Productivity and Semantic Transparency

An exploration of compounding in Mandarin

Harald Baayen and Tian Shen (申甜)

Tübingen University and Shanghai Jiao Tong University

Perspectives on Productivity, May 26, 2021

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Productivity

- productivity is “the central mystery of word formation" Aronoff (1976)
- this is especially true for compounds:
  - compounding in English and Mandarin is extremely productive
  - but at the same time, the relation between form and meaning is not very regular at all compared to inflection and derivation in English
  - the Dutch structuralists who worked on productivity (Uhlenbeck, Schultink) argued that compounds do not constitute a ‘morphological category’, as there is no change in form that goes hand in hand with a change in meaning

- compounds typically serve word formation,
  creating names for things and events in the world

- things and events in the world are not necessarily compositional,
  and they can be described/referenced in many different ways

- compounds are 'labels' that are a priori unpredictable,
  but a-posteriori understandable (in culture and/or context);
  they often look remarkably like mnemonics

- fire engine : a truck that has equipment for putting out fires
- 火车 (*huo3che1, fire engine*) : a train

the riddle of compounding:
how can compounding be so unpredictable and yet so productive?

- in Mandarin, compounding is even more productive than in English

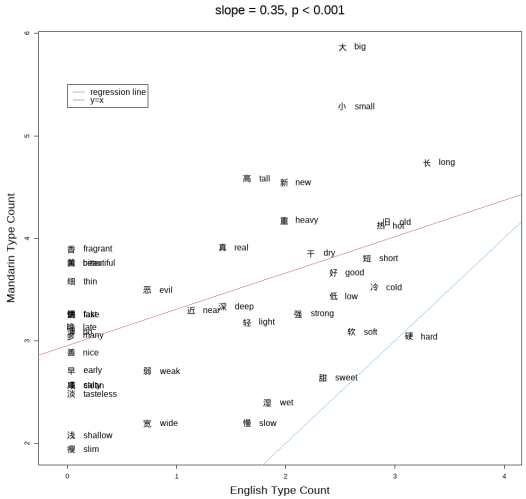
- in the following example, -er is a recurring exponent for English, but translation equivalents in Mandarin are often compounds, although a subject exponent also exists:
  - teacher 老师 (*lao3shi1*) old master
  - worker 工人 (*gong1ren2*) do person
  - worker bee 工蜂 (*gong1feng1*) work bee

    ...
  - dancer 舞者 (*wu3zhe3*) dance subject
  - writer 作者 (*zuo4zhe3*) do subject
  - the dead 死者 (*si3zhe3*) die subject

    ...

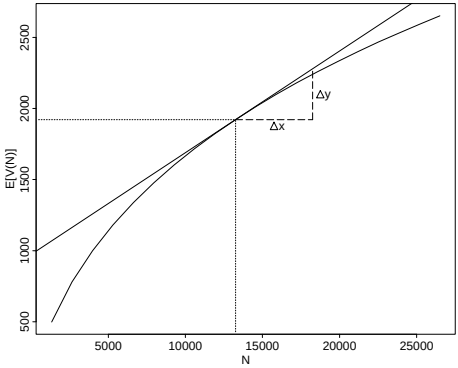
- constituent family: all compounds sharing the same constituent in the same position (e.g., 大 as first constituent)

- our hypothesis is that constituent families in Mandarin compounds are characterized by different degrees of productivity, and that their degree of productivity is positively correlated with their semantic transparency

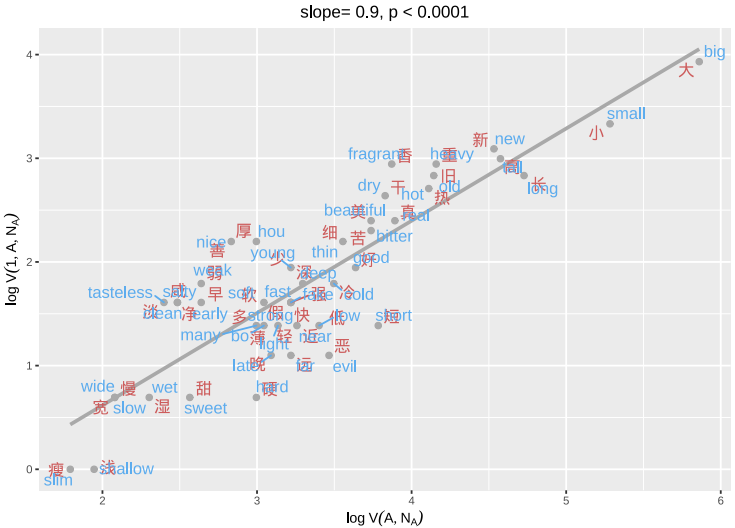


- 大家 big family, ‘everyone’
- 大会 big meeting, ‘conference’
- 大亨 big prosperity, ‘magnate’
- . . .

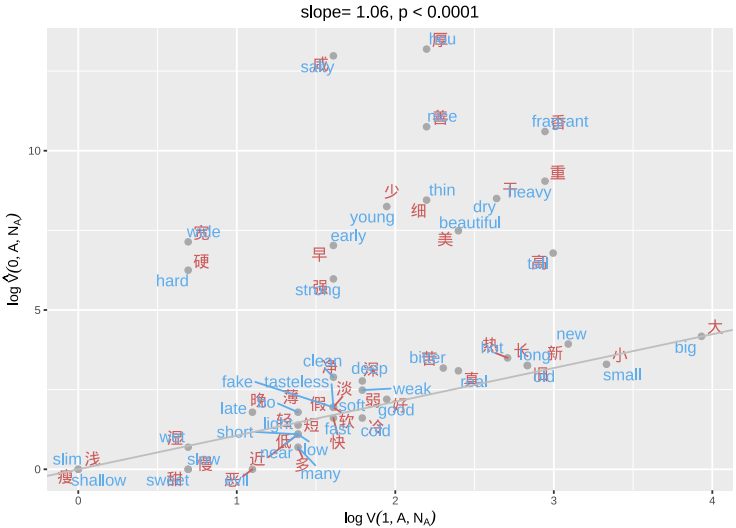
- 甜瓜 sweet melon, ‘sweet melon’
- 甜点 sweet dot, ‘dessert’
- 甜酒 sweet wine, ‘sweet wine’
- . . .



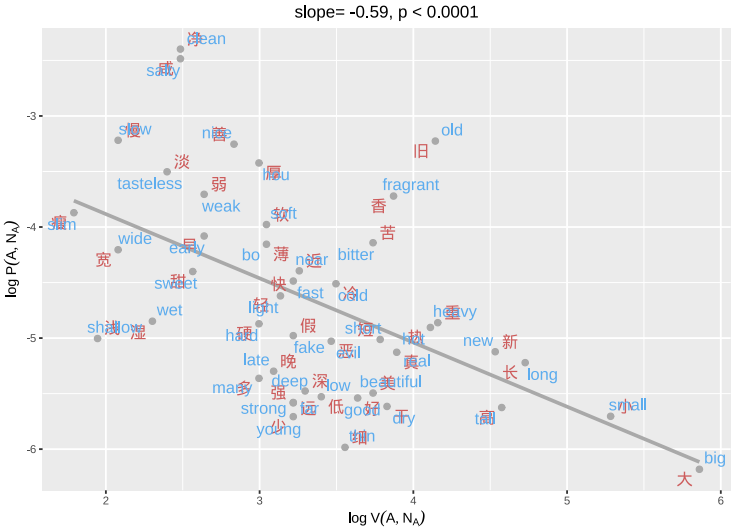
- growth curve of the vocabulary size V(N)
- the growth rate P(N) is given by slope of the tangent to the curve at token size N
- this slope is estimated by V(1)/N
- as $V(N)$ increases, $P(N)$ decreases



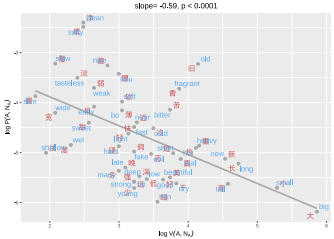
- V: extent of use (cf. Corbin's profitability)
- V(1): proportional to the growth rate of the total vocabulary

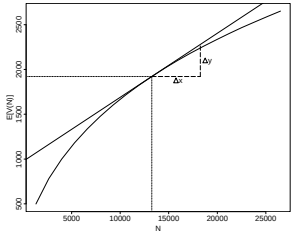


- using probability theory, we can estimate the number of unseen types $V(0)$
- estimates can be too high due to insufficient data
- a Gaussian location-scale model shows that V and V(0) are positively correlated



as expected given the mathematical properties of the growth curve of the vocabulary, the growth rate decreases with increasing V





- this negative correlation is a new empirical finding
- for English derivation, as studied by Baayen and Lieber (1991); Hay and Baayen (2002) no such negative correlation exists

why are V and P negatively correlated in Mandarin, but not in English?

- the adjectives in this study are semantically much more similar than the collection of English derivational affixes investigated by Baayen and Lieber (1991); Hay and Baayen (2002), which are semantically very diverse

- we are probably closer to matching morphological categories for semantics (cf. the work on rival affixes such as -ness and -ity, or un- and in- in English, which attempted to study productivity while controlling for semantics)

however, rival affixes can have rather different semantics (Riddle, 1984), and one could hypothesize that:

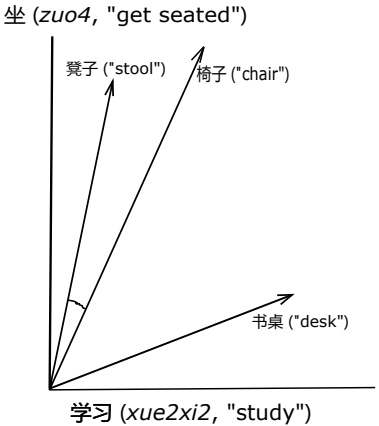
- semantic transparency -ness $>$ semantic transparency -ity
- productivity of -ness $>$ productivity of -ity

but for this to be convincing, we need to properly quantify the notion of semantic transparency

- distributional hypothesis : words with similar meanings tend to occur in similar distributions of contexts (Firth, 1957; Harris, 1954)
- word meanings are represented as high-dimensional vectors derived from the co-occurrence of words in diverse corpora
- semantic similarities or semantic relatedness of two words can be approximated with geometric methods by gauging the cosine of the angle formed by the semantic vectors of the two target words in semantic spaces

distributional semantics: many implementations available

- LSA (Landauer & Dumais, 1998)
- word2vec, fasttext, ...
- grounded vectors (Shahmohammadi et al., 2021)


坐 (*zuo4*, "get seated")
凳子 ("stool")
椅子 ("chair")
书桌 ("desk")
学习 (*xue2xi2*, "study")

- semantic vectors
  - Tencent AI Lab Embedding Corpus for Chinese Words and Phrases: 200-dimensional vector representations for 8 million Chinese words and phrases (Song et al., 2018)
  - semantic vectors : 56 adjectives, 751 nouns, and corresponding 1482 adjective-noun compounds

there are three semantic comparisons for a given adjective-noun compound:

- adjective and the compound ($r_{\text{A-AN}}$)
- adjective and the noun ($r_{\text{A-N}}$)
- noun and the compound ($r_{\text{N-AN}}$)

in addition, we can ask:

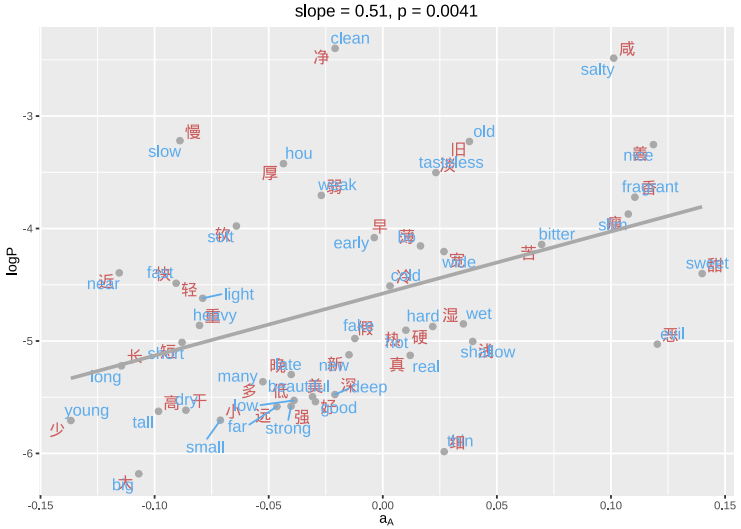
- how similar are the AN compounds to each other ($\bar{r}_{AN}$)

predictions:

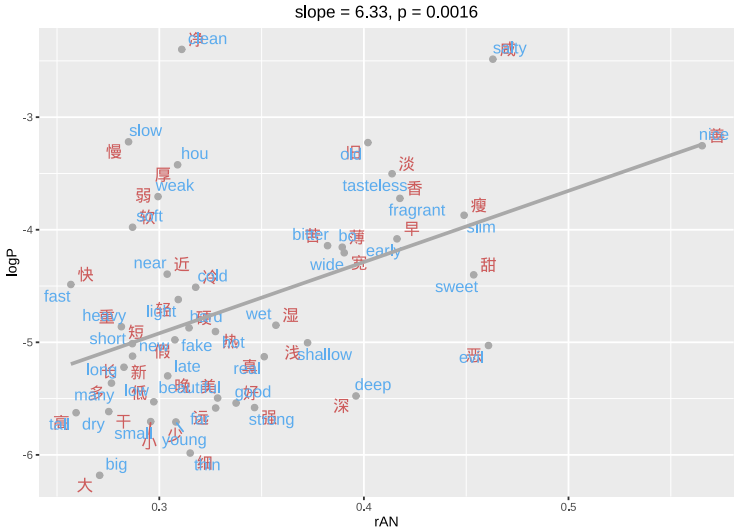
- greater transparency should predict greater degree of productivity P

  why? because P is conditioned on the morphological category itself, and hence is potentially sensitive to the dynamics of the form-meaning systematicities in the morphological category

- of the four measures, the best candidates are the ones that zoom in on the morphological category (the compound similarity measure, and the adjective-compound similarity measure)



- a_A: mean semantic transparency for a constituent family
- we observe the expected positive correlation



- r_{AN}: mean of pairwise semantic similarity of the compounds in the constituent family
- again, we observe a positive correlation, an expected

- semantic transparency correlates with P
- semantic transparency does not correlate with V or V(1)

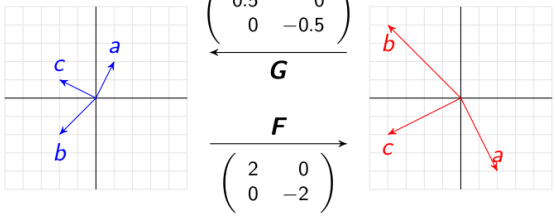
thus, the conditional probability that a word token represents a new type, given that it belongs to the morphological category, correlates with the semantic transparency of the adjectives and the compounds on the one hand, and the compounds between themselves on the other hand

- target constituents : antonymous verbs
  - Action verbs: 问 ('ask') vs. 答 ('answer')
  - Motion verbs: 来 ('come') vs. 去 ('go')
  - psychological verbs: 爱 ('love') vs. 恨 ('hate')
- no restrictions on the output category
- results:
  - a positive correlation between semantic similarity $r_{\text{V-VN}}$ and morphological productivity $\mathcal{P}$ (V, VN) ($p = 0.02$).
  - a positive correlation between semantic coherence $\bar{r}_{VN}$ and morphological productivity $\mathcal{P}$ (V, VN) ($p = 0.03$).

- target constituents : near-synonyms of 人 ('man')
  - 农 ('peasant')
  - 匠 ('craftsman')
  - 鬼 ('ghost')
  - 迷 ('addict')
- no restrictions on the output category
- results:
  - No significant correlation between semantic similarity $r_{\text{N-XN}}$ and degree of productivity $\mathcal{P}$ (N, XN) ($p = 0.36$).
  - a positive correlation between semantic coherence $\bar{r}_{XN}$ and degree of productivity $\mathcal{P}$ (N, XN) ($p = 0.009$).

- for case study 1 (AN compounds), we had the strongest constraints in place, creating a semantically relatively homogeneous group with gradable adjectives
- by contrast, for case studies 2 and 3, selection criteria were less stringent, and results are somewhat less robust
- it seems to us that careful matching for general semantics (the idea that also motivated investigating rival affixes) is important, not only for assessment of degrees of productivity, but specifically for the study of the relation between productivity and semantic transparency

$$\boldsymbol{C} = \begin{pmatrix} 1 & 2 \\ -2 & -2 \\ -2 & 1 \end{pmatrix} \qquad \boldsymbol{S} = \begin{pmatrix} 2 & -4 \\ -4 & 4 \\ -4 & -2 \end{pmatrix}$$



Baayen, R. H., Chuang, Y. Y., Shafaei-Bajestan, E., and Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. Complexity, 2019, 1-39.

- form vectors were phone-based
- word2vec word embeddings
- comprehension task: predict meaning from form
  accuracy 95.4%
- for each adjective, we calculated the number of erroneous recognitions and the number of correct recognitions, and ran a logistic model with our transparency measures as predictors

- we observed a positive correlation between semantic transparency and model accuracy
  - for adjective-compound similarity ($r_{\text{A-AN}}$): $p = 0.001$
  - for pairwise compound similarity ($\bar{r}_{AN}$): $p = 0.0015$
- we also observed a positive correlation between model accuracy (for predicting the correct semantic vectors) and $\mathcal{P}(A, N_A)$: $p = 0.0001$

Conclusions

- distributional semantics makes it possible to quantify semantic transparency, and correlate this with productivity measures
- we have been able to show, to our knowledge for the first time, that productivity and semantic transparency go hand in hand
- with the new tools, we can now start probing the productivity of compounding through compound constituent families
- computational modeling with LDL clarifies that the relation between form and meaning in compounds is better learnable when semantic transparency is greater

for further details, see Shen, T., and Baayen, R. H. (2021). Adjective-Noun Compounds in Mandarin: A Study on Productivity. Corpus Linguistics and Linguistic Theory, https://www.degruyter.com/document/doi/10.1515/cllt-2020-0059/html

thank you for your attention

谢谢!

Baayen, R. H. and Lieber, R. (1991). Productivity and English derivation: a corpus-based study. *Linguistics*, 29:801–843.

Firth, J. R. (1957). *Studies in Linguistic Analysis*. Wiley-Blackwell.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hay, J. B. and Baayen, R. H. (2002). Parsing and productivity. In Booij, G. and Van Marle, J., editors, *Yearbook of Morphology 2001*, pages 203–235. Kluwer Academic Publishers, Dordrecht.

Riddle, E. A. (1984). A historical perspective on the productivity of the suffixes *-ness* and *-ity*. In *Conference on Historical Semantics and Word Formation*, pages 28–31, Błazejevko, Poland.

Song, Y., Shi, S., Li, J., and Zhang, H. (2018). Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180.