Words as Basic Lexical Units in Chinese

James Myers
National Chung Cheng University

1 Introduction

Words pose a theoretical challenge in Chinese, but words pose a challenge in any language. There is as much evidence for a level of representation somewhere between morphemes and phrases in Chinese as there is in English. Yet as with English, the evidence reveals Chinese words as dynamic, sometimes more like lexicalized wholes and sometimes more like decomposable complexes. Moreover, while semantics, phonology, syntax, and psycholinguistics all point roughly at a word level in Chinese (and English), they do not point in exactly the same direction.

In other words, it is important to place the Chinese wordhood problem in a wider cross-linguistic and theoretical context. I begin in section 2 by reviewing well-known problems with wordhood in English. In section 3 I review the copious evidence for the word level in Chinese, from linguistic analysis to psycholinguistic research, along with the signs that words in Chinese are as variable in nature as they are in English. In section 4 I sketch out a formalism for expressing the dynamism of wordhood, and apply it to Chinese. Section 5 gives some brief conclusions.

2 Wordhood in English

Chao (1968, p. 136) famously wrote that "[n]ot every language has a kind of unit which behaves in most (not to speak of all) respects as does the unit called 'word' when we talk or write in English about the subunits of English." While he is right that the universality of words cannot simply be assumed, he also seems to give the impression that the status of English words is obvious. As this section reviews, it is not.

There are good reasons for this. Words are central to linguistics, lying at the interface between the lexicon and grammar; they encode arbitrary semantic and phonological properties in the lexicon while simultaneously serving as the atoms of productive syntax. However, interfaces cannot be described completely within a single domain. Native speakers also have folk-linguistic notions about words, making the linguist's job that much harder.

Linguists have thus long recognized that words are tricky things. The morphology textbook of Matthews (1991) doesn't even ask "What are words?" until p. 208; Haspelmath (2011) questions whether words can be defined at all. Consider the influential early attempt in Bloomfield (1926), where a word is a "minimum free form." Yet it is difficult to come up with plausible contexts in which function words like *the* or *to* or even uninflected content words like *know* or *cat* can constitute utterances on their own.

Or consider the orthography-based definition of English folk linguistics: a word is a string of letters surrounded by spaces or punctuation (not counting apostrophes or hyphens). However, word spacing intuitions vary; there are many compounds about which even copyeditors (or copy editors?) cannot agree. More importantly, orthographic spacing doesn't correspond consistently with other wordhood tests. To cite a textbook example, *white house* and *White House* are both written with internal spaces, but the first is argued to be a phrase

because it is semantically compositional and has phrase-final stress, while the latter is argued to be a word because it has noncompositional semantics and compound-initial stress.

The arguments for the wordhood of *White House* themselves have problems, however. After all, phrases also seem capable of having noncompositional semantics. This is most obvious in the case of idioms like *kick the bucket* (die), where even the syntax is partially lexicalized: the unacceptability of **the bucket was kicked* parallels that of its figurative meaning **I was died* (Newmeyer 1974). English also has so-called phrasal verbs like *blow up* with special semantics (explode), despite their syntactically separable parts (*blow something up*).

Phonological tests are also limited. *White House* is stressed as it is not because it is a word generally speaking, but specifically because it is a phonological word, a prosodic unit containing one main stress. Phonological words can include syntactically free clitics: the familiar suffix voicing assimilation in *cats* [s] / *dogs* [z] also applies in *The cat's* [s] / *dog's* [z] *here* (contraction of *is*).

English words also receive ambiguous support from language processing. The most robust finding in all of psycholinguistics is the word frequency effect, whereby the speed and accuracy with which a word is processed increases with how often it has been processed before, as estimated from corpus token counts (Monsell 1991). While this suggests that words are treated as memorized wholes at some stage of processing, frequency effects are observed for larger and small constituents as well. Common phrases like *don't have to worry* are responded to more quickly than rarer ones like *don't have to wait*, even with word frequencies taken into account (Arnon and Snider 2010). Responses to words are also generally faster when they contain more common morphemes, again even with whole-word frequency taken into account, and exceptions to this pattern have led some researchers to argue that word frequency effects themselves arise from a late processing stage where morphemes are combined (e.g. Taft 2004).

To some linguists, the mixed results from semantics, phonology, and processing are irrelevant: if words serve as atoms in the discrete combinatorial system of grammar, then the only tests that matter are syntactic. In particular, a constituent is claimed to be a word if and only if syntax cannot manipulate or otherwise refer to its internal components, a principle known as the Lexicalist Hypothesis or the Lexical Integrity Hypothesis (Chomsky 1970, Huang 1984, Bruening forthcoming).

Unfortunately, the Lexicalist Hypothesis also faces empirical challenges. One class of exceptions consists of word-internal phrases, an impossibility under this hypothesis. An *atomic scientist*, for example, is someone who works on *atomic science*, not a scientist who is atomic; the suffix *-ist* is thus somehow attached to an adjective + noun phrase (Spencer 1988). Native speakers are also quite confused over how to apply derivational affixes to phrasal verbs, leading to nonstandard forms like *blower-upper* (Bauer 1983: 71). Phrase incorporation may be even more productive with compounding: if one *thinks outside the box*, one is an *outside-the-box thinker* (see Bruening forthcoming for many more examples).

Syntactic operations can also apply to "word"-internal morphemes. For example, *White House* and *Blue House* (the official residence of the South Korean president) can be coordinated as *the White and Blue Houses*.[1] Affixes can also be conjoined, as long as they are sufficiently productive, as in *pro- and anti-democracy* (Duanmu 1998:139; see also Bruening forthcoming). Even in the paper that first formalized the Lexicalist Hypothesis, Chomsky (1970) argued that the parallels between verb phrases like that in *he refused the offer* and

[1] Source: https://www.theguardian.com/commentisfree/2016/mar/23/north-korea-rhetoric-pyongyang-us-china-nuclear-destruction. Accessed 26 August 2017.

nominalized gerundive phrases like *his refusing the offer* demonstrate a transformational relationship between them. Nominalized forms that cannot be analyzed transformationally still preserve the syntactic behavior of their roots. For example, *destroy* is intrinsically causative, permitting a thematic object in *his destruction of the city*, but *grow* is intrinsically inchoative, forbidding **his growth of the tomatoes* (see discussion in Marantz 1997).

Even if we were somehow able to harmonize such evidence with the Lexicalist Hypothesis (see attempts in Newmeyer 2009), we would still face a fundamental conceptual problem. Syntactic tests for wordhood presume that words and phrases are built via completely distinct operations (syntax vs. morphology), but this just makes the wordhood question harder, not easier, by pushing it up one level of abstraction (Matthew, 1991: 208).

There are two lessons here. First, it is wrong to think that wordhood is somehow trivial in languages other than Chinese. Second, words nevertheless do retain an elusive reality. Surely it is not mere coincidence that the "words" revealed by semantics, phonology, processing, and syntax overlap to a great extent. This holds as much for Chinese as it does for English.

3 Wordhood in Chinese

As in English, the Chinese folk-linguistic "word" (or as Chao 1968: 136 calls it, the sociological word) is orthographic, but a Chinese character is more like the linguist's morpheme than the linguist's word. This has led some linguists to argue that Chinese has no English-like words at all (see review in Huang et al. in this volume). Nevertheless, as I show in this section, it is as correct to posit polymorphemic lexical units in Chinese as it is in English. In doing this I build on previous reviews of the Chinese wordhood question, including Chao (1968), Duanmu (1998, 2017), Packard (2000), and the many other works that they cite.

3.1 Some basic facts

A key typological feature of Chinese is the monosyllabicity of its morphemes. This aligns basic meanings with precisely those phonological units that organize articulatory gestures and leave the clearest acoustic traces. The resulting high salience of Chinese morphemes is likely responsible for a cascade of other typological features, including the mostly syllable-timed prosody (with caveats to be noted below), the preference for compounding over affixation (given that affixes tend to become destressed or even subsyllabic), and the morpheme-based writing system, with its concomitant lack of word boundary markings. Monosyllabicity is the driver, not orthography: Thai and Vietnamese also have monosyllabic morphemes, syllable-timed phonology, and a paucity of affixation, but their orthographies are alphabetic.

Nevertheless, Chinese linguists have recognized polymorphemic lexical units, christening them with the repurposed term 词 *ci*, for over a century (Zádrapa 2017), and modern linguists can now see that bimorphemic words were already present in the very earliest stages of written Chinese and quickly came to dominate the lexicon (see Feng, this volume). While traditional Chinese dictionaries give separate entries for each character, contemporary word-oriented dictionaries are common, as are word-segmented corpora. Using such sources, it is often noted that the type frequency for multi-character words, that is, the number of distinct lexical items, is far higher than for one-character words (e.g. Zhou and Marslen-Wilson 1995). Of course token frequency, or the number of times a lexical item appears in a corpus, is highest for one-character words, but this is merely the consequence of Menzerath's Law (Schindelin 2017); in particular, function words universally tend to be both frequent and short.

In my own calculations using the Academia Sinica Balanced Corpus of Modern Chinese (version 4.0, with around ten million word tokens, around half a million transcribed from speech; Huang et al. 1997), the overall mean word length is 1.6 characters. Two-character

words not only have the highest type frequency, but are also the most productive word size, having the steepest growth in number of types as one samples more tokens (see Myers and Tsay 2015 for similar results in Southern Min).

Of course these points simply take for granted the wordhood judgments of linguists, dictionary makers, and corpus compilers. An initial hint that Chinese words are truly comparable to those in other languages comes from an analysis of a parallel corpus compiled by Ziemski et al. (2016) of 8,000 translation-equivalent sentences randomly selected from United Nations documents in the six official UN languages: Arabic, Chinese, English, French, Russian, and Spanish. All but Chinese are written with word boundaries, making it easy to compare the number of characters per Chinese sentence with the number of orthographic words per matching sentence in the other five languages. Particularly interesting were the slopes of the best-fit lines, expressed as the regression coefficient *B*, for each language pair. When the number of English words was predicted from the number of words in each of the other languages with orthographic word marking, the slopes were all close to one, implying roughly one-to-one correspondences (Arabic: $B = 1.06$; French: $B = 0.82$; Russian: $B = 1.02$; Spanish: $B = 0.80$). However, when the number of Chinese characters was predicted from the number of words in each of the other five languages, the slopes hovered around 1.5 (Arabic: $B = 1.84$; English: $B = 1.63$; French: $B = 1.38$; Russian: $B = 1.75$; Spanish: $B = 1.34$), which happens to be the mean Chinese word length noted in the previous paragraph. Whatever Chinese "words" may be, at least we can say that they reflect concepts similar in size to those expressed by "words" in a variety of other languages.

3.2 Semantic tests

Just as in English, idiosyncratic meanings can be associated not just with morphemes but also with morpheme strings, as illustrated in (1).

(1)

a. 红花
   hong__hua
   red__flower
   safflower

b. 吃饭
   chi__fan
   eat__rice
   dine

Safflowers are indeed red flowers, but they may also be yellow, and anyway all other types of red flowers have their own names; *chifan* includes the eating of noodles. Similarly, 开车 *kaiche* is literally 'open car' but here *kai* means 'operate', and 头发 *toufa* is literally 'head hair' but only refers to hair on the top of the head, not beards.

Of course, as noted by Chao (1968), Duanmu (1998), Packard (2000) and others, idioms also have idiosyncractic semantics in Chinese, as in any language. To take an arbitrary example, 对牛弹琴 *duiniutanqin* literally means 'play a qin (a stringed instrument) to a cow' but figuratively means 'speak to somebody who does not understand.' Speakers thus have to memorize that it does not have some other figurative meaning, such as 'soothe an angry person with kind words.'

But does the idiosyncrasy of idioms mean that semantics fails to diagnose wordhood at all, or does it suggest that idioms truly are wordlike in some sense? Not all linguists dismiss the latter inference so easily; in the Sinica Corpus, for example, *duiniutanqin* is not only left non-decomposed but also tagged as an active intransitive verb, just like 跑 *pao* 'run'.

Idioms often behave like words in the syntax as well. A particularly dramatic example of this is given in (2) (from Shi 2000: 390). While *dayu chi xiaoyu* is structured like a full sentence, the syntactic context shows that it is actually a predicate (the adverb *zhuanmen* cannot appear before a subject, so the subject here must be *tamen*, not *dayu*). Thus Chinese idioms seem to share the same ambiguous status as English idioms like *kick the bucket*, which is phraselike in form but wordlike in its meaning and resistance to passivization.

(2) 他们专门大鱼吃小鱼。
tamen__zhuanmen__dayu__chi__xiaoyu
they__specially__big-fish__eat__little-fish
They are doing nothing but acting according to the law of the jungle.

3.3 Phonological tests

Despite the centrality of the syllable in Chinese, larger phonological constituents also play a role. When Tseng et al. (2005) studied the acoustic properties of read-aloud Taiwan Mandarin speech, they found that phonological words, distinct from syllables and prosodic phrases, were also necessary in their statistical model, with speakers shortening word-initial syllables and lengthening word-final ones, even when phrasal effects were taken into account.

Phonological words also affect speech planning. Chiu (2005) prompted native speakers of Taiwan Mandarin to construct sentences like those in (3), where the b and c sentences both have one more syllable than the a sentence, but in b this syllable is an unstressed function morpheme and in c it is a stressed lexical morpheme. Preparation time (from prompt to the onset of speech) was the same for the a and b sentences but longer for the c sentence, suggesting that Mandarin speakers, like the Dutch speakers tested with this method by Wheeldon and Lahiri (1997), mentally chunk their utterances into phonological words.

(3)

a. 他買課本
ta__mai__keben
he__buy__ textbook
He buys a textbook.

b. 他買了課本
ta__mai__le__keben
he__buy__ASP__textbook
He bought a textbook.

c. 他買錯課本
ta__mai__cuo__keben
he__buy__wrong__textbook
He buys the wrong textbook.

Due to the monosyllabicity of Chinese morphemes, bimorphemic compounds also correspond to another prosodic constituent: the metrical foot. Qin and Duanmu (forthcoming) used experimentally elicited native speaker judgments to confirm the old observation that when a disyllabic noun is compounded with a monosyllabic noun, it is preferable for the former to precede the latter, as in (4). This pattern suggests that feet are aligned with bimorphemic compounds, which in turn suggests that these compounds have psychological reality (see Duanmu 1998 for further arguments of this sort from Shanghai). Note, by the way, the elastic word sizes (Duanmu 2017), but this is yet another Chinese-typical phenomenon that also happens to be attested in English (*William~Will*, *laboratory~lab*, *telephone~phone*, *refrigerator~fridge*).

(4)

a. 技术工
   jishu__gong
   skill__work
   skilled worker
b. *技工人
   ji__gongren
   skilled__worker
   skilled worker

3.4 Psychological tests

While all linguistic evidence is ultimately psychological evidence, the latter also includes data that are not easy to classify within grammatical theory. Here I review two general types relevant to Chinese wordhood: native-speaker intuitions (3.4.1) and language comprehension experiments (3.4.2).

3.4.1 Intuitions

Although some linguists dismiss Chinese intuitions as too muddled to be of any use (e.g. Duanmu 1998), closer examination reveals regularities within the variability, just as in English.

An amusing example of this is the bilingual Chinese/English title on an old tourist map of the Lion's Head Hill Scenic Area in Hsinchu, Taiwan.[2]. In this bit of anonymous and undated ephemera (perhaps from the 1980s), the Chinese title is written horizontally as in (5a) (without the numbered brackets, of course, which I will explain later). Bizarrely, however, the English "translation" written beneath it appears on the map as in (5b).

(5)

a. [台省]$_1$ [名勝]$_2$ [獅頭山]$_3$ [遊覽圖]$_4$
   Taisheng__mingsheng__shitoushan__youlantu
   Taiwan-province__scenic-spot__Lion-Head-Hill__sightseeing-map
   Sightseeing map for Lion's Head Hill, a scenic spot of Taiwan Province
b. [WANDER PERUSAL FIGURE]$_4$ [LION HEAD HILL]$_3$ [TITLE WIN]$_2$ [STAGE PROVINCE]$_1$

There is method in this madness, though. First, the translator made the traditional assumption that each Chinese character corresponds to one word, so that 台省 *Taisheng* 'Taiwan Province' is translated character by character as 'stage province', and so on, presumably with the help of a character-based dictionary. Second, the translator had the idiosyncratic belief that English word order is the reverse of Chinese, possibly also inspired by orthography (horizontal text is traditionally written from right to left in Chinese, as in an earlier version of this map, before the English was added).

Despite these folk-linguistic influences, the translator also obeyed a third, perhaps more natural, principle: group characters together in words. This is indicated by the numbered brackets in (5), showing that the translator segmented the title into right-headed nominal compounds, just as I did in my English gloss in (5a).

---

[2] Source: https://0.share.photo.xuite.net/dbfish66/10c1b52/15612621/836163205_m.jpg. Accessed 24 August 2017.

Wordhood intuitions are also reflected in variant word segmentations in the Sinica Corpus. Despite the corpus creators' attempt to implement strict conventions (Huang et al. 1997) via well-trained human assistants and computer algorithms, splitting the corpus by punctuation reveals a small number of character strings that are segmented differently in different places. Typical examples are shown in (6) (|| marks word segmentations).

(6)

a. 伴我成長 ~ 伴 || 我 || 成長
   ban__wo__chengzhang
   ban__wo__chengzhang
   grow up with me

b. 不要 ~ 不 || 要
   bu__yao
   not__want
   do not want

These variations are not random. Example (6a) shows a proper name (the title of a Christian song) treated sometimes as a whole and sometimes as a syntactic phrase (in the latter case even when enclosed in quotation marks 「 」). Example (6b) shows the optional cliticization of a function morpheme (other variably segmented examples like this in the corpus involve 就 *jiu* 'thus', the modifier marker 的 *de*, and the final particle 了 *le*).

Word segmentation intuitions have also been explored experimentally, though not entirely satisfactorily. Hoosain (1992) interpreted the inconsistent word markings made by Cantonese-speaking students in (Mandarin) text as showing that there is no clear concept of word in Chinese. Yet based on the few examples he cites, the students' segmentations conformed to the same regular variability seen in the Sinica Corpus. For example, they often treated 就是 *jiushi* 'thus is' as a whole, even though Hoosain, following Chao (1968), considers it syntactically separable (as in 就一定是 *jiu yiding shi* 'thus definitely is'). The students' tendency to cliticize *jiu* to *shi* is nevertheless no more "wrong" than it would be for an English student to write *it is* as *it's*.

More recently, Lin et al. (2011) asked Taiwanese students of various ages to circle two- or three-character words (*ci*) randomly embedded in lines of characters that also contained random character strings and phrases. While the number of detected words increased with age (up to 95% for the oldest, university undergraduates), all of the students circled far more words than phrases. However, even the undergraduates treated 50% of the phrases as words. Of course, as with the Hoosain study, at most this shows a mismatch with the experimenters' own stipulated wordhood intuitions. More interestingly, the authors found that the participants' wordhood responses depended in part on the transition probabilities between characters, a variable that we will see reappear several more times in this chapter.

3.4.2 Language comprehension

Words have also been observed in Chinese language comprehension, particularly in reading. Spoken language comprehension is still understudied in Chinese, but word-driven models of Chinese listening comprehension are intrinsically more plausible than morpheme-driven ones, simply because whole words have far fewer homophones than do individual morphemes (Packard 1999). The ambiguity created by homophony is likely also why higher syllable frequency slows down the recognition of isolated spoken words (Zhou and Marslen-Wilson 1994). Chinese listeners also seem to segment words while listening to sentences; Ding et al. (2016) found that brain waves track bimorphemic constituents when listening to simple noun-

verb Chinese sentences, though they did not specifically test if these constituents were being treated as words and not merely as syntactic units.

In reading, characters make morphemes far more salient than words. Despite this, Chinese readers recognize characters more readily if they are embedded in real words (Mok 2009) and reading times are slowed if characters are split at places other than word boundaries (Bai et al. 2008). Moreover, as in English, the most robust finding in Chinese psycholinguistics is the facilitative effect of word frequency (Myers 2006, 2017).

Also as in English, however, Chinese shows frequency effects in constituents both larger and smaller than lexical words. Liu (2015) found that the frequency of idioms affected their acceptability and learning, and Myers et al. (2006) found that readers in Taiwan were faster to respond to a two-character verb followed by the durative aspect morpheme 著 *zhe* the more frequent the whole construction, even with character and verb frequencies factored out. Unsurprisingly, character frequency also affects written word recognition (Myers 2006, 2017). While common morphemes usually speed responses, as in English, they slow responses in semantically opaque compounds (where morpheme and word meanings conflict), and since compounds with less common characters tend to have higher cross-character transition probability, this may speed responses via enhanced word-internal cohesiveness.

Words also seem to play a central role in sentence reading. In a particularly data-rich study, Li et al. (2014) found that eye movements in Chinese reading are influenced by the same word-level variables affecting English reading: word length, frequency, and contextual predictability. Even the character-level properties that also affect eye movements, like frequency and visual complexity, do so by affecting the detection of words in upcoming text.

3.5 Syntactic tests

While semantic, phonological, intuitive, and processing data all suggest that Chinese has words, at least to the same fuzzy extent that English does, Chinese linguists, like linguists generally, are particularly interested in syntactic evidence (e.g. Chao 1968, Duanmu 1998, Packard 2000). Syntactic tests do work much better than one would expect if Chinese did not have words at all, but as we saw with English in Section 2, they also have limitations.

Here I focus on just two of the most notorious wordhood problems in Chinese: adjective-noun (AN) and verb-object (VO) constructions. Regarding the first, Duanmu (1998) reviews a variety of syntactic tests strongly suggesting that AN is a word and A 的 *de* N is a phrase. In contrast to A *de* N, AN is not fully productive (e.g. the semantically compositional English phrase *white hand* is not translated as 白手 *baishou* ‘white hand’). It also cannot be modified by a degree word, as shown in (7) (the same pattern is seen with 最 *zui* ‘most’, 这么 *zheme* ‘such a’, and 不 *bu* ‘not’), disallows internal phrases, as shown in (8), and disfavors coordination of internal parts, as shown in (9), these last three points being special cases of the Lexicalist Hypothesis (all examples are taken from or based on those in Duanmu 1998).

(7)

a. 新书
   xin__shu
   new__book
   new book

b. 新的书
   xin__de__shu
   new__DE__book
   new book

c. *很新书
hen__xin__shu
very__new__book
d. 很新的书
hen__xin__de__shu
very__new__DE__book
very new book

(8)

a. *新[三本书]
xin__san__ben__shu
new__three__CL__book
b. 新的[三本书]
xin__de__san__ben__shu
new__DE__three__CL__book
three new books
c. *[有名的作者]书
youming__de__zuozhe__shu
famous__DE__author__book
d. [有名的作者]的书
youming__de__zuozhe__de__shu
famous__DE__author__DE__book
book by a famous author

(9)

a. *旧跟新书
jiu__gen__xin__shu
old__and__new__book
b. 旧跟新的书
jiu__gen__xin__de__shu
old__and__new__DE__book
old and new books

As important as such evidence is for demonstrating some sort of reality for Chinese words, all of these tests have flaws. For example, A *de* N constructions can accrue lexical idiosyncrasies; (10) happens to be the title in Taiwan for the *Mission: Impossible* movies, a fact that speakers take into account when choosing to use it.

(10) 不可能的任務
bukeneng__de__renwu
impossible__DE__mission
impossible mission

Regarding the adverbial modification test, Duanmu admits that in certain fixed constructions *zui* can modify the A in AN constructions, as in (11), suggesting that some morphemes can occasionally change status from phrasal to word-forming.

(11) 最高级
zui__gao__ji
most__ high__level
the most high level

While Duanmu seems right to clam that numeral-classifier-noun and *de*-phrases cannot be embedded within words, other phrase-like constructions do so as readily as they do in English, as in (12) (from Wiese 1996: 185).

(12) 百花齊放运动
bai__hua__qi__fang__yundong
hundred__flower__together__bloom__movement
Hundred Flowers movement

Finally, Duanmu himself finds the coordination test unreliable, citing counterexamples in English (quoted earlier in Section 2). While he gives no Chinese counterexamples, the form in (13) seems to be one, merging 进口 *jinkou* 'import' and 出口 *chukou* 'export', though without an overt coordinator.

(13) 进出口
jin__chu__kou
enter__exist__mouth
import and export

Turning to VO constructions, the problem here is that syntactic tests show that some are consistently phrase-like (e.g. like 吃面 *chi mian* 'eat noodles'), others are consistently word-like (e.g. 出版 *chuban* 'publish', literally 'output version'), and others alternate between the two statuses (e.g. 担心 *danxin* 'worry', literally 'carry heart'; Chao 1968, Huang 1984, Packard 2000). This is illustrated in (14)-(16) using three syntactic diagnostics: phrases allow inflection on the verbal morpheme, topicalization of the object, and a constituent-external object, while words do not (judgments checked in an informal poll of native speakers).

(14)

a. 他们吃了面
tamen__chi__le__mian
they__eat__ASP__noodles
They ate noodles

b. 他们出版了
tamen__chuban__le
they__publish__ASP
They published (something)

c. *他们出了版
tamen__chuban__le
they__publish__ASP

d. 他们担了心
tamen__dan__le__xin
they__carry__ASP__heart
They worried

e. 他们担心了
tamen__danxin__le
they__worry__ASP
They worried

(15)

a. 面，他们一点都没吃
mian__tamen__yidian__dou__mei__chi
noodles__they__a-bit__all__not__eat
As for noodles, they didn't eat them at all
b. *版，他们一点都没出
ban__tamen__yidian__dou__mei__chu
version__they__a-bit__all__not__output
c. 心，他们一点都没担了
xin__tamen__yidian__dou__mei__dan
heart__they__a-bit__all__not__carry
As for worried, they were not at all.

(16)

a. *他们吃面蛋糕
tamen__chi__mian__dangao
they__eat__noodle__cake
b. 他们出版了一本书
tamen__chuban__le__yi__ben__shu
they__publish__ASP__one__CL__book
They published a book
c. 他们担心你
tamen__danxin__ni
they__worry-about__you
They are worried about you.

Chinese is hardly the only language with items that alternate between words and phrases (recall *think outside the box ~ outside-the-box thinker* from Section 2). Similarly, according to all three tests, *danxin* either behaves as composed in the syntax or as a syntactic atom, never both at once. Nevertheless, such phenomena suggest that, like the "word"-internal *zui* in (11), the word/phrase boundary is permeable, whether via lexicalization of phrases into words (Huang 1984), reanalysis of words as phrases (Packard 2000), or some combination of both. As usual, syntactic tests also need not correlate with tests from other linguistic domains; as mentioned earlier, *chifan* means dining in general, not just on rice, but syntactically it behaves just as phraselike as *chimian*.

It is not even clear exactly how consistent the syntactic tests are among themselves. For example, (17a) shows what Chao (1968: 433) called the "ionization" (splitting) of 幽默 *youmo* 'tease' (from English *humor*) that he observed in a Taiwanese newspaper, supplemented with examples in (17b) and (17c) created by Huang (1984: 65). But the Taiwan Mandarin speakers I consulted tended not to accept the "non-ionized" form in (17c) and only some accepted the topicalized structure in (17b); many accepted (17a), but this is not verb plus object, but verb plus verbal classifier. As far as I know, such variation has yet to be investigated, whether via corpus analysis or formal judgment experiments, in a properly representative sampling of speakers and constructions.

(17)

a. 還幽了他一默，說...
   hai__you__le__ta__yi__mo__shuo
   also__hu-__ASP__he__one__-mor__say
   (I) teased him again, saying...

b. 這種默，我想你最好還是不要幽
   zhe__zhong__mo__wo__xiang__ni__zuihao__hai__buyao__you
   this__kind__-mor__I__think__you__best__still__not__hu-
   This way of teasing, I think you'd better not do it.

c. 我常常幽默他
   wo__changchang__youmo__ta
   I__often__tease__he
   I often tease him.

To summarize all of section 3, then, the evidence for words is as strong in Chinese as in English, but as with English, we have to accept that wordhood tests do not entirely agree and that wordhood status also varies across the context of use.

4 A dynamic approach to wordhood

This shifting reality of words is just what we would expect if words serve multiple functions in the dynamic interaction of multiple linguistic forces. In 4.1 I argue that this may be our most promising theoretical option among the alternatives and suggest how the idea may be formalized. In 4.2 I show how this formalism captures several important aspects of Chinese wordhood.

4.1 Formalizing fuzziness

The fuzziness of the word/phrase distinction is so notorious that it has led to two diametrically opposed attempts to eliminate words altogether: either it's syntax all the way down (e.g. Bruening forthcoming; Marantz 1997, 2013), or the lexicon all the way up (e.g. Baayen and Ramscar 2015; Daelemans and van den Bosch 2005; Jackendoff and Audring 2016). Both approaches seem oversimplified: putatively all-syntax approaches actually shunt irregularities off to non-syntactic components, while all-lexicon approaches underestimate the difficulty of generating syntactic regularity by analogy alone (e.g. recursion and long-distance dependencies).

I favor a dynamic and pluralistic approach, viewing the shifty nature of words as a conspiracy where the conspirators do not quite agree. In Chinese, for example, single morphemes are too small to express the complex meanings, rhythmic prosody, syntactic architecture, and user-friendly expressions needed for effective communication, so the semantics adds hidden meanings, phonology builds clitic groups, and syntax and processing fight over their own favored constituent sizes.

Aspects of this approach have already been formalized. Jackendoff and Audring (2016) encode the interfacing of semantics, phonology, and morphosyntax in terms of schemas. Computational models deriving wordhood from the transition probability of lower-level units are presented in Bicknell and Levy (2010) for eye movements in reading, and independently in Huang and Xue (2012) for automated Chinese text segmentation. Baayen et al. (2015) show how a computational model of child language acquisition mapping phonemes to semantic units is capable of learning "words" without performing any overt segmentation procedure at all.

To keep the discussion concrete, I will focus just on the last of these models (see Baayen and Ramscar 2015 for a non-technical overview). Naive discriminative learning (NDL), motivated by general learning theory, consists of one layer of connections between "cues" (e.g. Chinese characters) and "outcomes" (e.g. meanings) for each learning "event" (e.g. a character string in a learning context where the meaning is clear). The learning algorithm is discriminative because a cue-outcome connection is strengthened only if the cue is informative (e.g. A→X and AB→X will not generalize to B→X); it is naive because each cue-outcome connection is adjusted while ignoring all other outcomes. A schematic NDL model is shown in Figure 1 (based on (6a)); training by events in an actual corpus would strengthen some connections more than others.

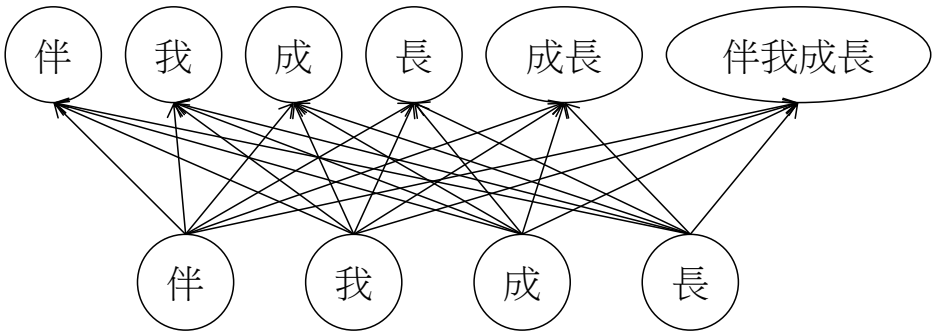

Figure 1. NDL model linking single-character cues to whole-word meaning outcomes

NDL makes an attractive formalization of dynamic wordhood for a number of reasons. As just noted, it can learn words from fluent speech without overt word segmentation, and it can also model distinct morpheme-level, word-level, and phrase-level effects within the same network (Baayen et al. 2013). It also incorporates insights from all-syntax approaches; the network nodes may be abstract, and as Marantz (2013) points out, in current practice its meaning outcomes are generally the size of morphemes. The content-neutral nature of NDL architecture also allows it to go beyond linking form with meaning, to linking pictures with meanings or even meanings with meanings (Hendrix et al. 2017). Putting these last two points together, NDL thus has the potential to include abstract syntactic elements as well, in case it turns out (as seems likely to me) that not all of syntax is reducible to analogy. Of course, by itself, NDL cannot explain where the patterns it learns come from in the first place, but perhaps it could do so by incorporating diachronic feedback loops (e.g. Kirby 2001), thereby capturing the dynamic nature of wordhood across generations as well.

4.2 Applications to Chinese

While NDL has yet to be applied systematically in Chinese, preliminary results already help explain some of the patterns discussed in sections 3.

Tsung-Ying Chen and I trained an NDL model, using the ndl package (Arppe et al. 2015) in R (R Core Team 2017), on the written transcription of the spoken portion of the Sinica Corpus, chosen primarily for its relatively small size (linking all possible cues with all possible outcomes makes NDL a memory hog, so modeling even this half-million-word corpus required around 28 GB of RAM). Events were strings of characters in the transcription, demarcated on each end by punctuation; cues were the individual characters within an event, and outcomes were the words in the event as segmented in the corpus. The *n* most activated outcomes per event, where *n* was the "actual" number of words, were taken as the trained model's word guesses. While the model correctly identified only 75% of the "actual" word tokens, this need not be the fault of NDL; our cues contained no sequential

information (e.g. ABCD was represented the same as BDCA) and by using whole words as outcomes, our model unrealistically assumes that all Chinese words are semantically opaque.

Nevertheless, even this simple model managed to capture several observations made in Section 3. For example, consistent with human wordhood intuitions and eye movements in reading, we found that the more predictable one character was from another within a two-character word, the more accurate the NDL model was at identifying this word (τ = .39, z = 58, p < .0001; we used the Kendall rank correlation coefficient due to the non-normality of both variables). This result is particularly striking given that the cues were unordered characters and thus transition probability was not coded directly.

Word frequency also improved the accuracy of our model in identifying words (τ = .32, z = 46, p < .0001). It could even do this simultaneously with modeling the effects of morpheme and phrase frequency (as Baayen et al. 2013 found for English): a multiple linear regression predicting the proportion of "actual" Chinese words detected per event showed not only a positive effect of mean word frequency (B = 0.68, t = 2323, p < .0001) but also independent effects of mean character frequency (B = -1.08, t = -151, p < .0001) and whole-string frequency (B = -0.02, t = -30, p < .0001), the latter two effects negative due to the model's human-like tendency not to decompose opaque words or common phrases.

We also explored how NDL can capture semantic tests for wordhood. Using a toy corpus, we linked cue pairs (simulating a two-character compound) with one outcome (opaque), two outcomes (transparent), or variably one or two outcomes (ambiguous). Unsurprisingly, in the last case the model distributed activations across both opaque and transparent meanings (like *honghua*, which may mean either 'safflower' or 'red flower').

Another toy model captured the word segmentation triggered by function morphemes, a key factor in many syntactic tests. We trained the events AB, BA, CD, DC, AfB, BfA, CfD, DfC, where capital letters represent content morphemes and f represents a function morpheme, with cues coded as bigrams (e.g. AB was coded as #A, AB, B#, and AfB as #A, Af, fB, B#). Outcomes assumed full semantic transparency (e.g. AB was linked with A and B, and AfB with A, f, B). When we tested the model on the untrained inputs AD and AfD, we found that activation of the individual content morphemes A and D was higher for AfD than for AD, just as in Chinese A *de* N and V *le* O are more decomposable than AN and VO. This behavior resulted from the fact that in training, the bigrams Af and fD also became associated with A and D, respectively, boosting their activation when prompted with AfD, whereas the bigram AD had never been encountered at all. In plain language, function morphemes trigger segmentation because they appear in more different contexts and thus have lower transition probabilities.

Since NDL cues are modality-specific form units, wordhood should not be the same for readers and listeners. While this has not yet been modeled in Chinese, predictions can be derived from the work of Pham and Baayen (2015) on Vietnamese, which also has monosyllabic morphemes, rampant homophony, and a predilection for compounding. They first report experimental results showing that morpheme frequency slow wordhood judgments in Vietnamese, the reverse of English. They then model this result in NDL, with Vietnamese coded in letter bigrams, and find that the lower activation of words with high-frequency morphemes is caused by homophony overloading the model's discriminative ability. NDL thus predicts that Chinese character frequency effects should generally be facilitative in reading, since like English morphemes, characters are readily discriminable, whereas for spoken Chinese, homophony should cause syllable frequency to slow responses. As we saw in Section 3.4.2, both predictions are correct.

5 Conclusions

Are words lexical units in Chinese? I think the evidence for this is as strong as it is in English or any other language. Are they basic units, though? The answer to this question depends on how one looks at it. I have argued that the variable behavior of words, in all languages, reveals them as attempts to satisfy sometimes competing, sometimes cooperating linguistic forces. In the sense that words are derived dynamically from these forces, they are not basic units. Yet given that all languages seem to have wordlike units, with roughly the same semantic, phonological, syntactic, and processing behavior (and roughly the same sizes in parallel translations), words seem to provide the optimal solution to a universal engineering problem. In the sense that words are inevitable for a well-functioning language, lying as they do at the nexus of all components of the linguistic machine, they are basic units. What remains to be seen is how precisely this machine, and the dynamic role of words within it, can be modeled.

References

Arppe, Antti, Peter Hendrix, Petar Milin, R. Harald Baayen, Tino Sering and Cyrus Shaoul. 2015. ndl: Naive discriminative learning. R package version 0.2.17. https://cran.r-project.org/web/packages/ndl/index.html

Arnon, Inbal, and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62 (1): 67–82.

Baayen, R. Harald, and Michael Ramscar. 2015. Abstraction, storage and naive discriminative learning. In *Handbook of cognitive linguistics*, ed. Ewa Dabrowska and Dagmar Divjak, 99–120. Berlin: De Gruyter Mouton.

Baayen, R. Harald, Peter Hendrix, and Michael Ramscar. 2013. Sidestepping the combinatorial explosion: An explanation of *n*-gram frequency effects based on naive discriminative learning. *Language and Speech* 56(3): 329–47.

Baayen, R. Harald, Cyrus Shaoul, Jon Willits, and Michael Ramscar. 2015. Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience* 31(1): 106–28.

Bai, Xuejun, Guoli Yan, Simon P. Liversedge, Chuanli Zang, and Keith Rayner. 2008. Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance* 34(5):1277–1287.

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

Bicknell, Klinton and Roger Levy. 2010. A rational model of eye movement control in reading. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, 1168–78. http://idiom.ucsd.edu/~rlevy/papers/bicknell-levy-2010-acl.pdf

Bloomfield, Leonard. 1926. A set of postulates for the science of language. *Language* 2(3): 153–64.

Bruening, Benjamin. Forthcoming. The Lexicalist Hypothesis: Both wrong and superfluous. To appear in *Language*.

Chao, Yuen-Ren. 1968. *A grammar of spoken Chinese*. Berkeley: University of California Press.

Chiu, Chenhao C. 2005. Phonological words in Mandarin speech production. *Berkeley Linguistics Society: Proceedings of the 31st Annual Meeting* 31 (1): 61–72.

Chomsky, Noam. 1970. Remarks on Nominalization. *Readings in English transformational grammar*, ed. by Roderick A. Jacobs and Peter S. Rosenbaum, 11–61. Waltham, MA: Ginn and Co.

Daelemans, Walter, and Antal van den Bosch. 2005. *Memory-based language processing*. Cambridge, UK: Cambridge University Press.

Ding, Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience* 19(1): 158–64.

Duanmu, San. 1998. Wordhood in Chinese. In *New approaches to Chinese word formation: Morphology, phonology and the lexicon in Modern and Ancient Chinese*, ed. Jerome L. Packard, 135–96. Berlin: Mouton de Gruyter.

Duanmu, San. 2017. Word and wordhood, modern. In *Encyclopedia of Chinese language and linguistics, vol. IV*, ed. by Rint Sybesma, 543–49. Leiden: Brill.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1):31–80.

Hendrix, Peter, Patrick Bolger, and Harald Baayen. 2017. Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43(1): 128–49.

Hoosain, Rumjahn. 1992. Psychological reality of the word in Chinese. *Language processing in Chinese*, ed. by Hsuan Chih Chen and Ovid J. L. Tzeng, 111–30. Amsterdam: North-Holland.

Huang, James C.-T. 1984. Phrase structure, lexical integrity, and Chinese compounds. *Journal of the Chinese Language Teachers Association* 19(2):53–78.

Huang, Chu-Ren, and Nianwen Xue. 2012. Words without boundaries: Computational approaches to Chinese word segmentation. *Language and Linguistics Compass* 6(8): 494–505.

Huang, Chu-Ren, Keh-jiann Chen, Feng-yi Chen, and Li-Li Chang. 1997. Segmentation standard for Chinese natural language processing. *Computational Linguistics and Chinese Language Processing* 2(2): 47–62.

Jackendoff, Ray and Jenny Audring. 2016. Morphological schemas: Theoretical and psycholinguistic issues. *The Mental Lexicon* 11 (3), 467–93.

Kirby, Simon. 2001. Spontaneous evolution of linguistic structure – an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation* 5(2): 102–10.

Li, Xingshan, Klinton Bicknell, Pingping Liu, Wei Wei, and Keith Rayner. 2014. Reading is fundamentally similar across disparate writing systems: A systematic characterization of how words and characters influence eye movements in Chinese reading. *Journal of Experimental Psychology: General* 143 (2): 895–913.

Lin, Tzu-Jung, Richard C. Anderson, Yu-Min Ku, Kiel Christianson, and Jerome L. Packard. 2011. Chinese children's concept of word. *Writing Systems Research* 3(1): 41–57.

Liu, Li. 2015. *Chinese quatra-syllabic schematic idioms: Description and acquisition*. Hong Kong: The Hong Kong Institute of Education dissertation.

Marantz, Alec. 1997. No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. *University of Pennsylvania Working Papers in Linguistics* 4(2): 201–25.

Marantz, Alec. 2013. No escape from morphemes in morphological processing. *Language and Cognitive Processes* 28(7):905–16.

Matthews, Peter. H. 1991. *Morphology* (2nd edition). New York: Cambridge University Press.

Mok, Leh Woon. 2009. Word-superiority effect as a function of semantic transparency of Chinese bimorphemic compound words. *Language and Cognitive Processes* 24(7–8): 1039–81.

Monsell, Stephen. 1991. The nature and locus of word frequency effects in reading. In *Basic processes in reading: Visual word recognition*, ed. Derek Besner and Glyn W. Humphreys, 148–197. Hillsdale, NJ: Erlbaum.
Myers, James. 2006. Processing Chinese compounds: A survey of the literature. In *The representation and processing of compound words*, ed. Gary Libben and Gonia Jarema, 169–96. Oxford University Press.
Myers, James. 2017. Morphological processing of compounds, behavioral studies. *Encyclopedia of Chinese language and linguistics, vol. III*, ed. by Rint Sybesma, 94–100. Leiden: Brill.
Myers, James, Yu-chi Huang, and Wenling Wang. 2006. Frequency effects in the processing of Chinese inflection. *Journal of Memory and Language* 54 (3): 300–23.
Myers, James and Jane Tsay. 2015. Trochaic feet in spontaneous spoken Southern Min. *Proceedings of the 27th North American Conference on Chinese Linguistics, Vol. 2*, Los Angeles, 368–87. https://naccl.osu.edu/proceedings/naccl-27
Newmeyer, Frederick J. 1974. The regularity of idiom behavior. *Lingua* 34:327–42.
Newmeyer, Frederick J. 2009. Current challenges to the lexicalist hypothesis: An overview and a critique. *Time and again: Theoretical perspectives on formal linguistics in honor of D. Terence Langendoen*, ed. by William D. Lewis, Simin Karimi, Heidi Harley, and Scott O. Farrar, 91–117. Amsterdam: John Benjamins.
Packard, Jerome L. 1999. Lexical access in Chinese speech comprehension and production. *Brain and Language* 68(1): 89–94.
Packard, Jerome L. 2000. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.
Pham, Hien, and Harald Baayen. 2015. Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience* 30(9): 1077–95.
Qin Zuxuan and San Duanmu. Forthcoming. A judgment study of word-length preferences in Chinese NN compounds. To appear in *Lingua*.
R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://cran.r-project.org/
Schindelin, Cornelia. 2017. Menzerath's Law. *Encyclopedia of Chinese language and linguistics, vol. III*, ed. by Rint Sybesma, 1–3. Leiden: Brill.
Shi, Dingxu. 2000. Topic and topic-comment constructions in Mandarin Chinese. *Language* 76(2):383–408.
Spencer, Andrew. 1988. Bracketing paradoxes and the English lexicon. *Language* 64(4):663–82.
Taft, Marcus. 2004. Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology Section A* 57(4): 745–65.
Tseng, Chiu-yu, Shao-huang Pin, Yehlin Lee, Hsin-min Wang, and Yong-cheng Chen. 2005. Fluent speech prosody: Framework and modeling. *Speech Communication* 46 284–309.
Wheeldon, Linda, and Aditi Lahiri. 1997. Prosodic units in speech production. *Journal of Memory and Language* 37:356–381.
Wiese, Richard. 1996. Phrasal compounds and the theory of word syntax. *Linguistic Inquiry* 27(1): 183–93.
Zádrapa, Lukáš. 2017. Word and wordhood, premodern. *Encyclopedia of Chinese language and linguistics, vol. IV*, ed. by Rint Sybesma, 549–54. Leiden: Brill.
Zhou, Xiaolin, and William Marslen-Wilson. 1994. Words, morphemes and syllables in the Chinese mental lexicon. *Language and Cognitive Processes* 9(3):393–422.
Zhou, Xiaolin and William Marslen-Wilson. 1995. Morphological structure in the Chinese mental lexicon. *Language and Cognitive Processes* 10(6):545–600.

Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. *Language Resources and Evaluation*, Portorož, Slovenia. https://conferences.unite.un.org/UNCorpus/Content/Doc/un.pdf