Corpus-Based Methods in Chinese Morphology and Phonology

中文構詞法和音韻學語料庫方法

Richard Sproat

AT&T Labs — Research

rws@research.att.com

www.research.att.com/∼rws

Chilin Shih

Bell Labs

cls@research.bell-labs.com

www.bell-labs.com/project/tts/cls.html

Lecture 1: Morphology

2001 LSA Institute – University of California, Santa Barbara
Subinstitute on Chinese Corpus Linguistics

Overview

- Intro to Chinese Morphology: What is a Word
- Word Frequency Distributions, Measures of Productivity, Measures of Association
- Applications

Overview of Lecture 1

- What is a Word?

- Some Chinese Morphological Phenomena

- Segmentation Standards

- Preview: What Corpus-Based Methods Can Do For Us

What is a Word?

- Having a good definition of what is a word seems like a prerequisite to any study of words in any language.

- In Chinese the problem is exacerbated by the lack of word boundaries in orthography.

- Human judges disagree: in (Sproat et al 1996) we only measured 0.76 (out of 1.00) agreement among human judges.

- Proposed segmentation standards don’t even agree.

What is a Morpheme?

- At least it is agreed that Chinese words are made up of morphemes, but: what is a morpheme?

- Morpheme = single syllable, single character?

- Problems:
  - Polysyllabic morphemes such as 東西*dōngxī* (east west) ‘things’; 馬上*mǎshàng* (horse on) ‘suddenly’
  - Borrowed morphemes: 巴基斯坦 *bājīsītǎn* ‘Pakistan’

A Clear Class of Disyllabic Morphemes

Orthography	Analysis	Pronunciation	Gloss
蹉跎	< **foot+cuōtuō** >	*cuōtuó*	‘procrastinate’
踉蹌	< **foot+liángcāng** >	*lángqiāng*	‘hobble’
蹂躪	< **foot+róulìn** >	*róulìn*	‘trample’
躊躇	< **foot+shòuzhù** >	*chóuchú*	‘hesitate’
躑躅	< **foot+zhèngshǔ** >	*zhízhú*	‘hesitate’
萵苣	< **grass+guǎjù** >	*wōjù*	‘lettuce’
菡萏	< **grass+hánxiàn** >	*hàndàn*	‘lotus’
蒹葭	< **grass+jiānjiǎ** >	*jiānjiā*	‘type of reed’
苜蓿	< **grass+mùsù** >	*mùsù*	‘clover’
慫恿	< **heart+cóngyǒng** >	*sǒngyǒng*	‘egg on’
忸怩	< **heart+niuní** >	*niǔní*	‘coy’
慇懃	< **heart+yīnqín** >	*yīnqín*	‘attentively’
蝙蝠	< **insect+biǎnfù** >	*biānfú*	‘bat’
蜉蝣	< **insect+fúyóu** >	*fúyóu*	‘mayfly’
蚯蚓	< **insect+qiūyǐn** >	*qiūyǐn*	‘earthworm’

Table 1: Pairs of characters that only cooccur with each other, and occur at least three times in a 20 million character corpus. Note that in each case both characters share the same semantic radical. See, for instance, the examples with the **foot** radical, underlined in the table above. The second column gives the component analysis following the schema in (Sproat 2000).

Packard's (2000) Notions of Word: I

- **Orthographic word:** Words as defined by delimiters in written text. This appears to have no relevance in Chinese since there are no such written delimiters (apart from punctuation marks, which mark ends of phrases, not words).

- **Sociological word:** Following Chao (1968: pp. 136–138), these are 'that type of unit, intermediate in size between a phoneme and a sentence, which the general, non-linguistic public is conscious of, talks about, has an everyday term for, and is practically concerned with in various ways.' In English this is the lay notion of 'word', whereas in Chinese this is the character (字 *zì*).

- **Lexical word:** This corresponds to Di Sciullo and Williams' (1987) *listeme*.

Packard’s (2000) Notions of Word: II

- **Semantic word:** Roughly speaking, this corresponds to a ‘unitary concept’.
- **Phonological word:** A word-sized entity that is defined using phonological criteria. (Yes, that’s circular.) Packard goes on to note that Chao considers prosodic issues, such as when one can pause in a sentence, to provide useful tests for phonological wordhood. In many languages, phonological words are the domain of particular phonological processes: in Finnish, for example, vowel harmony is restricted to phonological words. In dialects of Mandarin that have (stress) reduction, such phenomena are generally restricted to words, as is consonant lenition.

Packard’s (2000) Notions of Word: III

- **Morphological word:** following Di Sciullo and Williams’ notion, a morphological word is anything that is the output of a word-formation rule.
- **Syntactic word:** These are all and only those constructions that can occupy X^0 slots in the syntax. (The syntactic word is the definition of word that Packard uses as the basis for the bulk of his discussion.)
- **Psycholinguistic word:** This is the “‘word’ level of linguistic analysis that is . . . salient and highly relevant to the operation of the language processor” (2000, p. 13).

To these definitions could be added the extrinsically defined notions of word that have formed the basis of the various segmentation standards for Chinese that have been proposed.

Morphological Types

- Reduplication
- Affixation
- Compounding
- Proper names
- Abbreviations (縮寫 *suōxiě*)

Verbal reduplication

(1)

說說	shuō-shuō	(speak speak)	‘speak a little’
看看	kàn-kàn	(look look)	‘have a little look’
走走	zǒu-zǒu	(walk walk)	‘have a little walk’
磨磨	mó-mó	(rub rub)	‘rub a little’
討論討論	tǎolùn-tǎolùn	(discuss discuss)	‘discuss a little’
請教請教	qǐngjiào-qǐngjiào	(ask ask)	‘ask a little’
研究研究	yánjiù-yánjiù	(research research)	‘research into’

Verbal reduplication

(2)

說一說	shuō-yi-shuō	(say one say)	‘go ahead and say it’
看一看	kàn-yi-kàn	(look one look)	‘have a look’
走一走	zǒu-yi-zǒu	(walk one walk)	‘have a little walk’
磨一磨	mó-yi-mó	(rub one rub)	‘rub a little’

(3)

說說看	shuō-shuō-kàn	(speak speak look)	‘talk about it and see’
走走看	zǒu-zǒu-kàn	(walk walk look)	‘walk and see’

Adjectival Reduplication

(4)	紅紅	*hóng-hóng*	(red red)	‘red’
	慢慢	*màn-màn*	(slow slow)	‘slow’

(5)

舒服	*shūfú*	舒舒服服	*shūshū-fúfú*	‘comfortable’
乾淨	*gānjìng*	乾乾淨淨	*gāngān-jìngjìng*	‘clean’
胡塗	*hútú*	胡胡塗塗	*húhú-tútú*	‘muddle-headed’
快樂	*kuàilè*	快快樂樂	*kuàikuài-lèlè*	‘happy’
漂亮	*piàoliàng*	漂漂亮亮	*piàopiào-liàngliàng*	‘pretty’

Adjectival Reduplication

(6)	亮晶	*liàngjīng*	亮晶晶	*liàng-jīngjīng*	‘bright’
	白花	*báihuā*	白花花	*bái-huāhuā*	‘white’

(7)	∗重重要要	*zhòngzhòng-yàoyào*	‘important’
	∗偉偉大大	*wěiwěi-dàdà*	‘majestic’
	∗粉粉紅紅	*fēnfēn-hónghóng*	‘pink’
	∗?美美麗麗	*měiměi-lìlì*	‘beautiful’
	∗透透明明	*tòutòu-míngmíng*	‘transparent’

Measure word reduplication

(8)	磅磅	*bàngbàng*	‘every pound’
	條條魚	*tiáotiáo yú*	‘every fish’
	套套西裝	*tàotào xīzhuāng*	‘every suit’
	件件衣服	*jiànjiàn yīfú*	‘every piece of clothing’

(9)	*雙雙手	*zhīzhī shǒu*	‘every hand’
	*冊冊書	*cècè shū*	‘every book’
	*打打蛋	*dádá dàn*	‘every dozen eggs’
	?座座山	*zuòzuò shān*	‘every mountain’
	?隻隻雞	*zhīzhī jī*	‘every chicken’

Measure word reduplication

(10) 我們的雞,隻隻都病了

wǒmen-de jī, zhīzhī dōu bìng le

(our-DE chicken CL-CL all sick ASP)

‘Every one of our chickens is sick.’

(11)	?日日	*rìrì*	‘every day’ (okay in poetic language)
	天天	*tiāntiān*	‘every day’
	年年	*niánnián*	‘every year’

(12)	*公里公里	*gōnglǐ gōnglǐ*	‘every kilometer’

Prefixation

(13)

老	*lǎo-*	老王	*lǎo wáng*	‘old Wang’
小	*xiǎo-*	小張	*xiǎo zhāng*	‘little Zhang’
第	*dì-*	第一	*dìyī*	‘first’
初	*chū-*	初三	*chū sān*	‘the third’
好/難	*hǎo-/nán-*	好吃/難吃	*hǎochī/nánchī*	‘tasty/bad-tasting’

(14)	可	*kě-*	可愛	*kě-ài*	‘lovable, cute’
			可靠	*kě-kào*	‘reliable’

Suffixation

(15) “Diminutive” suffixes:

兒	*-er*	鳥兒	*niǎo-er*	‘bird’
子	*-zi*	猴子	*hóu-zi*	‘monkey’
頭	*-tou*	饅頭	*mán-tou*	‘steamed bread’

(16) Other deriational suffixes:

學	*-xué*	心理學	*xīnlǐ-xué*	‘psychology’
家	*-jiā*	物理學家	*wùlǐxué-jiā*	‘physicist’
化	*-huà*	美化	*měi-huà*	‘Americanization’
率	*-l`ü*	生蛋率	*shēng-dàn-l`ü*	‘egg production rate’
主義	*-zhǔyì*	馬克斯主義	*mǎkèsī-zhǔyì*	‘Marxism’

Suffixation

(17)	了	*-le*	吃了	*chī-le*	‘have eaten’
	過	*-guò*	吃過了	*chī-guò-le*	‘have eaten’
	們	*-men*	孩子們	*háizi-men*	‘children’

Compounding

(18) **location**

客廳 沙發	*kètīng shāfā*	‘living room sofa’
河 馬	*hémǎ*	‘hippopotamus (river horse)’
海 狗	*hǎigǒu*	‘seal (sea dog)’

(19) **used for**

指甲油	*zhǐjiǎ yóu*	‘nail polish’
乒乓球	*pīngpāng qiú*	‘pingpong ball’
太陽眼鏡	*tàiyáng yǎnjìng*	‘sunglasses’
飯碗	*fànwǎn*	‘rice bowl’

(20) **material**

大理石 地板	*dàlǐshídìbǎn*	‘marble (Dali rock) floor’
紙老虎	*zhǐlǎohǔ*	‘paper tiger’

Compounding

(21) **powered by**

電 燈	*diàn dēng*	‘electric light’
風 車	*fēng chē*	‘windmill (wind cart)’

(22) **organization**

大學 校長	*dàxué xiàozhăng*	‘university president’
北京 大學	*bĕijīng dàxué*	‘Beijing University’

(23) **coordinate**

爸爸 媽媽	*bàbà māmā*	‘father and mother’
風水	*fēng shuĭ*	‘fengshui (wind water)’
紙筆	*zhĭ bĭ*	‘paper and pen’
牛羊	*niú yáng*	‘livestock (cattle sheep)’

Root Compounding

(24)

螞蟻 *mǎyǐ* ‘ant’	蟻王 *yǐwáng* ‘queen ant’	工蟻 *gōngyǐ* ‘worker ant’
腦子 *nǎozi* ‘brain’	腦水腫 *nǎoshuǐzhǒng* ‘hydrocephaly’	後腦 *hòunǎo* ‘hindbrain’
蒼蠅 *cāngyíng* ‘fly’	蠅屍 *yíngshī* ‘fly corpse’	地中海蠅 *dìzhōnghǎiyíng* ‘Mediterranean fly’
蘑菇 *mógū* ‘mushroom’	菇傘 *gūsǎn* ‘pileus’	金菇 *jīngū* ‘golden mushroom’

(25) 蟻有蟻國,蜂有蜂國

yǐ yǒu yǐguó, fēng yǒu fēngguó

(ant have ant country, bee have bee country)

‘Ants have Antland, bees have Beeland’

Resultative Compounding

(26) 上 *shàng* ‘up’

下 *xià* ‘down’

進 *jìn* ‘enter’

出 *chū* ‘(go) out’

起來 *qǐlái* inchoative

回 *huí* ‘return’

過 *guò* ‘pass, across, beyond’

開 *kāi* ‘open’

完 *wán* ‘finish’

好 *hǎo* ‘good, complete’

死 *sǐ* ‘die’

見 *jiàn* ‘see, perceive’

破 *pò* ‘break’

清楚 *qīngchǔ* ‘clearly’

Resultative Compounding

(27)	**result**	打破	*dǎpò*	‘break by hitting’
		拉開	*lākāi*	‘open’
	achievement	寫清楚	*xiěqīngchǔ*	‘write clearly’
		買到	*mǎidào*	‘succeed in buying’
	direction	跳過去	*tiàoguòqù*	‘jump across’
		走進來	*zǒujìnlái*	‘come walking in’

(28)	跳得過去	*tiàodéguòqù*	‘able to jump across’
	跳不過去	*tiàobùguòqù*	‘not able to jump across’

Parallel Verbal Compounds

(29)	購買	*gòu-mǎi*	(buy-buy)	‘buy’
	建築	*jiàn-zhù*	(build-build)	‘build’
	檢查	*jiǎn-chá*	(examine-examine)	‘examine’
	治療	*zhì-liáo*	(treat-treat)	‘treat (a sickness)’

Subject-Verb Compounds

(30)	頭疼	*tóu-téng*	(head hurt)	‘(have a) headache’
	嘴硬	*zuǐ-yìng*	(mouth hard)	‘stubborn’
	眼紅	*yǎn-hóng*	(eye red)	‘covet’
	心酸	*xīn-suān*	(heart sour)	‘feel sad’
	命苦	*mìng-kǔ*	(life bitter)	‘tough straits’

(31) 我頭疼.
wǒ tóu téng
(I head hurt)
‘I have a head ache.’ (= ‘My head hurts.’)

(32) 你真嘴硬!
nǐ zhēn zuǐ-yìng
(you really mouth hard)
‘You are really stubborn!’ (≈ ‘Your mouth is really hard!’)

Verb-Object Compounds

(33)

出版	*chū-bǎn*	(emit edition)	‘publish’
睡覺	*shuì-jiào*	(sleep sleep)	‘sleep’
畢業	*bì-yè*	(finish course of study)	‘graduate’
開刀	*kāi-dāo*	(operate knife)	‘operate’
開玩笑	*kāi-wánxiào*	(operate joke)	‘make fun of’
照像	*zhào-xiàng*	(shine image)	‘take a photo’

Verb-Object Compounds

(34) 他們出版了那本書

tāmen chūbǎn-le nèiběn shū

(they publish-PERF that-CL book)

‘They published that book.’

(35) *他們開刀心臟

tāmen kāidāo xīnzàng

(they operate heart)

‘They are operating on the heart’

Verb-Object Compounds

(36) 我睡了覺

wǒ shuì-le jiào

(I sleep-PERF sleep)

‘I fell asleep’

(37) 開一個玩笑

kāi yīge wánxiào

(operate one-CL joke)

‘make fun of’

照兩張像

zhào liǎngzhāng xiàng

(shine two-CL photo)

‘take two photos’

Verb-Object Compounds

(38) 我睡了一個小覺

wǒ shuì-le yīge xiǎo jiào

(I sleep-PERF one-CL little sleep)

‘I took a little nap.’

(39) 開他的玩笑

kāi tāde wánxiào

(operate his joke)

‘make fun of him’

(40) 一個便我都沒有大

yīge biàn wǒ dōu méiyǒu dà

(one-CL convenience i all NEG have big)

‘I haven’t defecated at all.’

(cf. 大便 *dàbiàn* (big convenience) ‘defecate’)

(41) 你開甚麼刀?

nǐ kāi shénmo dāo

(you operate what knife)

‘What kind of operation are you having?’

Personal Names.

(42)	1	word	⇒	name
	2	name	⇒	1-char-family 2-char-given
	3	name	⇒	1-char-family 1-char-given
	4	name	⇒	2-char-family 2-char-given
	5	name	⇒	2-char-family 1-char-given
	6	1-char-family	⇒	char$_i$
	7	2-char-family	⇒	char$_i$ char$_j$
	8	1-char-given	⇒	char$_i$
	9	2-char-given	⇒	char$_i$ char$_j$

(Would need to expand this to cover “double-barreled” family names, which are still commonly used to refer to married women: thus 張王金蘭 *zhāng wáng jīnlán* ‘Mrs. Jinlan (Wang) Zhang’)

Abbreviations.

(43) 亞洲乒乓球聯盟 亞乒聯

yǎzhōu pīngpāng qiú liánméng *yǎ pīng lián*

Asian Ping-Pong Association

(44) 工業研究院 工研院

gōngyè yánjiù yuàn *gōng yán yuàn*

Industrial Research Center

(45) 以色列巴勒斯坦和談 以巴和談

yǐsèliè bālèsīdàn hètán *yǐ bā hètán*

Israel-Palestinian discussions

Abbreviations.

(46)	台灣香港	台港
	táiwān xiānggǎng	*tái gǎng*
	Taiwan-Hong Kong	
	中國石油	中油
	zhōngguó shíyóu	*zhōng yóu*
	China Oil	
(47)	上海杭州鐵路	滬杭鐵路
	shànghǎi hángzhōu tiělù	*hù háng tiělù*
	Shanghai-Hangzhou Railway	

Segmentation Standards

- Deciding what is a word in Chinese has practical consequences for a wide variety of applications both “sociological” and technological.

- The sociological applications arose early in the twentieth century in the context of various attempts to romanize Chinese orthography (see Pan, Yip and Han 1993):
  - 國語羅馬字運動 *guǒyǔ luómǎzì yùndòng* ‘Mandarin Romanization Movement’ (Y. R. Chao was involved in this)
  - 拉丁化運動 *lādīnghuà yùndòng* ‘Latinization Movement’

- Technological applications include IR, TTS, and ASR.

Proposed Standards

- Mainland Standard (GB 1993).

- ROCLING Standard (Huang et al 1997).

- University of Pennsylvania/Chinese Treebank (Xia 1999).

Mainland Standard

Lots of specific rules or principles, not all of them explained.

- Nouns derived in 家 *jiā*, 手 *shǒu*, 性 *xìng*, 員 *yuán*, 子 *zi*, 化 *huà*, 長 *zhǎng*, 頭 *tóu*, 者 *zhe* are segmented as one word: 科學家 *kēxuéjiā* ‘scientist’

- This seems reasonable enough, but for some reason 們 *men* is segmented separately, except in 人們 *rénmen* ‘people’, and words with *erhua*: 哥儿們 *gērmen* ‘brothers’.

- Personal names are also split: 毛 澤東 ‘Mao Zedong’.

- AABB reduplicants are one word: 高高興興 *gāogāoxìngxìng* ‘happy’. On the other hand ABAB reduplicants are two words: 雪白 雪白 *xuěbáixuěbái* ‘snow white’.

ROCLING Standard: I

The ROCLING standard seeks a standard that is “linguistically felicitous, computationally feasible, and must ensure data uniformity.” More of a meta-standard than a standard

- A string whose meaning cannot be derived by the sum of its components should be treated as a segmentation unit.
- A string whose structural composition is not determined by the grammatical requirements of its components,or a string which has a grammatical category other than the one predicted by its structural composition should be treated as a segmentation unit.

More specific guidelines:

- Bound morphemes should be attached to neighboring words to form a segmentation unit when possible.

ROCLING Standard: II

- A string of characters that has a high frequency in the language or high cooccurrence frequency among the components should be treated a sa segmentation unit when possible.
- Strings separated by overt segmentation markers should be segmented.
- Strings with complex internal structures should be segmented when possible.

A reference lexicon, based on a reference corpus is assumed.

U Penn Standard

The University of Pennsylvania standard is more akin in spirit to the Mainland standard in that it has less philosophy than the ROCLING standard, and more detailed rules.

- Sensitivity to phonological criteria: 室内 *shìnèi* ‘in the room’ is considered to be a single word, 中午 以後 *zhōngwǔ yǐhòu* ‘after noon’ is considered to be two words.

- Internal structure is marked: 打得破 *dǎ-dé-pò* ‘able to break’ is tagged as [打/V 得/DER 破/V]/V

Comparison of Standards

Form	UPenn	Mainland	ROCLING	Example
ABAB	ABAB	AB-AB	ABAB	研究研究 ‘research (a bit)’
AA-看	[AA/V kàn/V]/V	AA kàn	AA kàn	說說看 ‘talk about it and see’
Pers. Names	One Seg	Two Segs	One Seg	史立璿 Shi Lixuan
Noun + 們	One Seg	Two Segs	Two Segs	朋友們 ‘friends’
Ordinals	One Seg	Two Segs	Two Segs	第一 ‘first’

Table 2: Some differences between the segmentation standards, from (Xia 1999).

Segmented Corpora

- ROCLING Standard: 5 million words

- Penn Standard (Penn Chinese Treebank): 100K Words

Preview: What Corpus-Based Methods can do for us.

- Statistical basis for claims about wordhood.

- Measures of morphological productivity.

- Practical applications in various natural-language processing tasks.

Readings for Lecture 1

“Required” readings for this course are attached at the end of the notes. Optional readings are on reserve in the library.

- Packard (2000): Chapters 2–3.
- Morphology chapter from Li and Thompson 1981 (*on reserve*)

For those who can read Chinese, we can also recommend:

- Tang (1989), chapters 1, 3. (*on reserve*)

Finally, please take a look at the segmentation standards (*on reserve*)

- The Mainland standard (GB 1993)
- The ROCLING Standard (Huang et al 1997)
- The UPenn/Chinese Treebank Standard (Xia 1999)

There is no lab assignment for this section.