Measure schematicity through information content

A quantitative approach to grammaticalization

Liulin Zhang and Jiajun Tao
Soochow University

Apropos of the level of specificity, schematicity is the key indicator of grammaticalization in linguistics; compared to lexical items, the information provided by grammar patterns tends to be more abstract. With recourse to the notion of the quantifiable information content in information theory, the schematicity of a schema can be quantified by comparing the information content provided by the elements occurring in the open slots to that provided by the schema itself. A formula is thereby proposed to measure schematicity. This schematicity measure is able to illustrate the gradience and gradualness of grammaticalization in its applications in synchronic English data and diachronic Chinese data. Compared to previous measures of grammaticalization, there is a notable improvement in reliability and applicability.

Keywords: schematicity, information content, grammaticalization, quantitative

1. Introduction

Grammaticalization refers to the change whereby lexical items or constructions come to serve grammatical functions, or grammatical items develop new grammatical functions (Hopper & Traugott 2003: 18). A cline is proposed to illustrate the common path of grammaticalization; that is to say, content item > grammatical word > clitic > inflectional affix (Hopper & Traugott 2003: 7), entailing that each item to the right of the cline is more clearly grammatical and less lexical than the item to the left of it.

Despite the large bulk of literature dedicated to the specific cases of grammaticalization, few attempts have been made to quantify it. The most relevant work is Correia Saavedra (2021), who employed five variables, i.e. token frequency, let-

https://doi.org/10.1075/lali.00189.zha | Published online: 3 February 2025
Language and Linguistics 26:2 (2025), pp. 323–354. ISSN 1606-822X | E-ISSN 2309-5067

ter count, collocate density, colligate density, and dispersion, to predict whether a word is a grammatical item as opposed to a lexical one. Binary logistic regression models generated quantitative results ranging from 0 to 1, with 0 representing purely lexical, and 1 for grammatical. In the synchronic sample of 528 target words in written English, 73 words received the same score of 1, including articles (e.g., *a, the*), conjunctions (e.g. *and, albeit, 'cuz*), prepositions (e.g. *in, with, of*), pronouns (e.g. *I, it, ya*), etc. It needs to be noted that the model trained on the basis of the diachronic data turned out to be different from the one used for the synchronic data (Correia Saavedra 2021: 171), and the results were not really able to highlight changes in grammaticalization over time (Correia Saavedra 2021: 174), suggesting that the effects of those five variables are in fact case-specific: a general model is yet to be found. More importantly, although the author claims that the proposed approach is applicable to all kinds of linguistic elements (Correia Saavedra 2021: 182), collocate diversity and colligate diversity are clearly word/phrase-based measures as both pertain to the neighboring words: it is meaningless to talk about the collocate/colligate diversity of affixes or syntactic constructions. Moreover, as many grammatical words received the highest score of 1, this approach does not seem to be able to differentiate highly grammaticalized items, let alone affixes that are conventionally believed to be more grammaticalized than grammatical words.

Furthermore, the models proposed to measure productivity have also been used to estimate the level of grammaticalization (e.g. Arcodia & Basciano 2012; Perek 2018). Baayen's hapax-based P index (Baayen 1989, 1992; Baayen & Lieber 1991) is among the most commonly used, which is calculated by dividing the number of hapaxes of the process in question (n_1) by the token frequency of that process (N). Despite the ease of calculation, limitations of the P index have been thoroughly discussed. In the first place, Baayen "seems to assume perfectly prepared corpora" (Lüdeling et al. 2000), despite the fact that corpora typically contain a fair number of errors that are not negligible. As errors oftentimes occur only once, they would be counted as *hapax legomena*, becoming a dramatic noise for the estimation. Secondly, the N figures are extremely sensitive to corpus size (Bauer 1983: 148), rendering it meaningless to compare results from corpora of different sizes. Related to this issue, this method is more reliable for derivational affixes than for lexical items as the former typically have higher frequencies than the latter. In the last place, as the result of a statistical-probabilistic model, the P value represents the prediction of productivity, instead of a mere indicator of past activities (Fernández-Domínguez 2013). For this reason, Baayen (2009) introduced the term "potential productivity", which is not directly related to grammaticalization. If a grammatical form becomes obsolete, only preserved in a few fixed

expressions, it is no longer productive, with no more *hapax legomena*, but it is still a grammatical form.

To sum up, there is currently no quantitative model that can cover the complete spectrum of grammaticalization. From a theoretical perspective, the current models ignored the essential variable of grammaticalization, schematicity.

Cognitive constructionists maintain that the basic linguistic units are constructions, which are defined as form-meaning pairs (Goldberg 1995: 4). The term "syntax-lexicon" continuum is introduced to explain the relationship between words and syntactic patterns: syntactic patterns, are form-meaning pairs, but at a more abstract (schematic) level than words (Langacker 1987: 37, 2008: 19; Tuggy 2007). There is no unitary "grammar" of language but rather a continuum of constructions ranging from low frequency, highly specific, and lexical to high frequency, highly abstract, and general (Bybee 2008). According to Langacker (1987: 132–135; 2008: 19), a schema is abstract relative to its elaborations in the sense of providing less information and being compatible with a wider range of options, and thus the notion of schematicity pertains to level of specificity, i.e., the fineness of detail with which something is characterized. As illustrated in (1a) and (1b), → can be read as 'is schematic for'; and ← as 'is an elaboration of'. The categories/constructions to the right of → provide more details, than those categories/constructions to the left of it.

(1) a. word → content word → noun → *_ment* → *treatment*
b. transitive construction → ditransitive construction → *give* + recipient + patient → *give it a* + patient → *give it a try*

The difference between (1a) and (1b) resides in complexity: items listed in (1b) are syntagmatically more complex than those in (1a). Therefore, symbolic structures sketch along two parameters, namely complexity and schematicity, as represented in Figure 1, wherein the dashed line indicates the absence of any sharp boundary.

Examples of symbolic structures falling in different positions of Figure 1 are presented below in Table 1, in which *atomic* stands for low complexity and *substantive* for low schematicity.

In the view of the syntax-lexicon continuum, grammaticalization essentially pertains to an increase in schematicity. Therefore, the key to measuring the level of grammaticalization precisely resides in the quantification of schematicity, which has never been done before, and thus motivates the present study. The present study resorts to information theory to crystalize the notion of schematicity, and thereby proposes a viable quantitative model to measure schematicity based on the pre-existing information content formula. The relevance of information content to schematicity is discussed at length in § 2, leading to our quantita-

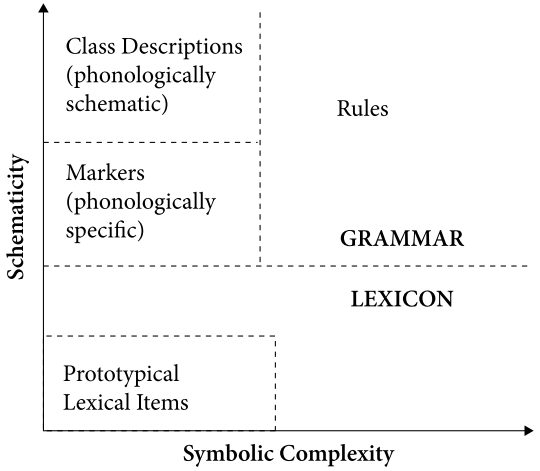

Figure 1. The syntax-lexicon continuum (Langacker 2008: 21)

Table 1. The syntax-lexicon continuum (Croft & Cruse 2004: 255)

Construction type	Traditional name	Examples
Complex and (mostly) schematic	syntax	[SBJ *be*-TNS V-*en by* OBL]
Complex, substantive verb	subcategorization frame	[SBJ *consume* OBJ]
Complex and (mostly) substantive	idiom	[*kick*-TNS *the bucket*]
Complex but bound	morphology	[N-*s*], [V-TNS]
Atomic and schematic	syntactic category	[DEM], [ADJ]
Atomic and substantive	word/lexicon	[*this*], [*green*]

tive schematicity measure. Synchronic applications of the schematicity measure in English data are presented in § 3, while diachronic applications of the schematicity measure in Chinese data across two millennia are presented in § 4. § 5 compares our schematicity measure to previous models including Correia Saavedra's (2021) multivariate models and Baayen's (1989; 1992) hapax-based measures, ending with a discussion of implications and future directions.

2. Quantification of schematicity

2.1 Schematicity in information theory

A key concern of information theory deals with the *informational value* of a communicated message, which is related to how surprising this message is, *aliās dictus*

the *information content*. If the message is about something that is very likely to happen, it carries very little information. In contrast, it will be much more informative if a highly unlikely event occurs. Therefore, the information content of an event X, represented as h(X), is related to the probability of its occurrence, p(X). Another feature of information content is that if X and Y are two independent events, the information content of these two events to co-occur (which is less likely to happen) equals to the sum of their individual information contents:

(2) $h(X, Y) = h(X) + h(Y)$

In (2), h(X, Y) is related to p(X, Y), as previously explained. Meanwhile, since X and Y are independent, $p(X, Y) = p(X) \cdot p(Y)$. A resemblance with the logarithmic function can be observed:

(3) $\log_a(X \cdot Y) = \log_a X + \log_a Y$

Based on the above observation, a logarithmic relationship can be conjectured between information content and probability, which explains Shannon's measure of information content (Shannon 1948a; 1948b, see also MacKay 2003: 32):

(4) $h(X) = \log_2 (\frac{1}{p(X)}) = -\log_2(p(X))$

In (4), 2 is chosen as the logarithmic base so that the resulting units may be called binary digits, i.e. bits. If base 10 is used, the units may be called decimal digits (Shannon 1948a; 1948b). For reasons to be discussed in § 3, the remainder of this paper will use 10 as the base.

The notion of information content can be easily applied to schematicity, as schematicity pertains to the specificity of information by definition (Langacker 1987: 132–135, 2008: 19; Tuggy 2007). Taking *_ment* and *treatment* as examples, as previously mentioned in § 1, *_ment* is schematic for *treatment*, while *treatment* is an elaboration of *_ment*. The word *treatment* provides more specific information than the schema *_ment*, and information theory tells us that the information content of *treatment* equals to the sum of the information contents of *_ment* and the morpheme *treat*, as shown in the following equation:

(5) $h(\textit{treatment}) = h(\textit{\_ment}) + h(\textit{treat})$

In (5), h(*treatment*) is related to the probability of this word to occur in the corpus, p(*treatment*), which equals $p(\textit{\_ment}) \cdot p(\textit{treat})$. Plugging in the formula of information content, logarithm is able to connect $p(\textit{\_ment}) \cdot p(\textit{treat})$ with $p(\textit{\_ment}) + p(\textit{treat})$, as shown below in (6).

(6) $\lg(p(\textit{treatment})) = \lg(p(\textit{\_ment})) + \lg(p(\textit{treat}))$

With this knowledge, the schematicity continuum presented in (1a) can be interpreted from the perspective of information theory, as shown below in (7).

(7) h(word) < h(content word) < h(noun) < h(*_ment*) < h(*treatment*)

It is clearly shown in (7) that the more schematic categories/constructions provide less information content for a specific case: for the word *treatment*, the category of noun provides less information content than the schema *_ment*, but meanwhile provides more information content than the category of word. Therefore, to estimate schematicity, a possible way is to look at how little information content is provided by the target schema for a typical instantiation of this schema.

2.2 Measurement of morphological grammaticalization

To illustrate Hopper & Traugott's grammaticalization cline, i.e. content item > grammatical word > clitic > inflectional affix, the focus of the present study is set on morphemes. Based on whether a morpheme can stand alone as a word, morphemes can be classified as free morphemes and bound morphemes. It needs to be noted that the distinction between free morphemes and bound morphemes is not clear-cut, and the boundedness of morphemes resides precisely in the level of grammaticalization: morphological grammaticalization involves phonetic erosion and semantic bleaching (Sweetser 1988; Heine 1993: 89, 106; Coussé et al. 2018), whereby the morpheme becomes less informative and more dependent on other elements (Lehmann 2002: 110; 2015: 132). For this reason, the present study sets morpheme as the target unit of analysis: bound morphemes are analyzed together with free morphemes, aiming to illustrate the complete spectrum of grammaticalization.

More and more researchers are coming to the conclusion that grammaticalization always takes place in specific contexts (Lehmann 2002; Traugott 2008; Hüning & Booij 2014): morphemes do not grammaticalize alone, but in schemata. For a target morpheme X, it grammaticalizes as the schema X_ or _X becomes more schematized, wherein the information content provided by the schema for its specific instantiations decreases. Without a schema, it is meaningless to talk about schematicity. With Shannon's measure of information content, presented in (4), we can easily compare the information content provided by the schema to that provided by the element occurring in the open slot. For example, X*a* is an instantiation of the schema X_; the frequencies of X and *a* are F(X) and F(*a*) respectively, and thus the probability of their occurrence can be represented as F(X)/T and F(*a*)/T, T corresponding to the corpus size (total token frequency). As previously mentioned in § 2.1, the schematicity of a schema pertains to how little information content is provided by this schema for a typical instantiation. To

get this, we can simply compare the information content of X_ to that of *a* by subtraction:

$$(8)\quad \begin{aligned} h(a) - h(X\_) &= -\lg\left(\frac{F(a)}{T}\right) + \lg\left(\frac{F(X)}{T}\right) \\ &= \lg\left(\frac{F(X)}{T} \Big/ \frac{F(a)}{T}\right) \\ &= \lg\left(\frac{F(X)}{F(a)}\right) \end{aligned}$$

Different situations are listed below in Table 2.

Table 2. The relationship between frequency and information content

Frequencies F(X) and F(*a*)	Information content	Interpretation
F(X) > F(*a*)	h(*a*) − h(X_) > 0	X_ is less informative than *a*
F(X) = F(*a*)	h(*a*) − h(X_) = 0	X_ and *a* are equally informative
F(X) < F(*a*)	h(*a*) − h(X_) < 0	X _is more informative than *a*

As in a cloze test, given a context that is rarely seen, those high-frequency elements can be easily filled out. On the other hand, it is much more difficult to guess the specific details based on commonly-seen information.

However, a schema typically has more than one instantiation: besides X*a*, X_ is also schematic for X*b*, X*c*, X*d*, and so forth. Simply comparing h(X_) to h(*a*) does not suffice to estimate the schematicity of X_: all its instantiations (tokens) need to be taken into consideration to formulate a general picture. A convenient way is to look at the mean, and thereby we can get a general formula for the schematicity index for morpheme-based schemata, as shown below in (9).

$$(9)\quad \begin{aligned} S(X\_) &= \text{Avg\_}\{h(\text{element occurring in the open slot}) - h(X\_)\} \\ &= \text{Avg\_}\left\{\lg\left(\frac{\text{total token frequency of X\_}}{\text{token frequency of the element occurring in the open slot}}\right)\right\} \end{aligned}$$

If we are interested in the schematicity of the *_ment*, in corpus, the token frequency of *_ment* adds up to 16,934, in which 227 types of elements occur in the open slot. we can aggregate the frequencies of those elements, as shown below in Table 3.

In Table 3, the second column "token frequency of this type of *_ment*" refers to the token frequency of each type of the instantiations of *_ment*, i.e. *government/development/environment/…*, it needs to be differentiated from the third column "frequency of the element", which refers to the token frequency of *govern/develop/environ/…* in the corpus, whether used as words or word-forming elements. Some elements invariably occur in *_ment*, such as *environ, parlia*, and *mismanage*, so the token frequency of these types of *_ment* is the same as that of

Table 3. Elements occurring in *_ment*

Element occurring in *_ment*	Token frequency of this type of *_ment*	Token frequency of the element
govern	3,485	3,806
develop	844	1,787
environ	741	741
parlia	729	729
manage	638	817
move	558	1,468
equip	495	561
invest	440	451
treat	440	707
employ	309	673
pay	301	1,322
... (214 types, 7,952 tokens of *_ment* omitted)		
mismanage	1	1
refurbish	1	3

the element. In contrast, some elements often occur in contexts other than *_ment*, such as *develop, manage* and *move*, so the token frequency of the element is significantly higher than that of those types of *_ment*.

As previously discussed, schematicity corresponds to how little information content is provided by the target schema for its instantiations, i.e., h(element occurring in the open slot) – h(X_), in which h(element occurring in the open slot) is related with the probability of this element to occur in the corpus, i.e. F(element occurring in the open slot)/T. Therefore, only "token frequency of the element", listed in the third column of Table 3, is relevant. For example, the token frequency of *move* is much higher than that of *environ*: although the token frequency of *movement* is lower than that of *environment, move* is not providing so much information content for *movement* as *environ* does for *environment*, as shown below in (10):

(10) a. $h(move) - h(\_ment) = \lg(\frac{F(\_ment)}{F(move)}) = \lg(16934 \div 1468) \approx 1.0620$

b. $h(environ) - h(\_ment) = \lg(\frac{F(\_ment)}{F(environ)}) = \lg(16934 \div 741) \approx 1.3589$

In an extreme situation, some elements only occur once in the corpus, and that is in the schema *_ment*, e.g. *mismanagement* and *refurbishment*, these elements would naturally be highly informative for the corresponding instantiations of *_ment*, as shown below in (11):

(11) $h(mismanage) - h(\_ment) = \lg(\frac{F(\_ment)}{F(mismanage)}) = \lg(16934 \div 1) \approx 4.2288$

Noticeably, for different types of instantiations, the value of {h(element occurring in the open slot) — h(*_ment*)} varies dramatically. Nevertheless, *movement, environment, mismanagement*, etc., are all instantiations of the schema *_ment*. To estimate the schematicity of *_ment*, each token of its instantiations needs to be taken into account. It must be noted that the information in Table 3 is listed by types, not tokens — there are 3,485 tokens of *government*, 844 tokens of *development*, etc. — but the schematicity index actually calls for the mean of {h(element occurring in the open slot) — h(X_)} for each token. Therefore, to calculate S(*_ment*) based on the information in Table 3, the result from each row needs to be weighted according to token frequency (information listed in the second column), as shown below in (12):

(12) $$\begin{aligned} &S(\_ment) \\ &= \text{Avg\_}\{\lg\ (\frac{\text{total token frequency of \_ment}}{\text{token frequency of the element occurring in the open slot}})\} \\ &= \lg(\frac{F(\_ment)}{F(govern)}) \cdot \frac{F(goverment)}{F(\_ment)} + \lg(\frac{F(\_ment)}{F(develop)}) \cdot \frac{F(development)}{F(\_ment)} + \dots \\ &= \lg(16934 \div 3806) \times 3485 \div 16934 + \dots \\ &\approx 1.3671 \end{aligned}$$

The value of the schematicity index can be above or below zero. As illustrated in Table 2, S(X_) = 0 (i.e. h(*a*) equals h(X_) in a typical instantiation of this schema) indicates that the target schema is typically as informative as the elements occurring in the open slot. In other words, the meaning of X_ is no more abstract than the elements occurring in it, and the schema X_ can thus hardly be called a grammatical (either word-forming or syntactic) pattern. If S(X_) < 0 (i.e. h(*a*) is smaller than h(X_) in a typical instantiation of this schema), the schema X_ typically provides more information content than the elements occurring in it, so X is better conceived of as a contentful element occurring in other grammatical patterns, and *_a* is likely to be more grammaticalized than X_. Only when S(X_) > 0 can X_ be considered a grammatical pattern in which the target schema is not so informative as the elements occurring in the open slot in a typical instantiation, which means that the target schema X_ typically relies on the elements occurring in the open slot to express specific meanings. The higher the S value, the more grammaticalized the schema.

As previously mentioned, morphemes do not grammaticalize alone. Grammaticalization always happens in context, entailing that for a target morpheme X, the schemata X_ and _X may have different levels of schematicity, which is exactly the case to be discussed in the next section. From the linguistic perspective, syntax is hierarchical and recursive, which means all kinds of elements, ranging from morphemes to clauses, may occur in the open slots of syntactic schemata. However, to minimize manual intervention, the computations in § 3 and § 4 only recognize the neighboring morphemes/words as the elements occurring in the open slots of the schemata X_ and _X, without considering the hierarchical structure of sentences.

3. Synchronic application of the schematicity measure

This section applies the schematicity measure in the Baby Edition of the British National Corpus (available at ⟨http://www.natcorp.ox.ac.uk⟩), consisting of four million words. The morphemes of interest (see below in (13)) are manually segmented, e.g. *endowment* is manually segmented as *endow -ment*.

(13) Target morphemes (besides free morphemes that are automatically segmented): *(-)able, auto(-), (-)berry, (-)dom, -ed* (past tense), *-ed* (past participle), *-ful, (-)graph, ing, inter(-), -ion, -ism, -logy, macro(-), -ment, micro(-), -ness, -ous, (-)over(-), -s* (plural), *-s* (3rd person singular), *(-)ship, tele(-)*.

These morphemes are representative prefixes, suffixes, and inflectional affixes frequently discussed in previous works (Bauer 2001: 148–149; Fernández-Domínguez 2013) that can be clearly identified and segmented. Some morphemes can be used alone as words, or combine with other elements to form words; e.g. *micro*(-), *auto*(-), (-)*berry*, (-)*over*(-). As long as it is the same morpheme, whether used alone or not, it is treated as one target. The frequency of each type of X_ and _X is generated by the *n*-gram function in AntConc 4.2.0: instantiations of X_ and _X can be perceived as bigrams when target morphemes are segmented.

The schematicity indices of X_ and X_, with the X being each of the target morphemes, are listed below in Table 4. The results of some high-frequency free morphemes are also listed as references.

Data in Table 4 can be plotted in Figure 2.

Table 4. Schematicity indices of X_ and _X

Morpheme	S(X_)	S(_X)	Morpheme	S(X_)	S(_X)
(-)able	0.152	1.577	*-ism*	−1.312	1.065
and	1.585	1.536	*keep*	−0.868	−1.233
at	0.303	0.527	*-logy*	−1.170	0.796
auto(-)	0.505	−1.615	*macro(-)*	0.452	−1.015
(-)berry	−1.000	−0.699	*may*	−0.099	−0.394
best	−0.330	−1.776	*-ment*	−0.085	1.367
book	−1.462	−0.958	*micro(-)*	0.231	−1.416
but	0.300	0.840	*must*	−0.612	−0.808
can	0.300	0.840	*never*	−0.001	−0.802
(-)dom	−1.196	0.163	*-ness*	−0.506	0.923
door	−0.908	−1.079	*of*	1.438	1.477
-ed (pp)	0.914	2.396	*on*	0.406	0.723
-ed (pt)	0.519	1.867	*-ous*	0.523	1.745
even	−0.249	−0.448	*(-)over(-)*	−0.006	0.279
face	−0.846	−0.851	*-s (3sg)*	0.399	1.320
fact	−1.332	−1.746	*-s (plural)*	1.277	2.513
from	0.380	0.435	*say*	−0.414	−0.390
-ful	0.016	0.702	*school*	−0.664	−0.292
go	−0.238	−0.119	*(-)ship*	−0.958	0.499
(-)graph	−0.170	0.480	*tele(-)*	0.261	−1.474
have	0.545	0.174	*that*	0.736	0.779
he	0.796	0.821	*the*	2.596	1.153
help	−0.813	−0.939	*this*	0.705	0.225
in	1.023	1.130	*to*	1.483	1.176
-ing	1.007	1.888	*want*	−0.659	−0.562
inter(-)	0.826	−0.565	*with*	0.656	0.694
-ion	0.333	2.239			

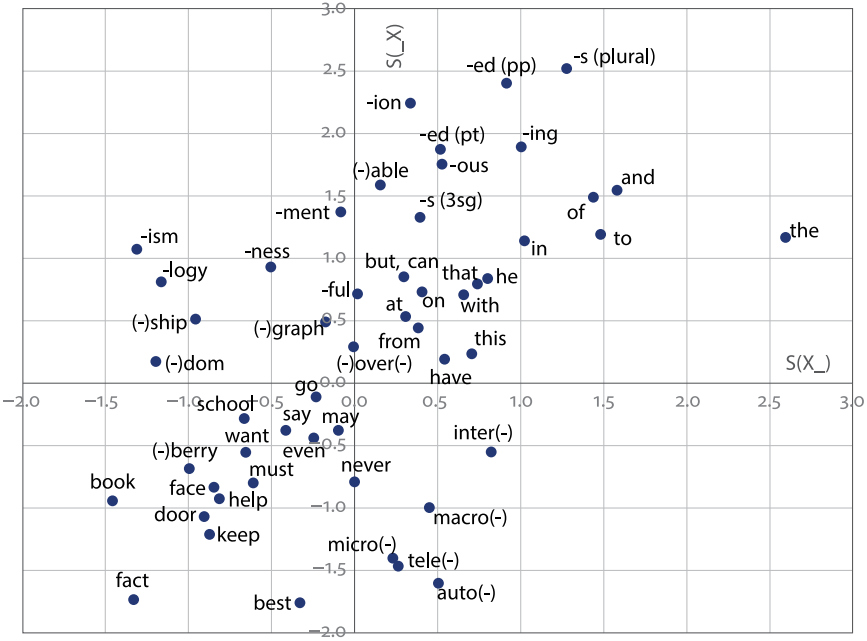

Figure 2. Schematicity indices of X_ and _X

Despite an overall positive correlation, an asymmetry between S(X_) and S(_X) can be observed from many cases. To crystalize the symmetry and asymmetry, we can divide the target morphemes into five categories (i.e. content word, function word, prefix, suffix, and inflectional affix)[1] and analyze the schematicity indices for each category. Results are shown below in Table 5 and Figure 3.

Table 5. Mean schematicity index of each category

	Content word	Function word	Prefix	Suffix	Inflectional affix
S(X_)	−0.529	0.731	0.455	−0.336	0.777
S(_X)	−0.679	0.662	−1.217	0.980	2.024

1. Following Correia Saavedra's (2021: 88) treatment, nouns, verbs, adjectives, and adverbs are categorized as content words, while function words include prepositions, articles, determiners, conjunctions, pronouns, and negations. Inflectional affixes include *-ed* (past tense), *-ed* (past participle), *-ing*, *-s* (plural), *-s* (3rd person singular). *Auto(-), inter(-), macro(-), micro(-)*, and *tele(-)* are categorized as prefixes, while *(-)able, (-)dom, -ful, (-)graph, -ion, -ism, -logy, -ment, -ness, -ous*, and *(-)ship* are treated as suffixes. However, it should be borne in mind that the distinctions between these categories are not clear-cut, which is to be discussed at length.

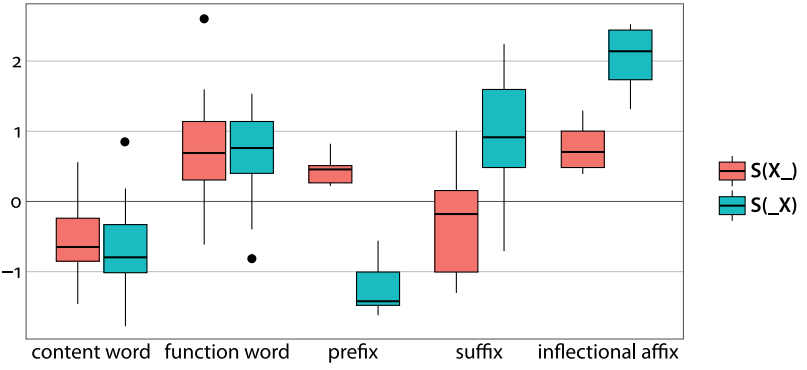

Figure 3. Schematicity indices of each category

The asymmetry between S(X_) and S(_X) corresponds to the unbalanced dependence of the target morpheme on the elements before and after it. If S(X_) is significantly higher than S(_X) — the element following X is typically providing more information than the element preceding X — X relies more on the element after it to express specific information, which characterizes prefixes, adjectives, and adverbs. In contrast, if S(X_) is significantly lower than S(_X), it is the elements preceding X that provide more information, which characterizes suffixes. Accordingly, morphemes that depend equally on both sides bring comparable S(X_) and S(_X) values. Function words made up of free morphemes are semantically abstract, relying heavily on the elements on both sides to provide contentful information, making S(X_) and S(_X) both high. Contentful nouns and verbs with specific meanings also have close values of S(X_) and S(_X), both low, oftentimes below zero, indicating that they are always providing contentful information.

In the meantime, the continua of the schematicity indices demonstrate the gradience of grammaticalization. Although clitics are not included in the present analysis, the greatest part of Hopper & Traugott's (2003) grammaticalization cline (i.e. content item > grammatical word > inflectional affix) is clearly demonstrated in Figures 2 and 3. Simply looking at S(_X) (the vertical axis in Figure 2), *book* < (-)*berry* < (-)*ship* < *-ment* < *-s* (plural) shows incremental semantic abstractness. This finding supports the hypothesis that affixes are grammaticalized from content words, and there are no clear cut boundaries in between: the boundedness of morphemes is a continuum. A morpheme can be completely bounded, such as *-s*, *-ed*, *-ing*, *-ment*, *-ion*, *-ous*, with high schematicity indices, or somewhere between free and bounded. Many of our target morphemes can stand alone as words or combine with other elements, such as (-)*able*, (-)*ship*, (-)*graph*, (-)*dom*, (-)*berry*, and their varied schematicity indices correspond to the varied levels of bounded-

ness. For example, S(_*berry*) is lower than S(_*ship*), reflecting that (-)*berry* typically provides more information content for the preceding element than (-)*ship*, and thus more free. The morpheme *-ship* is sometimes taken as an affix as in *friendship, relationship* etc., while *strawberry* is normally treated as a compound, in which (-)*berry* is virtually never perceived as an affix. Affixes are also characterized by different levels of grammaticalization: both as derivational affixes, *ship* is not so schematic and semantically bleached as *-ment*, yet *-ment* is still not so grammaticalized as inflectional affixes.

The reason why 10 is chosen as the logarithmic base can also be explained by the data shown in Table 4: with 10 being the base, prototypical function words and derivational affixes have S(X_) or S(_X) above 0.5, and prototypical inflectional affixes feature S(_X) above 1.5. The correspondence between the schematicity index and levels of grammaticalization is listed below in Table 6.

Table 6. Correspondence between the schematicity indices and levels of grammaticalization

Schematicity index		Level of grammaticalization	Example
S(_X) < 0.5	S(X_) < 0.5	content word	*help, face, keep*
	S(X_) > 0.5	prefix	*inter-, auto-*
0.5 < S(_X) < 1.5	S(_X) > S(X_)	suffix	*-ism, -logy, -ness*
	S(_X) = S(X_)	function word	*with, in, of, that*
	S(_X) < S(X_)	prefix (if any)	–
S(_X) > 1.5	S(_X) <> S(X_)	inflectional affix	*-ing, -ed, -s*
	S(_X) = S(X_)	function word	*the, and, of, in*

Importantly, the correspondence presented in Table 5 is a general tendency: 0.5 and 1.5 should not be understood as cutoff points. The theory of grammaticalization entails that content words, function words, derivational affixes, and inflectional affixes are all radial categories without clear-cut boundaries. As presented in Table 4 and Figure 2, the schmaticity indices form a continuum demonstrating the gradience of grammaticalization.

To compare the present results to previous measures of grammaticalization, it is worth reiterating that no previous measure can cover the entire spectrum of grammaticalization: Correia Saavedra's (2021) multivariate measure is proposed to estimate how likely a lexical item is a grammatical word, while Baayen's (1989; 1992) hapax-based measures have mainly been used to estimate the productivity of affixes. Bearing this in mind, the schematicity indices of free morphemes are

compared to Correia Saavedra's (2021) results, and those of bound morphemes are compared to the hapax-based P indices.

Correia Saavedra (2021) calculated the grammaticalization indices for 528 high-frequency words in the written portion and the spoken portion of the British National Corpus. Since the corpus size of the written portion is larger than the spoken portion, which is also the situation of the present study, we shall use his results based on written English for comparison, as shown below in Table 7.

Table 7. Correia Saavedra's (2021: 101–108) results for the targets

Word	***face***	***fact***	***door***	***help***	***want***	***book***
CS's result	0.011	0.016	0.017	0.032	0.035	0.062
Word	***keep***	***best***	***school***	***never***	***say***	***go***
CS's result	0.07	0.076	0.076	0.086	0.087	0.108
Word	***must***	***may***	***may***	***can***	***and***	***at***
CS's result	0.63	0.847	0.847	0.976	1	1
Word	***but***	***even***	***from***	***have***	***he***	***in***
CS's result	1	1	1	1	1	1
Word	***of***	***on***	***that***	***the***	***this***	***to***
CS's result	1	1	1	1	1	1

The correlation is 0.7860 between Correia Saavedra's results and the S(X_) indices, and 0.8322 between Correia Saavedra's results and the S(_X) indices, both high without a significant difference. Major variations come from function words, i.e., free morphemes that have relatively high levels of grammaticalization. Since Correia Saavedra's (2021) measure was proposed to estimate how likely a lexical item is a grammatical word, it is self-explanatory that many function words will receive the highest score of 1, with no further differentiation. In contrast, the present schematicity metric attempts to illustrate the entire spectrum of grammaticalization, from content words to inflectional affixes, so the nuanced difference between function words can also be captured. For example, as a function word, *the* is in fact more grammaticalized than *this* and *that*.

The hapax-based P indices are calculated for bound morphemes that occur at the end of words, as shown below in Table 8.

Since the boundedness of morphemes is a continuum, the sample listed in Table 7 includes absolute bound morphemes that never stand alone as words (listed in the first two rows of Table 7), and free-bound morphemes that do stand alone in some cases (listed in the last row of Table 7). When absolute bound

Table 8. Hapax-based P indices for the target bound morphemes

Target	***-ed (pp)***	***-ed (pt)***	***-ful***	***-ing***	***-ion***	***-ism***
P index	0.136	0.144	0.080	0.114	0.072	0.262
Target	***-logy***	***-ment***	***-ness***	***-ous***	***-s (3sg)***	***-s (plural)***
P index	0.152	0.054	0.197	0.148	0.134	0.095
Target	***(-)able***	***(-)berry***	***(-)dom***	***(-)graph***	***(-)over(-)***	***(-)ship***
P index	0.122	0.021	0.069	0.188	0.262	0.128

morphemes and free-bound morphemes are analyzed together, the correlation between the P indices and the S(_X) indices is −0.0159; when the comparison is limited to absolute bound morphemes, the correlation between the P indices and the S(_X) indices is −0.3357. No positive correlation can be observed, and the results of P indices deviate significantly from Hopper & Traugott's grammaticalization cline: the P indices of inflectional affixes turn out to be lower than many derivational affixes.

4. Diachronic application of the schematicity measure

This section applies the schematicity measure in the diachronic data of Chinese. The Chinese language has an uninterrupted history of documentation for over two millennia, bringing rich materials to the diachronic study of grammaticalization. Moreover, despite several attempts at orthographic reform, the writing system has always consisted of morphosyllabic Chinese characters, with each sign corresponding roughly to a morpheme pronounced as one syllable (Saalbach & Stern 2004; Hung 2012; also referred to as "logographic writing", see DeFrancis 1984: 72), so that the functional evolution of Chinese morphemes is easily traceable. By contrast, the definition of words has been controversial in Chinese (Dixon & Aikhenvald 2002). Packard (1998) points out that the notion of "word" did not exist in China until it was imported from the West in the 20th century. Experimental studies show that Chinese native speakers can only reach about 75% agreement in word segmentation, and have difficulties to replicate their own previous segmentation (e.g., Hoosain 1992; Sproat et al. 1996; Miller 2002; Bassetti 2005; Liu et al. 2013). Since Chinese texts are not word-segmented, compared to English, it is more challenging to distinguish bound morphemes from free morphemes in Chinese: the distinction between morphemes, words, and phrases is fluid (Hoosain 1992). This problem is further complicated by diachronic variations as morpheme compounds may have various levels of con-

ventionality in different historical strata (Norman 1988: 86). For example, *qīzǐ* 妻子 'wife' is recognized as a word in Modern Mandarin, but in Old Chinese it is typically analyzed as a phrase consisting of two free morphemes *qī* 妻 'wife' and *zǐ* 子 'son'. For this reason, the notion of "word" will be avoided in this section. Instead, morpheme, as represented by each character, is the basic unit of analysis.

The difficulty in word segmentation casts doubt on the applicability of Correia Saavedra's (2021) measure and Baayen's (1989; 1992) hapax-based measures, as both approaches presuppose the basic unit to be the word. Besides, both approaches turned out to be ineffective in highlighting changes in grammaticalization over time (Arcodia & Basciano 2012; Correia Saavedra 2021: 174). This section will thus not apply them to Chinese data.

Focusing on the grammaticalization of two morphemes, i.e., *zài* 在, *zhe* 着, we shall look at their use and schematicity indices in different historical periods. Corpora were constructed using materials from five historical periods, i.e. (i) *Zuǒ Zhuàn* (左傳, about 180,000 characters, late fourth century BC), (ii) *Shìshuō Xīnyǔ* (世說新語, about 79,000 characters, fifth century AD), (iii) six vernacular stories from *Biànwén* (變文, about 40,000 characters, c. 700–900 AD), (iv) four novels from *Sānyán Èrpāi* (三言二拍, about 70,000 characters, c. 1640 AD), and (v) five novels written by Shuò Wáng (王朔, about 490,000 characters, c. 1990 AD). There has been a considerable gap between literary Chinese and spoken Chinese for a long time (Norman 1988: 111). The selected materials all represent the vernacular style of their particular time. For the data of each of these historical periods, the method is generally the same as we have done for the English data presented in §3; the only difference is that Chinese data are inherently morpheme-segmented as written characters roughly corresponding to morphemes. It must be noted that *zài* and *zhe* are used to represent the pronunciations merely for convenience: in Modern Mandarin, *zài* is the standard pronunciation of the character 在, while the character 着 has various pronunciations (see §4.2 for details), among which *zhe* is the most common one; given the rich regional and historical varieties of Chinese (Norman 1988: 1), there is virtually no way to know the exact pronunciations of these characters at any given time and place.

Zài 在 and *zhe* 着 are two commonly recognized imperfective aspect markers in Modern Mandarin (Li & Thompson 1981: 185). Their diachronic S(X_) and S(_X) values are presented below in Table 9 and Figure 4.

To instantiate the above results, §4.1 will illustrate the diachronic usages of *zài* 在, while §4.2 focuses on the functional evolution of *zhe* 着.

Table 9. S(X_) and S(_X) of *zài* 在 and *zhe* 着

	Zuǒ Zhuàn	*Shìshuō Xīnyǔ*	*Biànwén*	*Sānyán Èrpāi*	Wáng's novels
S(在_)	0.000	0.666	0.410	0.688	0.774
S(_在)	−0.2000	0.340	0.202	0.350	0.788
S(着_)	–	0.030	−0.128	−0.040	0.795
S(_着)	–	−0.322	0.105	0.460	1.281

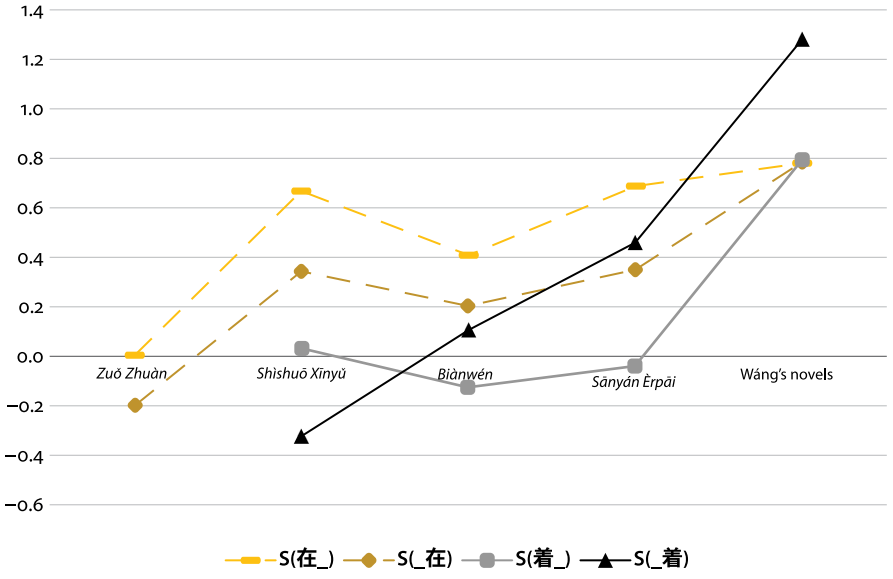

Figure 4. S(X_) and S(_X) of *zài* 在 and *zhe* 着

4.1 The grammaticalization of *zài* 在

Zài 在 functions mainly as a verb indicating 'to exist; to be at (location)' in *Zuǒ Zhuàn*, as exemplified below:

(14) 諸侯 之 賓，問 疾 者 多 **在**。
Zhūhóu zhī bīn, wèn jí zhě duō ***zài.***
marquises SUB guest ask sickness person many **ZAI**
'There were many marquises' guests who went to see doctors.'
(*Zuǒ Zhuàn*)

(15) 我 先 君 簡 公 **在** 楚。
*Wǒ xiān jūn Jiǎn Gōng **zài** chǔ.*
my former monarch Jian Gong **ZAI** Chu
'My former monarch Jian Gong is at Chu.' (*Zuǒ Zhuàn*)

(16) 子大叔 之 廟 **在** 道 南，其寢 **在** 道 北。
*Zǐ Tàishū zhī miào **zài** dào nán, qí qǐn **zài** dào běi.*
Zi Taishu SUB temple **ZAI** road south, his house **ZAI** road north
'Zi Taishu's temple is to the south of the road, and his house is to the north of the road.' (*Zuǒ Zhuàn*)

More abstract meanings are occasionally seen in *Zuǒ Zhuàn*: *zài* 在 in the following examples can be interpreted as 'to reside in; to be at the hands of':

(17) 師 克 **在** 和， 不 在 眾。
*Shī kè **zài** hé, bú **zài** zhòng.*
army victory **ZAI** harmony, not **ZAI** number-of-people
'The victory of an army resides in harmony, not in the number of people.' (*Zuǒ Zhuàn*)

(18) 二 子 之 不 欲 戰 也 宜， 政 **在** 季氏。
*Èr zǐ zhī bú yù zhàn yě yí, zhèng **zài** Jì Shì.*
two person SUB not willing fight also self-explained power **ZAI** Ji Shi
'It is self-explained that they do not want to fight as the power is at the hands of Ji Shi.' (*Zuǒ Zhuàn*)

Accordingly, both S(在_) and S(_在) values are close to 0, typical of verbs with relatively general and abstract meanings. After *Zuǒ Zhuàn*, the S(在_) and S(_在) values have been steadily increasing, corresponding to the rising prepositional use exemplified blow:

(19) 簡文 **在** 暗 室 中 坐，召 宣武。
*Jiǎnwén **zài** àn shì zhōng zuò, zhào Xuānwǔ.*
Jianwen **ZAI** dark room inside sit, call in Xuanwu
'Jianwen sat in the dark room, and called in Xuanwu.' (*Shìshuō Xīnyǔ*)

(20) 如來 生 **在** 南 天竺 國， 長 在 迦毗羅 城。
*Rúlái sheng **zài** nán Tiānzhú guó, zhǎng zài Jiāpíluó chéng.*
Buddha be-born **ZAI** south India country grow-up at Kapila city
'Buddha was born in India to the south of China, and grew up in the city of Kapila." (*Biànwén*)

(21) 只 聽 得 雞 **在** 籠 中 不 住 吱吱喳喳。
*Zhī tīng é jī **zài** long zhōng bú zhù zhīzhīzhāzhā.*
Only listen get chicken **ZAI** cage inside not stop squawk
'Only heard chickens squawking ceaselessly in the cage.'
(*Sānyán Èrpāi*)

The adverbial use of *zài* 在 appeared relatively late. In *Sānyán Èrpāi*, sporadic cases are captured where *zài* 在 compounds with *zhèng* 正 to indicate progressive, as follows:

(22) 且 說 朱 恩 同 母親 渾家 正 **在** 那裏 飼 蠶……
*Qiě shuō Zhū Ēn tóng mǔqīn húnjiā zhèng **zài** nàlǐ sì cán…*
Just say Zhu En with mother wife just **ZAI** there feed silkworm
'(Just talking about) Zhu En was there feeding silkworms with his mother and wife…' (*Sānyán Èrpāi*)

(23) 平 氏 正 **在** 打疊 衣箱，內 有 珍珠 衫 一 件。
*Píng Shì zhèng **zài** dǎdié yīxiāng, nèi yǒu zhēnzhū shān yī jiàn.*
Ping Shi just **ZAI** pack suitcase inside have pearl shirt one CL
'Ping Shi was just packing up his suitcase. There is a pearl shirt inside.'
(*Sānyán Èrpāi*)

Cases in which *zài* 在 functions as an imperfective aspect marker alone are mainly seen in Shuò Wáng's novels, representing Modern Mandarin, exemplified below:

(24) 雖 說 時代 **在** 變， 道德 還 是 古代 那 道德。
*Suī shuō shídài **zài** biàn, dàodé hái shì gǔdài nà dàodé.*
although say time **ZAI** change moral still is ancient-times that moral
'Although the time is changing, the moral is still the moral of ancient times.'
(Shuò Wáng's novel)

(25) 你 心 裏 總 有 個 小小 的 自我 **在** 作 怪。
*Nǐ xīn lǐ zǒng yǒu gè xiǎoxiǎo de zìwǒ **zài** zuò guài.*
your heart inside always have CL little SUB ego **ZAI** make trouble
'There is always a small ego making trouble in your heart.'
(Shuò Wáng's novel)

Noteworthily, even in the Modern Mandarin sample, the prepositional use is still the prototypical function of *zài* 在, with tokens far outnumbering its adverbial use.

The usages of *zài* 在 across different historical periods can be summarized below in Table 10.

It can be observed from Table 10 that the percentage of tokens wherein *zài* 在 is used as a verb has been steadily decreasing, while its prepositional use has been steadily increasing. Its adverbial use appeared relatively late. Although *zài*

Table 10. Diachronic usages of *zài* 在

	Zuǒ Zhuàn	*Shìshuō Xīnyǔ*	*Biànwén*	*Sānyán Èrpāi*	Wáng's novels
Verbal: to exist; to be in/on/at	97.44%	82.25%	82.88%	64.99%	37.20%
Prepositional: in/on/at	2.35%	17.75%	16.22%	34.20%	40.40%
Adverbial: imperfective aspect (constrained)	—	—	—	1.16%	1.80%
Adverbial: imperfective aspect (independent)	—	—	0.90%	0.29%	7.60%
Other	0.21%	—	—	0.87%	13.00%

在 can be used as an aspect marker in Modern Mandarin, this is in fact not its prototypical use. Besides, there are a few compounds with relatively high frequencies in Wáng's novels, such as *xiànzài* 現在 'now' and *shízài* 實在 'really', coded as "other" in the sample. The high conventionality of these compounds undoubtedly affects the schematicity scores. Importantly, although *zài* 在 has been grammaticalizing, its verbal use has always been there: it can still be used as a verb indicating 'to exist' in Modern Mandarin. Multiple layers of grammaticalization co-present for this morpheme.

4.2 The grammaticalization of *zhe* 着

The character *zhe* 着 appeared relatively late in Chinese history. It is not seen in *Zuǒ Zhuàn*, and not common in *Shìshuō Xīnyǔ* either. Its appearances in *Shìshuō Xīnyǔ* and *Biànwén* are invariably verbal, denoting 'to wear' or 'to touch; to be in contact with'. Based on the phonetic standard of Modern Mandarin, these usages of *zhe* 着 are pronounced as *zhuó* or *zháo*, as exemplified below:

(26) 太傅 時 年 七 八 歲， **着** 青 布 綺。
Tàifù shí nián qī bā suì, ***Zhuó*** *qīng bù kù.*
Taifu time age seven eight year-old **ZHUO** green cloth pants
'Taifu was seven or eight years' old at that time, wearing green-cloth pants.'
(*Shìshuō Xīnyǔ*)

(27) 奴家 愛 **着** 綺羅 裳。
Nújiā ài ***zhuó*** *qǐluó cháng.*
I (humble) love **ZHUO** silk skirt
'I love to wear silk skirts.' (*Biànwén*)

(28) 以 舌 着 上顎。
*Yǐ shé **zhuó** shàng'è.*
use tongue **ZHUO** palate
'Use your tongue to touch your palate.' (*Biànwén*)

As a verbal element, *zhe* 着 also functions as the resultative complement for other verbal elements, as in the following examples:

(29) 藍田 愛 念 文度。雖 長大， 猶 抱 着 膝 上。
*Lántián ài niàn Wéndù, suī zhǎngdà, yóu bào **zhuó** xī shàng.*
Lantian love think-of Wendu although grow-up still hold **ZHUO** knee top
'Lantian loves Wendu. Although Wendu has grown up, Lantian still holds him on his knees.' (*Shìshuō Xīnyǔ*)

(30) 忽然 逢 着 夜叉王。
*Hūrán féng **zháo** yèchāwáng.*
suddenly meet **ZHAO** Yaksa
'Suddenly ran into Yaksa.' (*Biànwén*)

It can be noticed that *zhe* 着 are transitive in all the above examples — the nominal elements after *zhe* 着 are invariably the objects of it — the relationship between *zhe* 着 and the elements following it is fairly close. This is a typical case in *Shìshuō Xīnyǔ*, which explains the relatively high value of S(着_), as compared with S(_着), albeit both below 0.1 as presented in Table 9 and Figure 4. However, starting from *Biànwén* in our sample, there began to be cases in which *zhe* 着 does not take objects, as shown below:

(31) 將士 夜 深 渾 睡 着。
*Jiàngshì yè shēn hún shuì **zháo.***
officer-soldier night late all sleep **ZHAO**
'Officers and soldiers have all fallen into sleep at night.' (*Biànwén*)

No physical contact is expressed by *zhe* 着 in (31). 'Contact' needs to be understood in a more abstract sense: to get into a state. As the meaning of *zhe* 着 became increasingly abstract, its association with the preceding verbal elements strengthened, corresponding to the rising S(着_) value, while the S(_着) value did not rise so fast as the association between *zhe* 着 and the following elements has been relatively loosened: *zhe* 着 no longer needs to take objects itself.

From the verbal complement use of *zhe* 着 exemplified in (29), (30) and (31), the meaning of it is further bleached. In many cases in *Sānyán Èrpāi*, *zhe* 着 indicates the continuation of a state or the progression of an activity, thus becoming an aspect marker, as shown in the following examples.

(32) 家 家 都 閉 着 門兒。
*Jiā jiā dōu bì **zhe** ménr.*
home home all close **ZHE** door
'The door of every home is closed.' (*Sānyán Èrpāi*)

(33) 三巧兒 指 着 床 前 一 個小 小 藤 榻兒，道……
*Sānqiǎor zhǐ **zhe** chuáng qián yí gè xiǎo xiǎo téng tàr, dào …*
Sanqiao point **ZHE** bed front one CL small small wicker couch say
'Sanqiao was pointing at a small wicker couth in front of the bed when she said …' (*Sānyán Èrpāi*)

This is the dominant use of *zhe* 着 in Modern Mandarin, making the S(_着) value above 1 (see Table 9 and Figure 4), while the verbal uses of it are only preserved in a few fixed expressions such as *zhuólù* 着陸 'to land' and *bùzhuóbiānjì* 不着邊際 'wide of the mark'. According to the phonetic standard of Modern Mandarin, *zhe* 着 needs to be pronounced as *zhe* as an aspect marker. The neutral tone and the reduced vowel are clearly evidence of phonetic erosion accompanying grammaticalization.

Overall, the diachronic usages of *zhe* 着 are summarized below in Table 11. Unlike *zài* 在, the verbal use of *zhe* 着 has been dramatically shrinking. In Modern Mandarin, it can barely function as a verb anymore. Predominantly aspectual, the function of *zhe* 着 is much more specialized than *zài* 在.

Table 11. Diachronic usages of *zhe* 着

	Zuǒ Zhuàn	***Shìshuō Xīnyǔ***	***Biànwén***	***Sānyán Èrpāi***	**Wáng's novels**
verbal *zhuó*: to wear	—	33.33%	15.39%	1.49%	0.40%
verbal *zhuó*/*zháo*: to touch; to be in contact with	—	26.67%	34.62%	4.98%	1.00%
verbal complement *zhuó*/*zháo*: to be in contact with	—	40.00%	38.46%	15.92%	3.00%
verbal complement *zháo*: to get into a state	—	—	11.54%	2.99%	1.40%
imperfective aspect marker *zhe*	—	—	—	72.64%	91.20%
Other	—	—	—	1.99%	3.00%

4.3 Summary

In §4 we have seen the correspondence between the schematicity indices and the grammaticalization of *zài* 在 and *zhe* 着: as *zài* 在 and *zhe* 着 evolved into aspect markers from contentful verbs, their schematicity indices have been increasing. Furthermore, the present schematicity measure can clearly reflect the gradualness of grammaticalization — for the target morphemes, it takes a long time for their grammatical uses to appear and to gain frequency — grammaticalization does not take place overnight (Traugott & Trousdale 2010). The diachronic schematicity indices of *zài* 在 and *zhe* 着 correspond well with the diachronic usages of these two morphemes.

However, the difference between *zài* 在 and *zhe* 着 should not be neglected. Although both originated from verbs, *zài* 在 derived adverbial uses preceding verbal elements; while *zhe* 着 developed into an aspect marker following verbal elements. Their difference can be easily observed from the following example:

(34) 大家 都 **在** 爭 **着** 向 馮 先生 獻媚， 你
*Dàjiā dōu **zài** zhēng **zhe** xiàng Féng Xiānsheng xiànmèi, nǐ*
everybody all **ZAI** strive **ZHE** towards Feng Mr. flatter you
為什麽 不 去？
wèishénme bú qù?
why not go
'Everybody is striving to flatter Mr. Feng. Why don't you go?'
(Shuò Wáng's novel)

Besides, it can also be observed that the overall S(X_) and S(_X) values of *zài* 在 and *zhe* 着 are lower than English inflectional affixes: the S(_*ing*) is 1.888. This difference can be attributed to the multiple layers of uses preserved in Chinese owing to the logographic writing system that is not sensitive to phonetic erosion. In English, and virtually all other languages using alphabets, if a morpheme is phonetically eroded in grammaticalization, the erosion is sometimes reflected in the writing; e.g., *going to* > *gonna*. Moreover, the complicated history of language contact makes it impossible to trace the origin of every English morpheme: *-ing* may also originate from content items, but there is virtually no way to know. In contrast, since Chinese characters are logographic by nature — they never serve as accurate records of phonemes — the written form of a morpheme always remains the same even if it is phonetically eroded with grammaticalization. Multiple layers of a morpheme can thereby co-present, which is exactly the case of *zài* 在. The co-presence of the contentful use and the functional use surely affects the estimate of schematicity, which also explains the reason why S(_着) is higher than S(在_) in Modern Mandarin: the verbal use of *zhe* 着 is generally obsolete, but *zài* 在 can still function as a verb.

5. Discussions

5.1 Comparison with previous measures

With recourse to the notion of information content in information theory, this paper argues that the schematicity of a construction can be understood as how little information content is provided by the target schema for a typical instantiation of this schema, and thus can be measured by comparing the information content provided by the elements occurring in the open slots to that provided by the schema itself. A formula is thereby proposed to quantify schematicity, and shown effective to characterize the grammaticalization of morphemes with synchronic English data and diachronic Chinese data.

Compared to previous measures of schematicity, the present schematicity measure shows incomparable applicability. From the synchronic perspective, the entire spectrum of grammaticalization from content items to affixes are clearly reflected by the schematicity indices, demonstrating the gradience of grammaticalization. From the diachronic perspective, the schematicity indices well correspond to the diachronic usages of target morphemes among which the grammatical uses gradually gain frequency, and thus illustrate the gradualness of grammaticalization. With the present approach, morphemes with various degrees of boundedness can be analyzed together, and in fact the boundedness of morphemes can be estimated by the S(X_) and S(_X) indices. This is beyond the ability of Correia Saavedra's (2021) multivariate measure and the hapax-based productivity measure. Besides, reliability is another notable advantage of the present schematicity measure. As presented in (9), the formula for the schematicity index is essentially based upon token-token ratio. Compared to previous methods drawn upon type-token ratio, including various types of productivity measures, this approach is not so sensitive to corpus size and sporadic errors in the corpus, making the results from different corpora roughly comparable.

To understand the improvement in applicability, it must be noted that the present approach taps directly on schematicity, which is the essential parameter underlying the syntax-lexicon continuum, while previous studies take token frequency, letter count, collocate density, colligate density, dispersion, and potential productivity as indicators of grammaticalization. Admittedly, these factors are related with schematicity. Simply looking at the formula for the schematicity index, presented in (9), "total token frequency of X_" occurs in the numerator position, while the "token frequency of the element occurring in the open slot" occurs in the denominator position, so we are definitely not negating the relationship between frequency and grammaticalization. For the schematicity of free morphemes, "the element occurring in the open slot" is exactly the "collocate"

in Correia Saavedra's (2021) measure: if many types of elements only occur in the target schema but nowhere else, the collocate diversity is high, and {h(element occurring in the open slot)-h(X_)} is also high. For bound morphemes, a large number of hapax legomena means that many types of low-frequency elements occur in the open slot, and thus is likely to bring a high value for {h(element occurring in the open slot)-h(X_)}, resulting in a high schematicity score. However, these are merely concomitant indicators. The essential factor of grammaticalization — schematicity — has been ignored. Schematicity is a feature of schemata with open slots, and it needs to be measured in relation to specific instantiations of the target schema. Therefore, even if we are investigating the grammaticalization of morphemes, we are actually looking at the schematicity of X_ and _X, but not the morpheme itself. In the meantime, "token frequency of the element occurring in the open slot", distinguished from "token frequency of this type of instantiation" (see Table 3 in § 2.2), needs to be taken into consideration, as the target schema is oftentimes not the only context in which the elements can occur. This view radically distinguishes the present study from Correia Saavedra's (2021) approach, and only in this way can we capture the difference between S(X_) and S(_X), which is the key to understanding the boundedness of morphemes.

5.2 Implications and future directions

It is worth reiterating that the computations presented in § 3 and § 4 only consider the morphemes/words adjacent to X as the elements occurring in the open slots of the schemata X_ and _X. This approach minimizes manual intervention, facilitates autonomic computing, and thus maximizes the replicability of the present study. However, it does not necessarily represent how the human brain processes language, and tends to underestimate the schematicity of elements with broad syntactic scopes, such as modal verbs, complementizers, and sentence final particles. In fact, experimental studies consistently show that brain systems track hierarchical syntax incrementally in addition to sequential processing (Ding et al. 2016; Henderson et al. 2016; Martin & Doumas 2017; Brennan & Hale 2019), so token-by-token identification of constructions undoubtedly produces more accurate results. With token-by-token identification, the schematicity index can be computed for all kinds of constructions. Taking the ditransitive construction as an example, unlike morpheme-based schemata, the ditransitive construction does not have any fixed constituents. Instead, there are three open slots: the verb, the indirect object, and the direct object. As long as the three elements appear together as shown below in Table 12, they always form an instantiation of this

construction. Therefore, there is no need to distinguish "token frequency of the elements" from "token frequency of this type of instantiation".

Table 12. Instantiations of the ditransitive construction

Type of instantiations	Token frequency of this type
give me five	a
give me a break	b
give you an example	c
teach me a lesson	d
show me the meaning	e
... (n types in total)	... (N tokens in total)

Nonetheless, schematicity still pertains to how little information content is provided by the schema itself, so we still need to compare the information content provided by the schema to that provided by the elements occurring in the three open slots, i.e. h(elements occurring in the open slots)—h(ditransitive construction). The only difference is that in this case, elements in the three open slots precisely form different types of instantiations, so h(elements occurring in the open slots) = h(instantiation of the schema). The original schematicity formula can thereby be further simplified for constructions with no fixed constituents, as shown below in (35).

(35) $$\begin{aligned} S(\text{construction}) &= \text{Avg\_}\{h(\text{instantiation}) - h(\text{construction})\} \\ &= \text{Avg\_}\left\{\lg\left(\frac{\text{total token frequency of the construction}}{\text{token frequency of each type of instantiation}}\right)\right\} \end{aligned}$$

Despite the distinct advantages, there are still a few points that must be pointed out. A big problem pertains to the determination of morphemes. As previously mentioned in §4, Chinese morphemes are determined by the written characters: even if a morpheme is phonetically eroded, such as *zhe* 着, it is always counted as the same morpheme as long as the same character is used. This is not a common way for languages using alphabets, e.g., *gonna* is typically taken as one morpheme, but *going to* obviously contains more than one morpheme. For this point, writing system seems to play a role in the determination of morphemes: languages using logographic writing systems have the semantic evolution of morphemes clearly documented, while the exact pronunciation is never accurately represented, and thus the determination of morphemes has to rely more on semantics instead of phonetics; in contrast, it is more challenging for languages using alphabets to keep track of the semantic evolution of elements, so the determination of morphemes relies more on phonetics. Additionally, homonyms, written variations,

and internal inflections add more layers to the difficulty of morpheme determination. Ultimately, the schematicity measure can only be used to describe, but not to predict the development of constructions. In §4, we have seen that both originated from verbs, *zài* 在 and *zhe* 着 went on different paths of grammaticalization: *zài* 在 becomes an adverbial preceding verbal elements, while *zhe* 着 becomes a particle following verbal elements. The schematicity indices of earlier historical periods provide no hints for this divergence. Besides, data in the present study are drawn solely from corpora; future studies can be conducted comparing the schematicity indices to human behavioral performances in schematicity judgment, lexical decision, self-paced reading, priming, etc. Jäger & Rosenbach (2008) hypothesized that less grammaticalized items tend to prime the more grammaticalized ones, but not the converse, which is compatible with our understanding that grammaticalized elements provide less information content than contentful elements. Empirical investigation is called for along this line.

Funding

This study is supported by the National Social Science Fund of China (Grant no. 20FYYB043), and the Interdisciplinary Research Team in Humanities and Social Sciences at Soochow University (Grant no. 5033720623).

Acknowledgements

The general idea has benefited from the discussions with Zhihui Yang, Huan Li, and Yihan Zhou, to whom we express our sincere gratitude.

List of abbreviations

ADJ	adjective
CL	classifier
DEM	demonstrative
N	Noun
OBJ	object
OBL	oblique
SBJ	subject
SUB	subordinative marker
TNS	tense
V	Verb
ZAI	morpheme *zài* 在
ZHE (ZHUO/ZHAO)	morpheme *zhe* 着 (*zhuó*/*zháo*)

References

Arcodia, Giorgio Francesco & Basciano, Bianca. 2012. On the productivity of the Chinese suffixes –兒 *–r*, –化 *–huà* and –頭 *–tou*. *Taiwan Journal of Linguistics* 10(2). 89–117.

Baayen, R. Harald. 1989. *A corpus-based study of morphological productivity: Statistical analysis and psychological interpretation*. Amsterdam: Vrije Universiteit Amsterdam. (Doctoral dissertation.)

Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In Booij, Geert & van Marle, Jaap (eds.), *Yearbook of morphology 1991*, 109–149. Dordrecht: Springer.

Baayen, R. Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Lüdeling, Anke & Kytö, Merja (eds.), *Corpus linguistics: An international handbook*, vol. 2, 899–919. Berlin: Mouton de Gruyter.

Baayen, R. Harald & Lieber, Rochelle. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* 29(5). 801–843.

Bassetti, Benedetta. 2005. Effects of writing systems on second language awareness: Word awareness in English learners of Chinese as a foreign language. In Cook, Vivian & Bassetti, Benedatta (eds.), *Second language writing systems*, 335–356. Clevedon: Multilingual Matters.

Bauer, Laurie. 1983. *English word-formation*. Cambridge: Cambridge University Press.

Bauer, Laurie. 2001. *Morphological productivity*. Cambridge: Cambridge University Press.

Brennan, Jonathan R. & Hale, John T. 2019. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS One* 14(1). 1–17. (Article e0207741.)

Bybee, Joan L. 2008. Usage-based grammar and second language acquisition. In Robinson, Peter & Ellis, Nick C. (eds.), *Handbook of cognitive linguistics and second language acquisition*, 216–236. New York: Routledge.

Correia Saavedra, David. 2021. *Measurements of grammaticalization: Developing a quantitative index for the study of grammatical change*. Berlin: De Gruyter Mouton.

Coussé, Evie & Andersson, Peter & Olofsson, Joel. 2018. Grammaticalization meets construction grammar: Opportunities, challenges and potential incompatibilities. In Coussé, Evie & Andersson, Peter & Olofsson, Joel (eds.), *Grammaticalization meets construction grammar*, 3–19. Amsterdam: John Benjamins.

Croft, William & Cruse, D. Alan. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press.

DeFrancis, John. 1984. *The Chinese language: Fact and fantasy*. Honolulu: University of Hawai'i Press.

Ding, Nai & Melloni, Lucia & Zhang, Hang & Tian, Xing & Poeppel, David. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience* 19(1). 158–164.

Dixon, R. M. W. & Aikhenvald, Alexandra Y. 2002. Word: A typological framework. In Dixon, R. M. W. & Aikhenvald, Alexandra Y. (eds.), *Word: A cross-linguistic typology*, 1–41. Cambridge: Cambridge University Press.

Fernández-Domínguez, Jesús. 2013. Morphological productivity measurement: Exploring qualitative versus quantitative approaches. *English Studies* 94(4). 422–447.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: The University of Chicago Press.

Heine, Bernd. 1993. *Auxiliaries: Cognitive forces and grammaticalization*. Oxford: Oxford University Press.

Henderson, John M. & Choi, Wonil & Lowder, Matthew W. & Ferreira, Fernanda. 2016. Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage* 132. 293–300.

Hoosain, Rumjahn. 1992. Psychological reality of the word in Chinese. In Chen, Hsuan-Chih & Tzeng, Ovid J. L. (eds.), *Language processing in Chinese* (Advances in Psychology 90), 111–130. Amsterdam: North-Holland.

Hopper, Paul J. & Traugott, Elizabeth Closs. 2003. *Grammaticalization*. 2nd edn. Cambridge: Cambridge University Press.

Hung, Yueh-Nu. 2012. How a morphosyllabic writing system works in Chinese. In Goodman, Ken & Wang, Shaomei & Iventosch, Mieko Shimizu & Goodman, Yetta (eds.), *Reading in Asian languages: Making sense of written texts in Chinese, Japanese, and Korean*, 16–31. New York: Routledge.

Hüning, Matthias & Booij, Geert. 2014. From compounding to derivation: The emergence of derivational affixes through "constructionalization". *Folia Linguistica* 48(2). 579–604.

Jäger, Gerhard & Rosenbach, Anette. 2008. Priming and unidirectional language change. *Theoretical Linguistics* 34(2). 85–113.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar, volume 1: Theoretical prerequisites*. Stanford: Stanford University Press.

Langacker, Ronald W. 2008. *Cognitive grammar: A basic introduction*. New York: Oxford University Press.

Lehmann, Christian. 2002. New reflections on grammaticalization and lexicalization. In Wischer, Ilse & Diewald, Gabriele (eds.), *New reflections on grammaticalization*, 1–18. Amsterdam: John Benjamins.

Lehmann, Christian. 2015. *Thoughts on grammaticalization*. 3rd edn. Berlin: Language Science Press.

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

Liu, Ping-Ping & Li, Wei-Jun & Lin, Nan & Li, Xing-Shan. 2013. Do Chinese readers follow the national standard rules for word segmentation during reading? *PLOS One* 8(2). 1–13. (Article e55440.)

Lüdeling, Anke & Evert, Stefan & Heid, Ulrich. 2000. On measuring morphological productivity. In Schukat-Talamazzini, Ernst G. & Zühlke, Werner (eds.), *KONVENS-2000 sprachkommunikation*, 57–61. Berlin: VDE-Verlag.

MacKay, David J. C. 2003. *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Martin, Andrea E. & Doumas, Leonidas A. A. 2017. A mechanism for the cortical computation of hierarchical linguistic structure. *PLOS Biology* 15(3). 1–23. (Article e2000663.)

Miller, Kevin F. 2002. Children's early understanding of writing and language: The impact of characters and alphabetic orthographies. In Li, Wenling & Gaffney, Janet S. & Packard, Jerome L. (eds.), *Chinese children's reading acquisition: Theoretical and pedagogical issues*, 17–29. Dordrecht: Kluwer.

Norman, Jerry. 1988. *Chinese*. Cambridge: Cambridge University Press.

Packard, Jerome L. 1998. Introduction. In Packard, Jerome L. (ed.), *New approaches to Chinese word formation: Morphology, phonology and the lexicon in Modern and Ancient Chinese*, 1–34. Berlin: Mouton de Gruyter.

Perek, Florent. 2018. Recent change in the productivity and schematicity of the *way*-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory* 14(1). 65–97.

Saalbach, Henrik & Stern, Elsbeth. 2004. Differences between Chinese morphosyllabic and German alphabetic readers in the Stroop interference effect. *Psychonomic Bulletin & Review* 11(4). 709–715.

Shannon, Claude E. 1948a. A mathematical theory of communication. *The Bell System Technical Journal* 27(3). 379–423.

Shannon, Claude E. 1948b. A mathematical theory of communication. *The Bell System Technical Journal* 27(4). 623–656.

Sproat, Richard W. & Shih, Chilin & Gale, William & Chang, Nancy. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics* 22(3). 377–404.

Sweetser, Eve E. 1988. Grammaticalization and semantic bleaching. In Axmaker, Shelley & Jaisser, Annie & Singmaster, Helen (eds.), *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, 389–405. Berkeley: Berkeley Linguistics Society.

Traugott, Elizabeth Closs & Trousdale, Graeme. 2010. Gradience, gradualness and grammaticalization: How do they intersect? In Traugott, Elizabeth Closs & Trousdale, Graeme (eds.), *Gradience, gradualness and grammaticalization*, 19–44. Amsterdam: John Benjamins.

Traugott, Elizabeth Closs. 2008. Grammaticalization, constructions and the incremental development of language: Suggestions from the development of degree modifiers in English. In Eckardt, Regine & Jäger, Gerhard & Veenstra, Tonjes (eds.), *Variation, selection, development: Probing the evolutionary model of language change*, 219–250. Berlin: Mouton de Gruyter.

Tuggy, David. 2007. Schematicity. In Geeraerts, Dirk & Cuyckens, Hubert (eds.), *The Oxford handbook of cognitive linguistics*, 82–116. New York: Oxford University Press.

Address for correspondence

Jiajun Tao
School of Chinese Language and Literature
Soochow University
199 Ren'Ai Road, Suzhou Industrial Park
Suzhou, Jiangsu Province 215123
China

taojiajun@suda.edu.cn

Co-author information

Liulin Zhang
Soochow University
liulinz@suda.edu.cn

Publication history

Date received: 1 March 2023
Date accepted: 12 October 2023
Published online: 3 February 2025