Chinese character form and the mathematics of productivity

James Myers
National Chung Cheng University
Taiwan
Lngmyers@ccu.edu.tw

International Conference on Asian Linguistics
Nguyen Tat Thanh University - Ho Chi Minh City - December 15-16, 2016

Thanks!

- Grants from Taiwan's National Science Council / Ministry of Science and Technology
  - NSC 97-2410-H-194-067-MY3, NSC101-2410-H-194-115-MY3, MOST 103-2410-H-194-119-MY3
- Lab assistants
  - Ko Yuguang, Yang Chentsung, Hsu Chiung-Wen, Hong Guo-Ming, Hsu Zi-Ping, Du Pei-Fen, Su Yu-Ting, Liu Yu-Jay, Chuang Wei-Chiang, Hsieh Yu-Yi, Chen Tsung-Ying, Pan Hsiao-Yin

Overview

1. The "grammar" of Chinese characters
   - Regularities in radical positions
2. Evidence for radical position productivity
   - Experiment: Novel character acceptance
   - Corpus: Novel character coinage
3. Triggers for radical position productivity
   - A simple mathematical model of rule learnability

1. "Grammar" beyond speech

- Structural regularities in learned systems
  - Sign language (Sandler & Lillo-Martin, 2006)
  - Music (Lerdahl & Jackendoff, 1983)
  - Comics (Cohn et al., 2012)
  - ...
- Is this "real" grammar?
  - Is it mentally active?
  - If so, is it learned in ways similar to speech?
  - Does it use the same mental devices as speech?

Chinese character "grammar"

- An old yet controversial idea
  - E.g., Rankin (1965), Wang (1983), Stalph (1989), Sproat (2000), Hsieh (2006), Kordek (2013), Ladd (2014), Myers (2016), ...
- My own radical position
  - Character form regularities are mentally active
  - Learned from the statistics of characters
  - But learning uses statistics in a biased way
  - Hence character "rules" are abstract

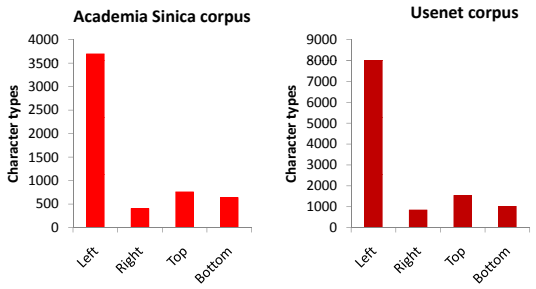
Radical position patterns

- Semantic radicals
  - Closed-class components hinting at meaning
  - Akin to "affixes"...?
- They tend to appear in consistent positions
  - Left: 詞 *cí* "word" (cf. 言 *yán* "speech")
  - Right: 鵝 *é* "goose" (cf. 鳥 *niǎo* "bird")
  - Top: 花 *huā* "flower" (cf. 艸 *cǎo* "grass")
  - Bottom: 盒 *hé* "box" (cf. 皿 *mǐn* "dish")
  - ...

My sources for traditional characters

- Academia Sinica Balanced Corpus of Modern Chinese (Chen et al., 1996)
  - From a variety of sources available in Taiwan
  - 6,608 character types (83% with L/R/T/B radicals)
  - 15,370,423 character tokens
- Usenet corpus (Tsai, 2006)
  - From 1993-1994 Usenet postings in Taiwan
  - 13,060 character types (87% with L/R/T/B radicals)
  - 171,894,734 character tokens

Radical position type frequencies

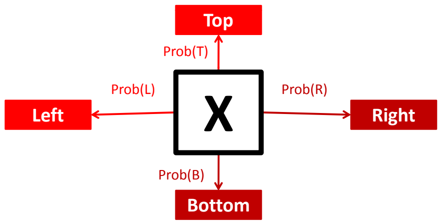


Position-dependent radical variation

- Radicals reduced in complexity at left and top
  心～忙　人～位　水～泊　手～拾
  艸～花　竹～箪
- Radicals not reduced at right or bottom
  忙～忘　泊～泉　拾～拿　加～功
- Exceptions...
  刀～刻　火～熱
- "Rules" addressed here:
  - "Small" radicals: left > right, top > bottom

2. Testing productivity

- Prob(Form|Context...)
  - Left vs. right, top vs. bottom "small" radical
  - I'll ignore semantic context, etc....



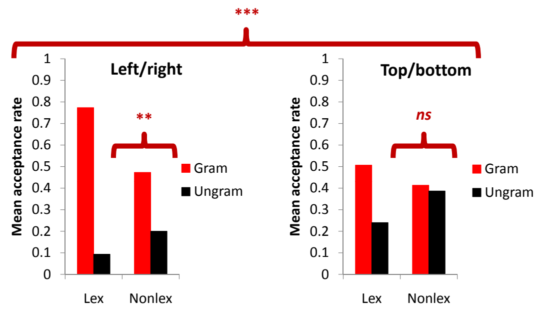
Fake character acceptability test

- Lexicality: Component is/isn't used as radical
- Grammaticality: Obey/violate patterns
  - Left (top) component is/isn't smaller
- Judge whether is/isn't like a character

	Lex		Nonlex	
	Gram	Ungram	Gram	Ungram
Left/Right	[illegible]	[illegible]	[illegible]	[illegible]
Top/Bottom	[illegible]	[illegible]	[illegible]	[illegible]

(Myers, 2011, 2012; see also Myers, 2016, for an experiment on a different pattern)

Only left/right rule generalizes



*** *p* < .001; ** *p* < .01; *ns* *p* > .7 in mixed-effects logistic regression models

Coining new characters

- Artistic examples from my office door
  - Guess which were invented by a Chinese writer!

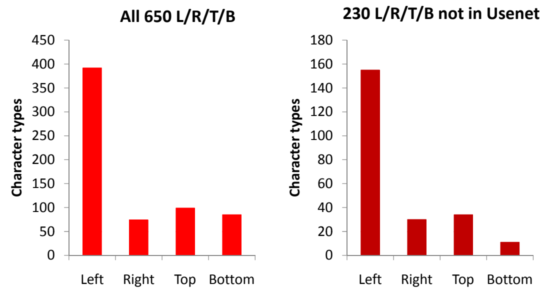
Natural character coinage

- Characters are not *quite* closed-class
  - Nonstandard variants are ancient
  - Standards in Japan, Korea, Hong Kong, the PRC
  - Chữ Nôm in pre-Romanization Vietnam
    (I'll return to this shortly)
- E.g. PRC characters obey positional allophony
  - Traditional 言: 警～詞 (no left reduction)
  - Simplified 言: 警～词 (left reduction!)

Chữ Nôm (𡨸喃)

- Source: www.chunom.org (Nguyen, 2015)
  - Incomplete and only coarse frequency estimates
  - Over 1000 characters/variants from classical texts
  - Over 400 not found in Usenet corpus
- Many apparent violations of radical patterns
  - Exceptions to left reduction: 手 [illegible] [illegible]
  - "Embedded" radicals: [illegible] [illegible] 𡨸
  - Chữ Nôm characters have a tendency to be formed by "compounding" instead of "affixation"

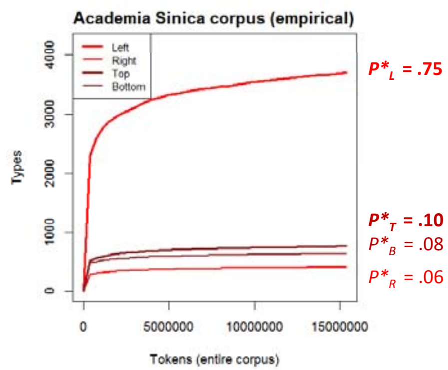
Radical positions in Chữ Nôm



Estimating productivity in AS corpus

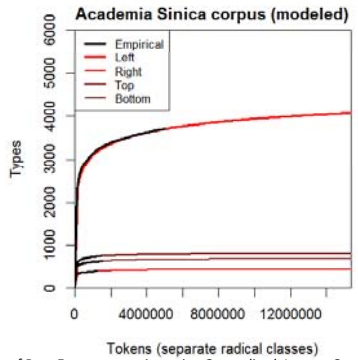
- Productivity as coinage rate (Baayen & Renouf, 1996)
  - **Hapax legomena** (words that appear only once in a corpus) potentially include novel coinages
  - $P^*_{N,c}$: Productivity ***P*** of word class ***c*** as number of hapax legomena of class ***c***, proportional to all hapax legomena in corpus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

Empirical growth curves (AS corpus)



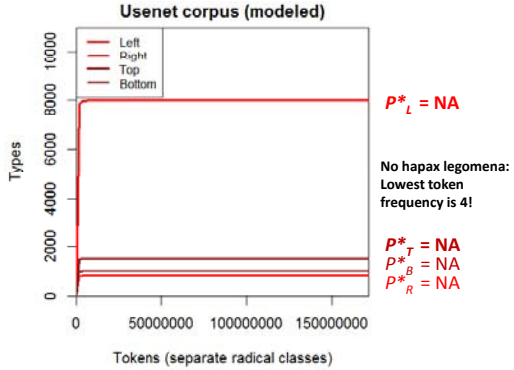
(Token order randomized to improve smoothness)

Modeled growth curves (AS corpus)



(Large Numbers of Rare Events regression using Generalized Inverse Gauss-Poisson modeling, which gave the best fit here: Baayen, 2001; Evert & Baroni, 2007)

Modeled growth curves (Usenet)



(Large Numbers of Rare Events modeling with Generalized Inverse Gauss-Poisson)

3. Triggering productivity

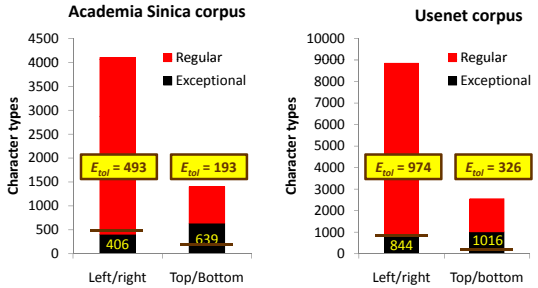
- Yang (2005, 2016): **The Tolerance Principle**
  - How many exceptions E_{tol} to a rule in a set of N word (item) types can a learner tolerate before the rule becomes unlearnable?
- Model assumes an Elsewhere principle
  - First search memory to see if item is an exception
  - If not, then apply the rule
  - Exception searching has a measurable cost
  - Too many exceptions mean rule gives no benefit

Yang's Tolerance Principle

$$E_{tol} \approx N / \ln N$$

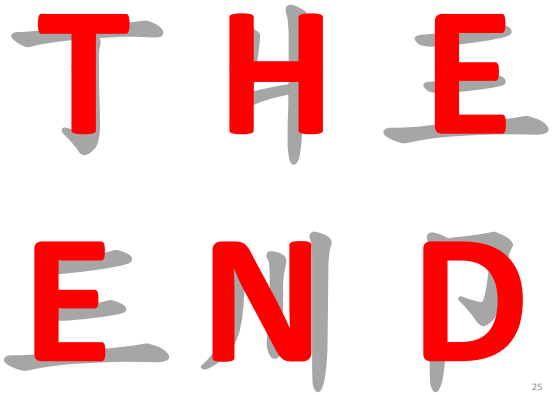
- Quick proof
  - Average steps to search all N items is N / ln N
  - If exceptions E exceed this cost, just list all N items
- See Yang's work for details
  - Copious empirical evidence from child language
  - Nonlinear learning: Smaller N permits larger E_{tol}
  - Slightly more detailed proof is given in Appendix

Implications for radical positions



Conclusions

- Left > right radical rule
  - Just barely passes Yang's Tolerance Principle
  - Readers generalize it to nonradicals
  - Huge L/R contrast in corpus-estimated coinage
- Top > bottom radical rule
  - Far from passing Yang's Tolerance Principle
  - Readers don't generalize it to nonradicals
  - No T/B contrast in corpus-estimated coinage



References (1/5)

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

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

Chen, K.-J., Huang, C.-R., Chang, L.-P. & Hsu, H.-L. (1996). Sinica Corpus: Design methodology for balanced corpora. *Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation*, Seoul, Korea, 167-176.

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology, 65* (1), 1-38.

References (2/5)

Evert, S. & Baroni, M. (2007). zipfR: Word frequency distributions in R. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, 29-32, Prague, Czech Republic.

Hsieh, S.-K. (2006). *Hanzi, concept and computation: A preliminary survey of Chinese characters as a knowledge resource in NLP*. University of Tübingen Ph.D. thesis.

Kordek, N. (2013). *On some quantitative aspects of the componential structure of Chinese characters*. Poznań, Poland: Wydawnictwo Rys.

Ladd, D. R. (2014). *Simultaneous structure in phonology*. Oxford, UK: Oxford University Press.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

References (3/5)

Myers, J. (2011, September). The psychological reality of formal regularities in Chinese characters. Presented at the 7th Conference of the European Association of Chinese Linguistics, Venice, Italy.

Myers, J. (2012, August). Levels of analysis in the generalization of Chinese character regularities. Presented at CogSci 2012: The 34th Annual Conference of the Cognitive Science Society, Sapporo, Japan.

Myers, J. (2016). Knowing Chinese character grammar. *Cognition, 147*, 127-132.

Nguyen, T. (2015). http://www.chunom.org. Accessed 2016/12/7.

References (4/5)

Rankin, B. K. (1965). *A linguistic study of the formation of Chinese characters*. University of Pennsylvania Ph.D. thesis.

Sandler, W., & Lillo-Martin, D. (2006). *Sign language and linguistic universals*. Cambridge, UK: Cambridge University Press.

Sproat, R. (2000). A computational theory of writing systems. Cambridge, UK: Cambridge University Press.

Stalph, J. (1989). *Grundlagen einer Grammatik der sinojapanischen Schrift* [Foundations of a grammar of the Sino-Japanese script]. Wiesbaden, Germany: Harrasowitz Verlag.

References (5/5)

Tsai, C.-H. (2006). Frequency and stroke counts of Chinese characters. http://technology.chtsai.org/charfreq/.

Wang, J. C.-S. (1983). *Toward a generative grammar of Chinese Character structure and stroke order*. University of Wisconsin-Madison, Ph.D. dissertation.

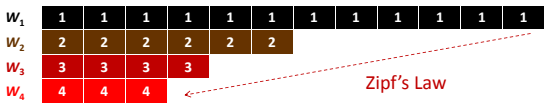
Yang, C. (2005). On productivity. *Linguistic Variation Yearbook, 5* (1), 265-302.

Yang, C. (2016). *The price of linguistic productivity: How children learn and break rules of language*. Cambridge, MA: MIT Press.

Appendix: Proving Tolerance (1/4)

Yang (2005, 2016):

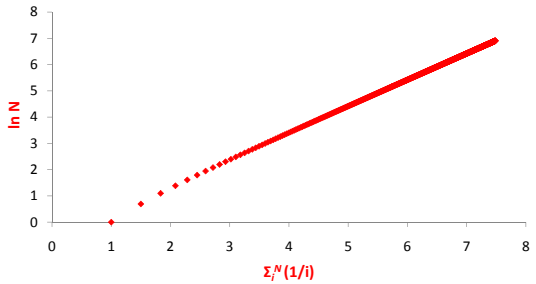
- **Learner hypothesis 1:** List all *N* words (items)
  - Assume frequency ordering and Zipf's Law
  - What is the average number of steps needed?



mean(1_1,...,1_{12}, 2_1,...,$2_{12/2}$, 3_1,...,$3_{12/3}$, 4_1,...,$4_{12/4}$) = $4 \times 12/(12 \times (1+1/2+1/3+1/4))$: $N/\Sigma_i^N(1/i)$

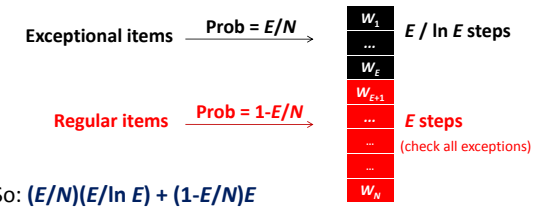
Appendix: Proving Tolerance (2/4)

$N/\Sigma_i^N(1/i) \approx N/\ln N$



Appendix: Proving Tolerance (3/4)

- **Learner hypothesis 2:** *E* exceptions, rest by rule
  - All items optimally sorted and obey Zipf's Law
  - What is the average number of steps needed?



So: $(E/N)(E/\ln E) + (1-E/N)E$

Appendix: Proving Tolerance (4/4)

- How many exceptions make Hyp 2 too costly?
  - Find *E* such that

    $N/\ln N$ = $(E/N)(E/\ln E) + (1-E/N)E$

    (no rule) (rule)
  - Set $x = E/N$, $f(x) = x(E/\ln E) + (1-x)E - N/\ln N$
  - To find *x* where $f(x) = 0$: Get derivative $f'(x) = 0$, exploit deriv(ln *x*) = 1/*x*, ignore small terms
- Resulting estimate: $E_{tol} \approx N / \ln N$
  - Same as number of steps for all-listing hypothesis