the participants of lower proficiency. It suggested that one terminable TC-unit consisted of more single TC-units in the output of higher proficiency participants, while the terminable TC-unit produced by lower proficiency participants consisted of less single TC-units.

Similar to the distribution on measure CTTCU/ATTCU for the FW task, the gap between the average scores of Group High and Group Native was stretched compared with the other four tasks on measure STCU/TTCU. This as well was caused by the relatively lower average score of Group High compared with the relatively higher score of Group Native. This indicated a relatively large gap between Group High and Group Native participants in terms of producing complex terminable TC-units consisting of more dependent single TC-units. Native Chinese speakers produced more complex terminable TC-units consisting of more dependent single TC-units than the advanced L2 Chinese speakers in writing.

Discriminant analysis

Direct discriminant analysis (also called discriminant function analysis) predicts group membership based on a set of predictors (Brown *et al.* 2001; Norris 2015). The study took a differential groups approach to validation by investigating how well the membership can be correctly predicted by the syntactic complexity measures. Three groups, Group Low, High, and Native, were predetermined by the participants' three Chinese proficiency levels based on their Mandarin EI scores or native speaker status. The four syntactic complexity measures were then applied by task as predictors of membership in the three proficiency groups, since higher proficiency Chinese speakers should be able to produce Chinese output with higher syntactic complexity. A separate discriminant analysis was conducted for each task using SPSS. For all the discriminant analyses, prior probabilities were computed from group sizes.

CS task output A direct discriminant function analysis identified two discriminant functions, the first accounting for the large majority (96.2 per cent) of observable between-groups variance across the three proficiency groups, and the second accounting for 3.8 per cent. An overall statistically significant effect was found for the combined functions (1 and 2), Wilks' lambda = 0.419, χ^2 (8, $N = 115$) = 96.201, $p = 0.000$, indicating that the combined predictor variables were able to account for around 58 per cent of the actual variance in proficiency level among the three groups. On its own, the second function did not provide additional statistically significant predictions, Wilks' lambda = 0.952, χ^2 (3, $N = 115$) = 5.491, $p = 0.139$. Figure 4 shows the individual cases and group centroids (average values for each group) displayed in two dimensions: (i) from left to right, Function 1 clearly distinguished between all three groups; (ii) from top to bottom, Function 2 additionally provided little distinguishment between Group High and the other two. The classification results for the CS task indicated that, overall, the combined Functions 1 and 2 were able to correctly classify 78 cases (or 67.8 per cent) as shown in Table 8. However, the accuracy of the classifications varied for the three levels. Group Low participants were classified with 73.0 per cent accuracy, while Group High participants were classified with 68.1 per cent accuracy, and Group Native participants were classified correctly with 61.3 per cent accuracy.

VR task output Both the combined functions (1 and 2) and the second function alone showed statistically significant effect in distinguishing proficiency level between the three groups. Function 1 accounted for the large majority (90.4 per cent) of observable between-groups variance across the three proficiency groups, and Function 2 accounting for 9.6 per cent. An overall statistically significant effect was found for the combined functions (1 and 2), Wilks' lambda = 0.270, χ^2 (8, $N = 115$) = 144.720, $p = 0.000$, indicating that the combined predictor variables were able to account for around 73 per cent of the actual variance in proficiency level between the three groups. On its own, the second function provided additional statistically significant predictions, Wilks' lambda = 0.822, χ^2 (3, $N = 115$) = 21.646, $p = 0.000$. Figure 4 shows the individual cases and group centroids (average values for each group) displayed in two dimensions: (i) from left to right, Function 1 clearly distinguished between all three groups; (ii) from top to bottom, Function 2 additionally distinguished Groups High from the other two. For the VR task,

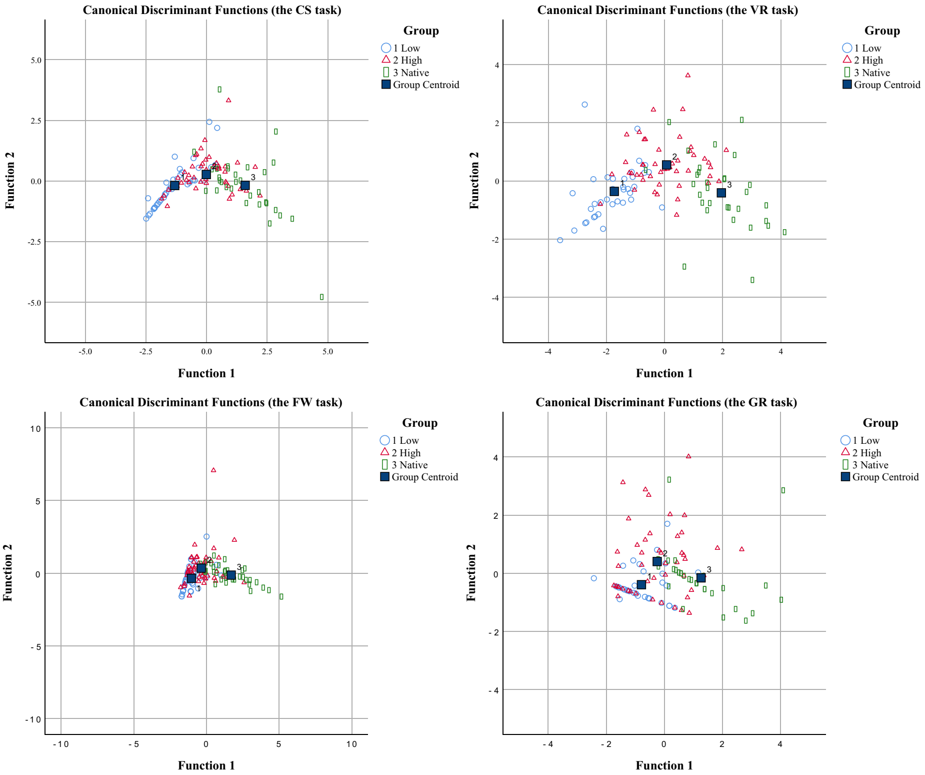

Figure 4: Predicting proficiency groups by four measures: cases and group centroids for two discriminant functions of the CS, VR, FW, and GR task.

the classification results indicated that, overall, 87 cases (or 75.7 per cent) were correctly classified as shown in Table 8. However, the accuracy of the classifications varied from 72.3 per cent to 81.1 per cent for the three groups.

FW task output Both the combined functions (1 and 2) and the second function alone showed statistically significant effect in distinguishing proficiency level between the three groups. Function 1 accounted for the large majority (92.7 per cent) of observable between-groups variance across the three groups, and Function 2 accounting for 7.3 per cent. An overall statistically significant effect was found for the combined functions (1 and 2), Wilks' lambda = 0.412, χ^2 (8, $N = 116$) = 98.848, p = 0.000, indicating that the combined predictor variables were able to account for around 59 per cent of the actual variance in proficiency level between the three groups. On its own, the second function also provided additional statistically significant predictions, Wilks' lambda = 0.912, χ^2 (3, $N = 116$) = 10.239, p = 0.013. Figure 4 shows the individual cases and group centroids (average values for each group) displayed in two dimensions: (i) from left to right, Function 1 distinguished between all

Table 8: Classification results for the CS, VR, FW, and GR task with four measures as the predictors

Actual group	Predicted group membership accuracy						Actual group	Predicted group membership accuracy					
	CS (67.8%)			VR (75.7%)				FW (67.2%)			GR (61.2%)		
	Low	High	Native	Low	High	Native		Low	High	Native	Low	High	Native
Low (37)	**27**	10	0	**30**	7	0	Low (36)	**22**	13	1	**25**	10	1
(%)	**73.0**	27.0	0.0	**81.1**	18.9	0.0	(%)	**61.1**	36.1	2.8	**69.4**	27.8	2.8
High (47)	8	**32**	7	6	**34**	7	High (48)	9	**35**	4	14	**27**	7
(%)	17.0	**68.1**	14.9	12.8	**72.3**	14.9	(%)	18.8	**72.9**	8.3	29.1	**56.3**	14.6
Native (31)	0	12	**19**	0	8	**23**	Native (32)	0	11	**21**	0	13	**19**
(%)	0.0	38.7	**61.3**	0.0	25.8	**74.2**	(%)	0.0	34.4	**65.6**	0.0	40.6	**59.4**

Note: The bold values are the correct predictions.

three groups, and much more so between Group Native and the other two; (ii) from top to bottom, Function 2 additionally distinguished Groups High from the other two. The classification results indicated that, overall, 78 cases (or 67.2 per cent) were correctly classified as shown in Table 8. However, the accuracy of the classifications varied from 61.1 per cent to 72.9 per cent for the three groups.

GR task output Both the combined functions (1 and 2) and the second function alone showed statistically significant effect in distinguishing proficiency level between the three groups. Function 1 accounted for the large majority (84.3 per cent) of observable between-groups variance across the three groups, and Function 2 accounting for 15.7 per cent. An overall statistically significant effect was found for the combined functions (1 and 2), Wilks' lambda = 0.530, χ^2 (8, N = 116) = 70.815, p = 0.000, indicating that the combined predictor variables were able to account for around 47 per cent of the actual variance in proficiency level between the three groups. On its own, the second function also provided additional statistically significant predictions, Wilks' lambda = 0.888, χ^2 (3, N = 116) = 13.191, p = 0.004. Figure 4 shows the individual cases and group centroids (average values for each group) displayed in two dimensions: (i) from left to right, Function 1 distinguished between all three groups, and much more so between Group Native and the other two; (ii) from top to bottom, Function 2 additionally distinguished Groups High from the other two. The classification results indicated that, overall, 71 cases (or 61.2 per cent) were correctly classified as shown in Table 8. However, the accuracy of the classifications varied from 56.3 per cent to 69.4 per cent for the three groups.

MLTTCU as the only predictor for speaking tasks As aforementioned, Function 1 accounted for 96.2 per cent (p = 0.000) and 90.4 per cent (p = 0.000) of the between-groups variance across the three proficiency groups in discriminating on the two speaking tasks respectively. In addition, for the CS and VR tasks, as shown in Table 9, the predictor variable MLTTCU showed the absolutely highest correlation, r = 0.96 and 0.99, with the first function. This indicated that in both discriminant analyses for the two tasks, each Function 1 was best represented by the global complexity measure MLTTCU. Therefore, MLTTCU was chosen to be applied as the only predictor and it generated approximately the same classification results as applying all the four predictors together. The reliability of the discriminant function for each of the two tasks was all found to be statistically significant when MLTTCU was applied as the only predictor variable. For the CS task, Wilks' lambda = 0.459, χ^2 (2, N = 115) = 87.312, p = 0.000, and a total of 78 (or 67.8 per cent) of the cases were correctly classified as shown in Table 10. For the VR task, Wilks' lambda = 0.332, χ^2 (2, N = 115) = 123.369, p = 0.000, and a total of 85 (or 73.9 per cent) of the cases were correctly classified as shown in Table 10. Comparing to

Table 9: Two functions of each discriminant analysis for the four tasks

Predictor variable	Correlations of Predictor Variables with Discriminant Functions							
	CS		VR		FW		GR	
	1	2	1	2	1	2	1	2
① MLTTCU	**0.96***	0.01	**0.99***	−0.01	0.72*	0.55	**0.93***	0.17
② CTTCU/ATTCU	0.70*	−0.12	0.64*	0.40	0.87*	−0.07	0.47	0.65*
③ MLSTCU	0.60*	0.34	0.67*	−0.11	0.05	0.96*	0.70*	−0.52
④ STCU/TTCU	0.71*	−0.12	0.58*	0.37	**0.95***	−0.23	0.48	0.73*
Canonical *R*	0.75	0.22	0.82	0.42	0.74	0.30	0.64	0.33
Eigen value	1.27	0.05	2.05	0.22	1.21	0.10	0.68	0.13

Note: * Marks the largest absolute correlation between each variable and any discriminant function.

Table 10: Classification results for the CS and VR task with MLTTCU as the only predictor

Actual group	*N*	CS Predicted group membership (67.8%)			VR Predicted group membership (73.9%)		
		Low	High	Native	Low	High	Native
Low	37	**27** **73.0**%	9 24.3%	1 2.7%	**31** **83.8**%	6 16.2%	0 0.0%
High	47	9 19.1%	**31** **66.0**%	7 14.9%	8 17.0%	**31** **66.0**%	8 17.0%
Native	31	1 3.2%	10 32.3%	**20** **64.5**%	0 0.0%	8 25.8%	**23** **74.2**%

utilizing all the four measures as predictors that generated 78 (or 67.8 per cent) and 87 (or 75.7 per cent) cases with group membership correctly predicted respectively for the CS and VR task, such approximate classification accuracy suggested that the measure MLTTCU by itself may be chosen as likely the most effective indicator of spoken Chinese syntactic complexity.

MLTTCU and STCU/TTCU as predictors for writing tasks The correlations between predictor variables and the two functions in the two writing tasks showed somewhat different patterns from the correlations noted in the two speaking tasks. As shown in Table 9, for the FW task it was not the measure

MLTTCU but the predictor variable STCU/TTCU that showed the highest correlation, 0.95, followed by CTTCU/ATTCU demonstrating a correlation of 0.87, with Function 1. For the GR task, measure MLTTCU showed the highest correlation, 0.93 with Function 1. Function 2 by contrast, was best represented by predictor variable measure STCU/TTCU, with a correlation of $r =$ 0.73. This might be pointing to a possibility that the composition of dependent single TC-units played a more important role in contributing to the written Chinese syntactic complexity than to the spoken Chinese syntactic complexity. In other words, the FW task might have elicited more dependent single TC-units to compose each complex terminable TC-unit as well as a higher percentage of such complex terminable TC-units out of all terminable TC-units.

Applying the length measure MLTTCU and the ratio measure STCU/TTCU as predictors in combination rather than as single predicators through SPSS CLASSIFY, generated approximately the same accuracy for group membership prediction in the two writing tasks as the accuracy generated when all the four measures were applied. A total of 79 (or 68.1 per cent) (Wilks' lambda = 0.434, χ^2 (4, $N = 116$) = 93.933, $p = 0.000$) and 73 (or 62.9 per cent) (Wilks' lambda = 0.554, χ^2 (4, $N = 116$) = 66.354, $p = 0.000$) of cases were correctly classified for the FW and GR tasks respectively, as shown in Table 11. Comparing to utilizing all the four measures as predictors that generated 78 (or 67.2 per cent) and 71 (or 61.2 per cent) cases with group membership correctly predicted respectively for the FW and GR task, such approximate classification accuracy suggested that a combination of measure MLTTCU and STCU/TTCU may be chosen as likely the most effective indicator of written Chinese syntactic complexity.

The different strength of length and ratio measures at classifying spoken and written Chinese syntactic complexity may have several possible causes. On the one hand, the spoken and written Chinese complexity might have taken varying developmental trajectory in terms of composing versus lengthening of single TC-units. Spoken Chinese complexity may have shown more salient development alongside the lengthening of terminable TC-units by composing more and longer dependent single TC-units; while for written Chinese complexity, greater salient development was shown in the form of composing even more yet potentially shorter dependent single TC-units into a longer complex terminable TC-unit, as well as in composing a greater portion of complex terminable TC-units out of all terminable TC-units. As previously shown in Figure 3, the distribution of the mean score on each complexity measure by task also confirmed such a difference between spoken and written Chinese output. For the spoken Chinese output elicited, the mean scores on all four measures across proficiency groups were quite evenly spread. For the two writing tasks, their contrasting distributions confirmed a reliance on combining rather than lengthening single TC-units at the advanced to native proficiency level. As the FW task presented no restriction on form, its distributions showed some overlap on the length measure MLSTCU and large gaps on the two ratio measures CTTCU/ATTCU and STCU/TTCU between Group High and

Table 11: Classification results for the FW and GR task with MLTTCU and STCU/TTCU as the predictors

Actual group	*N*	FW Predicted group membership (68.1%)			GR Predicted group membership (62.9%)		
		Low	High	Native	Low	High	Native
Low	36	**22** **61.1**%	13 36.1%	1 2.8%	**26** **72.2**%	9 25.0%	1 2.8%
High	48	9 18.8%	**35** **72.9**%	4 8.3%	13 27.1%	**28** **58.3**%	7 14.6%
Native	32	0 0.0%	10 31.3%	**22** **68.8**%	0 0.0%	13 40.6%	**19** **59.4**%

Group Native. Such a low length increase but a high ratio increase could be attributed to more but shorter dependent single TC-units produced by Group Native. By contrast, as the RW task engaged participants in revising the existing single TC-units but restricted them from creating additional ones, its distributions showed no large gaps on the ratio measures.

On the other hand, the task design in this study may have functioned differentially in eliciting Chinese spoken and written syntactic complexity. The time allocated for the writing tasks was relatively limited. Unlike the spontaneity of speaking output, for writing output people had the chance to review what they were writing, weigh their words, and revise their structures. All of this required more contemplation. In addition, for the FW task, a daily life topic such as '*my relationship with my father/mother/brother/sister/friend (choose any one of them)*' was selected in order to minimize the floor effect. However, words and syntactic structures of lower complexity were able to treat such a daily life topic sufficiently, which might have kept the higher proficiency participants from using words and structures of higher complexity and thus wrote at a lower complexity level than their maximum proficiency would have allowed. Further analysis comparing and contrasting the spoken and written Chinese output across different proficiency levels will provide more insights.

DISCUSSION

The four proposed Chinese syntactic complexity measures were able to distinguish at high efficiency (61.2–75.7 per cent) the proficiency level of Chinese speakers performing across tasks. Measure MLTTCU proved to be the most effective indicator of spoken Chinese syntactic complexity. The combination of measure MLTTCU with STCU/TTCU proved to be the most effective indicator of written Chinese syntactic complexity.

Organic complexity measures

For assessment of a multidimensional construct like syntactic complexity, inaccurate conceptualization and metrics may cause construct underrepresentation and threaten the validity of studies. Chinese syntactic complexity analysis based on clause subordination contributed to the invalid observation that native speakers produced less complex Chinese language than that of L2 Chinese learners (Jin 2006; Yuan 2009). In terms of the conceptualization of complexity, the present study argued that a valid unit for syntactic complexity analysis should be dependent on the typological characteristics of the language to which it is applied. Therefore, this study proposed and used the TC-unit to address the typological features of Chinese languages, such as topic-prominence, covert connectives, and the interpretive variability of sentence boundary. In terms of a metric for complexity, the proposed inventory of TC-unit-based measures further diminished the risk of construct underrepresentation by targeting Chinese syntactic complexity at multiple levels (i.e. global, clausal, subclausal, and specific forms) along different dimensions (i.e. length, ratio, and frequency). The four proposed TC-unit-based measures comprehensively captured the syntactic complexity of the Chinese language via dimensions of length and inner structure at different levels. Generally, it was found that with higher Chinese language proficiency, longer terminable TC-units consisting of more dependent single TC-units were produced. By contrast, with lower language proficiency, shorter complex terminable TC-units consisting of less dependent single TC-units or shorter single terminable TC-units consisting of only one independent single TC-unit were produced. Overall, the TC-unit appears across tasks to be an appropriate unit of analysis for Chinese syntactic complexity.

Dynamic complexity development

The elicited complexity of language output may be affected by task complexity (Skehan 1998; Robinson 2001, 2005; Jackson and Suethanapornkul 2013). For spoken Chinese, mean length of terminable TC-unit (MLTTCU) proved to be the most effective indicator of syntactic complexity. By completing two speaking tasks of different task complexity—CS and VR in this study—the mean length of terminable TC-units produced ranged from 8.88–9.82 characters for lower-proficiency participants, 12.91–14.65 characters for higher-proficiency participants, and 17.83–19.74 characters for native-speaker participants. Compared with spoken Chinese, the written Chinese elicited in this study exhibited more dynamic interactions between global-level complexity and clausal-level complexity in Chinese proficiency development. In addition to lengthened terminable TC-units as a result of increases in proficiency level, written Chinese produced a greater proportion of complex terminable TC-units out of all terminable TC-units. This came at the expense of smaller increases in the length of single TC-units or even shorter single TC-units at native proficiency level. For written Chinese, syntactic complexity was captured better using the combination of length and ratio measures. In completing the two writing tasks of FW and GR in this study, lower proficiency participants produced 10.73–13.27 character-long terminable TC-units with a ratio of 1.08–1.09 single TC-units per terminable TC-unit. Higher proficiency participants produced 14.23–15.45 character-long terminable TC-units with a ratio of 1.22–1.31 single TC-units per terminable TC-unit. Native-speaker participants produced 19.12–19.94 character-long terminable TC-units with a ratio of 1.39–1.82 single TC-units per terminable TC-unit. Future studies that involve more qualitative analysis of how TC-units develop as proficiency increases will provide more insights on the developmental stages of Chinese syntactic complexity and the interaction between these stages.

The same complexity measures also demonstrated a fluctuation in the accuracy of group member classification for tasks of different designs and complexity. Therefore, both spoken and written Chinese output collected from tasks of varied designs conducted in various contexts should be able to better assess complexity performance and development. In addition to complexity, additional analysis of accuracy and fluency will warrant a comprehensive analysis of language performance and development.

Pedagogical implications

Chinese and English employ different supra-clausal level units in complexity composition: topic chain and subordination, respectively. Accordingly, L1 Chinese speakers demonstrated different complexity-composing strategies compared to English-speaking L2 Chinese speakers. To increase complexity, the English-speaking L2 Chinese speakers mainly relied on lengthening single TC-units; in contrast, the L1 Chinese speakers equally and may have preferably relied on combining more single TC-units into complexity terminable TC-units. However, the current L2 Chinese teaching practice (Liu *et al.* 2017) typically introduces the topic-comment structure merely as one of the unique Chinese sentence patterns, despite long acknowledging its indispensability. Conceptualizing and operationalizing Chinese complexity based on TC-unit rather than on clause subordination may fundamentally turn around such teaching practice, instead promoting the teaching of clause combining via coreferential zeros. This distinct difference between how the English and Chinese languages constitute complexity sheds light on L2 learning and teaching methods for effective L2 Chinese or L2 English syntactic complexity development.

CONCLUSIONS

This study has provided a potential solution for organically operationalizing the complexity construct of typologically different languages. A syntactic complexity measure should be dependent on the typological characteristics of the language to which it is applied. This study offers an alternative to the prior practice of subordination-based complexity operationalization for topic-prominent languages such as Mandarin Chinese. In addition to clause subordination and topic chain, the study also suggests that it is possible to constitute other complexity forms based on the typological features of a target language.

Adopting Bulté and Housen's (2012) conceptualization of complexity, and taking into consideration the particular features of the investigated language, the present study proposed and validated TC-units to measure Chinese syntactic complexity appropriately in terms of: the number and the nature of the single TC-units that a terminable TC-unit consists of; and the number and the nature of the constituent relationships between the single TC-units. Four TC-unit-based Chinese syntactic complexity measures were validated, based on a high efficiency (61.2–75.7 per cent) with which measure-based results distinguished the proficiency level of Chinese speakers.

To continue the effort to design organic measures of complexity, further investigation of the subsystems of complexity and the interaction among them is needed. The four validated measures target global- and clausal-level complexity; other measures of phrasal-level complexity and lexical diversity would reveal complexity at varied levels and the interaction between levels and kinds of complexity. In addition, for a construct as multifaceted as complexity, a comprehensive investigation cannot be achieved without more studies employing tasks of different designs conducted in various contexts. With a clearer picture of *what* constitutes Chinese syntactic complexity provided here, future research can now investigate *how* it is developed longitudinally in varied contexts. Future studies can adequately outline and efficiently facilitate Chinese complexity development in terms of the length and inner-structure of the TC-units based on the topic-prominent structure of Chinese.

ACKNOWLEDGEMENTS

I am deeply indebted to my mentors Dr. John M. Norris and Dr. Song Jiang for their inspirational advice, without which this work would not have been possible. My thanks go to Dr. Anna Mauranen for her belief in the value of this contribution. I would like to also acknowledge my colleagues (Dr. Nelson Brunsting, Dr. Jonathan Smart, Dr. T. H. M. Gellar-Goad, Dr. Andrew Rodekohr, and Dr. Joshua Ziesel) at Wake Forest University for their support and comments on my revisions.

Conflict of interest statement. None declared.

REFERENCES

Altman, D. G. 1991. *Practical Statistics for Medical Research*. Chapman and Hall.

Bulté, B. and **A. Housen**. 2012. 'Defining and operationalising L2 complexity' in A. Housen, F. Kuiken, and I. Vedder (eds): *Dimensions of L2 Performance and Proficiency: Investigating Complexity, Accuracy and Fluency in SLA*. John Benjamins, pp. 21–46.

Brown, J. D., **G. Robson**, and **G. Rosenkjar**. 2001. 'Personality, motivation, anxiety, strategies, and language proficiency of Japanese students' in Z. Dörnyei, and R. Schmidt (eds):

Motivation and Second Language Acquisition. University of Hawai'i Press, Second Language Teaching & Curriculum Center, pp. 361–98.

Chu, C. C. 1998. *A Discourse Grammar of Mandarin Chinese*. Peter Lang Publishing.

Chao, Y.-R. 1968. *A Grammar of Spoken Chinese*. University of California Press.

De Clercq, B. and **A. Housen**. 2017. 'A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity,' *The Modern Language Journal* 101/2:315–34.

Ellis, R. 2003. *Task-Based Language Learning and Teaching*. Oxford University Press.

Ellis, R. 2008. The Study of Second Language Acquisition, 2nd edn. Oxford University Press.

Ellis, R. and **G. Barkhuizen**. 2005. *Analysing Learner Language*. Oxford University Press.

Erlam, R. 2006. 'Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study,' *Applied Linguistics* 27/3: 464–91.

Foster, P., **A. Tonkyn**, and **G. Wigglesworth**.2000. 'Measuring spoken language: A unit for all reasons,' *Applied Linguistics* 21/3: 354–75.

Huang, C.-T. J. 1984. 'On the distribution and reference of empty pronouns,' *Linguistic Inquiry* 15: 531–44.

Hunt, K. W. 1965. *Grammatical Structures Written at Three Grade Levels*. National Council of Teachers of English.

Housen, A., **F. Kuiken**, and **I. Vedder**. 2012. *Dimensions of L2 Performance and Proficiency: Investigating Complexity, Accuracy and Fluency in SLA*. John Benjamins.

Ishikawa, S. 1995. 'Objective measurement of low-proficiency EFL narrative writing,' *Journal of Second Language Writing* 4: 51–70.

Jin, H. 2006. 'Cong Hanyu xiezuo guocheng kan CFL yuyan jiegou fuzadu de fazhan [The complexity development of Chinese as a foreign language: A writing perspective]' in X. Li (ed.): *Hanyu Jiaoxue Xuekan [Journal of Teaching Chinese]* Book 2. Peking University Press, pp. 114–40.

Jackson, D. and **S. Suethanapornkul**. 2013. 'The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity,' *Language Learning* 63/2: 330–67.

Jiang, W. 2011. *Pianzhang Yuyanxue Yanjiu. [Studies on Discourse Linguistics]*. Peking University Press.

Jiang, W. 2013. 'Measurements of development in L2 written production: The case of Chinese L2,' *Applied Linguistics* 34/1: 1–24.

LaPolla, R. J.1995. 'Pragmatic relations and word order in Chinese' in P. Downing and M. Noonan (eds): *Word Order in Discourse*. John Benjamins, pp. 297–329.

Lennon, P. 1990. 'Investigating fluency in EFL: A quantitative approach,' *Language Learning* 40/3: 387–417.

Li, W. 2005. *Topic-Comment Chains in Chinese: A Discourse Analysis and Application in Language Teaching*. Lincom Enropa.

Li, C. N. and **S. Thompson**. 1976. 'Subject and topic: A new typology of language' in C. N. Li (ed.): *Subject and Topic*. Academic Press, pp. 457–89.

Li, C. N., and **S. Thompson**. 1981. *Mandarin Chinese: A Function Reference Grammar*. University of California Press.

Liu, D. 2004. 'Identical topics: A more characteristic property of topic prominent languages,' *Journal of Chinese Linguistics* 32/1: 20–64.

Liu, Y., **T. Yao**, **N. Bi**, **L. Ge**, and **Y. Shi**. 2017. *Integrated Chinese*, Volume 1 Textbook, 4th edn. Cheng & Tsui Company, Inc.

Lian, S. 1993. *Contrastive Studies of English and Chinese*. Higher Education Press.

Lü, S. 1979. *Hanyu Yufa Fenxi Wenti. [Issues on the Analysis of Chinese Grammar]*. The Commercial Press.

Monroe, J. H. 1975. 'Measuring and enhancing syntactic fluency in French,' *The French Review* 48: 1023–31.

Michel, M. C .2017. 'Complexity, accuracy and fluency (CAF),' in S. Loewen and M. Sato (eds): *The Routledge Handbook of Instructed Second Language Acquisition*. Routledge, pp. 50-68.

Norris, J. M. 2015. 'Discriminant analysis' in L. Plonsky (ed.): *Advancing Quantitative Methods in Second Language Research*. Routledge, pp. 309–32.

Norris, J. M., and **L. Ortega**. 2009. 'Towards an organic approach to investigating CAF in instructed SLA: The case of complexity,' *Applied Linguistics* 30: 555–78.

Ortega, L., **N. Iwashita**, **J. M. Norris**, and **S. Rabie**. 2002. 'An investigation of elicited imitation tasks in crosslinguistic SLA research'. Paper presented at the Second Language Research Forum, Toronto, Canada, 4–6 October.

Robinson, P. 2001. 'Second language task complexity, the Cognition Hypothesis, language learning, and performance' in P. Robinson (ed.):

Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance. John Benjamins, pp. 3–37.

Robinson, P. 2005. 'Cognitive complexity and task sequencing: Studies in a componential framework for second language task design,' *International Review of Applied Linguistics in Language Teaching (IRAL)* 45/3: 161–77.

Skehan, P. 1998. *A Cognitive Approach to Language Learning*. Oxford University Press.

Tsao, F. 1979. *A Functional Study of Topic in Chinese: The First Step towards Discourse Analysis*. Student Book.

Tsao, F. 1990. *Sentence and Clause Structure in Chinese: A Functional Perspective*. Student Book.

Tai, J. H.-Y. 1985. 'Temporal Sequence and Word Order in Chinese' in J. Haiman (ed.): *Iconicity in Syntax*. John Benjamins Publishing Company, pp. 49–72.

Wolfe-Quintero, K., **S. Inagaki**, and **H.-Y. Kim**. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity*. University of Hawaii at Manoa, Second Language Teaching and Curriculum Center.

Wu, S.-L. and **L. Ortega**. 2013. 'Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese,' *Foreign Language Annals* 46/4: 680–704.

Xu, L. and **D. Liu**.1998. *Huati de Jiegou he Gongneng [The Structure and Function of Topic]*. Shanghai Jiaoyu Chubanshe [Shanghai Education Press].

Xu, T. 1991. 'Yuyi yufa chuyi [On sememic syntax],' *Language Teaching and Linguistic Studies* 3: 38–62.

Yuan, F. 2009. 'Measuring learner language in L2 Chinese in fluency, accuracy and complexity,' *Journal of the Chinese Language Teacher Association* 44/3: 109–30.

Zhou, Y. 2012. 'Willingness to communicate in learning Mandarin as a foreign and heritage language,' Doctoral dissertation, ProQuest (Identifier: ISBN9781267500793).

Zhou, Y. and **S.-L. Wu**. 2009. *Can Elicited Imitation Be Used for the Measurement of Oral Proficiency in L2 Chinese? a Pilot Study*. Unpublished manuscript, Department of East Asian Languages and Literatures, University of Hawaii at Manoa, Honolulu.