What makes a good Chinese word?

A novel test of morphological productivity

James Myers
National Chung Cheng University
Lngmyers *at*
ccu.edu.tw
http://www.ccunix.ccu.edu.tw/~lngmyers/

EACL 9, 24 September, 2015, Stuttgart

Thanks!

- **Project codirector**
  Jane Tsay
- **Assistants**
  Chen Tsung-Ying, Ruan Jia-Cing, Chang Yu-chu, Jiang Guan-nan, Lo Chia-Wen, Yang Chen-Tsung, Hong Guo-Ming, Hsu Zi-Ping, Huang Yi-Ching, Du Pei-Fen, Chuang Wei-Chiang, Kuo Lin-Ju, Chen Yang-lien, Zhong Yu-Cheng, Liu Yu-Jay, Ko Yu-Guang, Wu Yi-Hsin
- **Grants**
  NSC 100-2410-H-194-109-MY3, NSC 101-2410-H-194-115-MY3, MOST 103-2410-H-194-119-MY3

Overview

- Estimating productivity from corpora
  – Some underused methods
- Estimating productivity from judgments
  – A brand-new (?) method

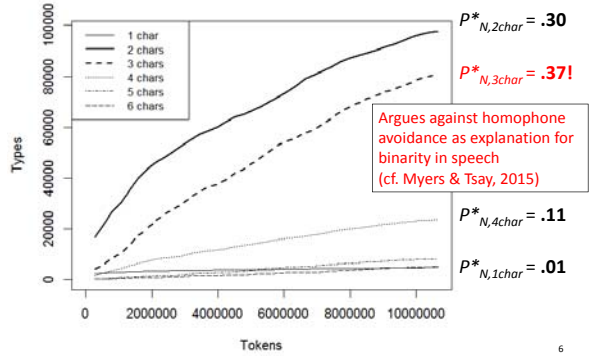
Mandarin corpus

- Academia Sinica Balanced Corpus (Huang et al., 1997)
  – (Taiwan) Mandarin (a "balanced" selection of texts)
- Around 10 million written word tokens
  – Segmented into words (compounds, affixed forms)
  – Tagged for word-level part of speech (POS)
- Transcribed in traditional characters
  ≈ morphemes, except for classic problems like 葡萄, 東西, 快樂 vs. 快速, ...

Corpus-based productivity estimates

- Productivity as coinage rate (Baayen & Renouf, 1996)
  – Hapax legomena (words that appear only once in a corpus) may include novel coinages
  – $P^*_{N,c}$: Productivity ***P*** of word class ***c*** as number of hapax legomena of class ***c***, proportional to all hapax legomena in corpus of ***N*** tokens
- Relative productivity can also be visualized as the slope of growth curves
  – Types vs. tokens as more of corpus is sampled
  – Steep slopes suggest word types still being added

Productivity in written Mandarin



$P^*_{N,2char}$ = **.30**

$P^*_{N,3char}$ = **.37!**

Argues against homophone avoidance as explanation for binarity in speech (cf. Myers & Tsay, 2015)

$P^*_{N,4char}$ = **.11**

$P^*_{N,1char}$ = **.01**

Productivity and part of speech

- Compound word components also have POS
  NN: 書店; VN: 飛機; VV: 尋找; NV: 頭痛; etc...
- Component POS can be tricky (Packard, 2000)
  verb-noun: **書**筆 vs. noun-**noun**: 國**書**
- How we defined component POS
  – Find one-character words in corpus
  – Look at their whole-word POS tags
  – Choose highest-frequency POS tag
- Caveats galore...

POS in Sinica corpus: Nouns

Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
Nb	Nba, Nbc	/*專有名稱*/
Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
Ncd	Ncda, Ncdb	/*位置詞*/
Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
Neu	*Neu*	/*數詞定詞*/.
Nes	*Nes*	/*特指定詞*/
Nep	*Nep*	/*指代定詞*/
Neqa	*Neqa*	/*數量定詞*/
Neqb	*Neqb*	/*後置數量定詞*/
Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
Ng	Ng	/*後置詞*/
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/

POS in Sinica corpus: Verbs

VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
VAC	VA2	/*動作使動動詞*/
VB	VB11,12,VB2	/*動作類及物動詞*/
VC	VC2, VC31,32,33	/*動作及物動詞*/
VCL	VC1	/*動作接地方賓語動詞*/
VD	VD1, VD2	/*雙賓動詞*/
VE	VE11, VE12, VE2	/*動作句賓動詞*/
VF	VF1, VF2	/*動作謂賓動詞*/
VG	VG1, VG2	/*分類動詞*/
VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
VHC	VH16, VH22	/*狀態使動動詞/
VI	VI1,2,3	/*狀態類及物動詞*/
VJ	VJ1,2,3	/*狀態及物動詞*/
VK	VK1,2	/*狀態句賓動詞*/
VL	VL1,2,3,4	/*狀態謂賓動詞*/
V_2	V_2	/*有*/

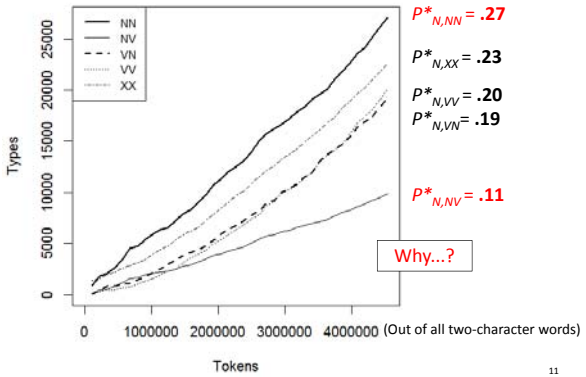
stative*
(≈ predicate adjective)

*Cf. A /*非謂形容詞*/ (non-predicate adjectives)

Examples of component POS

Word	Freq	POS	NumPOS1	NumPOS2	MaxPOS1	MaxPOS2	NounVerb
身法	17	Na	2	4	Na	Na	NN
律己	6	VA	1	3	Na	Nh	NN
窒悶	4	VH	1	2	VH	VH	VV
燭照	1	VC	1	5	Na	P	XX
老套	11	Na	8	3	VH	Nf	VN
默鐸	1	Nb	1	0	VH		XX
醫書	7	Na	3	2	Na	Na	NN
配合	2076	VC	5	8	VC	VJ	VV
丹青	7	Na	3	4	Na	VH	NV
自承	10	VE	5	3	P	P	XX
榜示	1	VE	2	1	Na	VE	NV

Productivity and POS



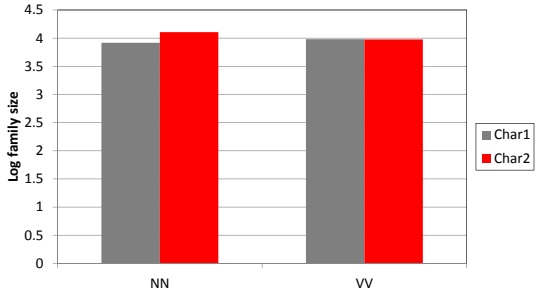
The role of headedness

- Headedness depends on whole-word POS (Packard, 2000)
- Nominal compounds tend to be right-headed
  - NN: 書店; VN: 飛機 (rare: NN: 父母)
- Verbal compounds more likely to be left-headed (or even coordinative)
  - VN: 開刀 (also common: VV: 尋找)
- This makes NV compounds anomalous
  - NV: 頭痛 (rare type)

Another factor

- Morphological family size (Schreuder & Baayen, 1997)
  - Number of words with a morpheme in some position
    書包: 書店、書桌、 ...; 書包: 皮包、肉包、...
- May also say something about productivity
  - Characters with a large family size are readily used to create new words
  - Word position and POS may also matter
- Technical note:
  - Logarithm used to make distribution more normal

Family sizes in NN vs. VV words



NN compounds right-headed

VV compounds more coordinative

Native-speaker judgments

- What makes a good Chinese word?
  - Why not ask native speakers?
  - This is what syntacticians do all the time....
- Wordlikeness judgments
  - Give people fake words, and ask if they're wordlike
  - For our previous work doing this, see Myers (2015)
- Caveat: What is a "fake word" in Mandarin?
  - Is 小房 a fake word or a real phrase?

Design and procedure

- Megastudy (Balota et al., 2012)
  - Huge random sample, explore results like a corpus
- 51 participants
  - Native speakers of (Taiwan) Mandarin
- Binary judgments: 像國語 vs. 不像國語
- 3000 two-character nonword items*
  - Real characters combined randomly in proportion to frequencies in real two-character words
  - Ruled out items with any special uses on the web
  - Included affixes like 們、子

*We accidentally repeated 6 items, so only 2994 distinct test items

Fake word component POS

- We determined component POS as before
  - No way to determine whole-"word" POS
- Sample items:
  - NN: 信同、子技、臺壤、度工、期聯
  - NV: 奧知、天雜、態建、人任、象學
  - VN: 疑程、遣盧、察歐、攻草、行系
  - VV: 低有、估療、限劃、委廣、似生
  - XX: 以限、免所、問如、局來、究日

Results: Best items

Item	Rating	Item	Rating	Item	Rating	Item	Rating	Item	Rating
長論	0.84	聽示	0.67	歡情	0.62	成學	0.59	立岸	0.57
小房	0.75	政才	0.65	示學	0.61	取理	0.59	回生	0.57
傳經	0.73	報實	0.65	求協	0.61	強用	0.59	長兒	0.57
播佈	0.72	世變	0.63	居之	0.61	處政	0.59	婚意	0.57
悔然	0.71	務時	0.63	師制	0.61	會制	0.59	結度	0.57
實論	0.71	諸年	0.63	處示	0.61	誘視	0.59	開民	0.57
示婚	0.69	適業	0.63	陸遊	0.61	應理	0.59	飯器	0.57
展台	0.69	正映	0.62	報政	0.61	總件	0.59	適務	0.57
受功	0.67	法勢	0.62	檢諮	0.61	避陣	0.59	思題	0.56
容理	0.67	前意	0.62	通戶	0.6	民傷	0.58	處家	0.56
眾謠	0.67	間樓	0.62	濫民	0.59	功教	0.57	富政	0.56

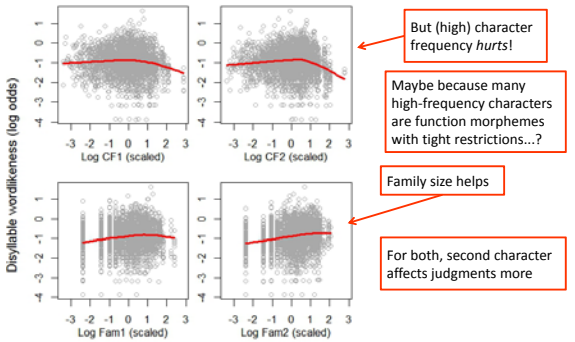
Worst items

Item	Rating	Item	Rating	Item	Rating	Item	Rating	Item	Rating
們大	0.08	慢媽	0.08	內較	0.06	第體	0.06	所什	0.04
們心	0.08	麼忙	0.08	她二	0.06	軟這	0.06	很於	0.04
們有	0.08	麼除	0.08	們問	0.06	就優	0.06	個易	0.04
們極	0.08	麼路	0.08	麼裂	0.06	達麼	0.06	們定	0.04
展但	0.08	影不	0.08	什比	0.06	網呀	0.06	夠們	0.04
時麼	0.08	導不	0.08	文麼	0.06	麼成	0.06	這化	0.04
場是	0.08	機十	0.08	吐們	0.06	麼例	0.06	開每	0.04
媽發	0.08	辦麼	0.08	股麼	0.06	麼根	0.04	陽不	0.04
溪不	0.08	體其	0.08	為特	0.06	水一	0.04	園這	0.04
裡好	0.08	改怕	0.06	們尚	0.06	且西	0.04	麼設	0.04
過爸	0.08	二在	0.06	級看	0.06	些他	0.04	否體	0.02

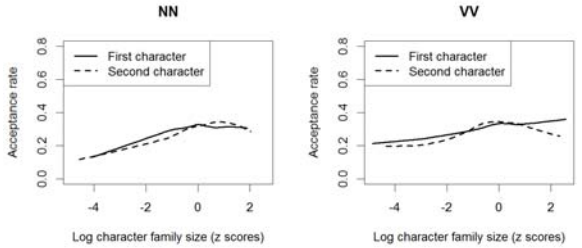
Modeling these judgments

- Dependent variable
  - Response choice ("like" vs. "unlike" Mandarin)
- Some of the independent variables
  - Character frequency
  - Family size
  - Part of speech
- Mixed-effects logistic regression
  - Both participants and items as random variables

Character frequency and family size



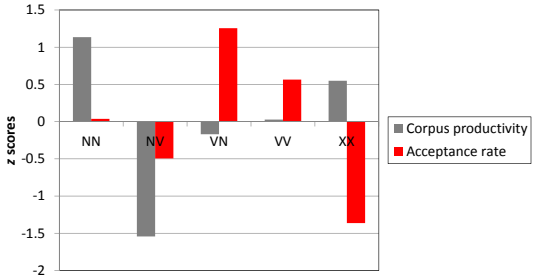
Family size x part of speech



2nd character has stronger effect than 1st character

Both effects weaker than in NN, 1st character has stronger effect

Part of speech: Corpus ≠ judgments



Acceptance rates cannot be derived from corpus-based $P^*_{N,c}$

Conclusions

- Word size (corpus)
  - In writing, two-character words are most *common*, but three-character words are most *productive*
- Headedness (corpus = judgments)
  - NN: More influenced by right "head" family size
  - VV: Family size influences are more balanced
- POS preferences (corpus ≠ judgments)
  - Corpus: **NN** > XX > VV > **VN** > NV
  - Judgments: **VN** > VV > **NN** > NV > XX
- Somebody should try to figure this out....

Your turn!

http://Lngproc.ccu.edu.tw/MWP/

Main search interface programmed by Ruan Jia-Cing

References (1/2)

Baayen, R.H., & Renouf, A. (1996). Chronicling the *Times*: Productive lexical innovations in an English newspaper. *Language, 72*, 69-96.

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell me about lexical processing? In J. S. Adelman (Ed.) *Visual word recognition 1: Models and methods, orthography and phonology* (pp. 90-115). Hove, England: Psychology Press.

Huang, C.-R., Chen, K.-J., Chen, F.-Y., & Chang, L.-L. (1997). Segmentation standard for Chinese natural language processing. *Computational Linguistics and Chinese Language Processing, 2* (2), 47-62.

References (2/2)

Myers, J. (2015). Markedness and lexical typicality in Mandarin acceptability judgments. *Language & Linguistics, 6*.

Myers, J., & Tsay, J. (2015). Trochaic feet in spontaneous spoken Southern Min. *Proceedings of the 27th North American Conference on Chinese Linguistics*.

Packard, J. L. (2000). *The morphology of Chinese*. Cambridge University Press.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language, 37*(1), 118-139.