Supplementary material for Historical semantic chaining and efficient communication: The case of container names

Yang Xu[*], Terry Regier[*], and Barbara C. Malt[†]

[*]Department of Linguistics, University of California, Berkeley
[†]Department of Psychology, Lehigh University

1. Time stamping

Procedure

For each container item in the dataset, we determined the modal modifier- and head-noun phrase for that item (e.g. *peanut jar*) from the naming data, and performed a corpus search for that phrase in a large historical corpus, the Google Ngram American English corpus (Michel et al., 2011), over the period 1800-2000. Table S1 provides the phrases for all 60 container stimuli. We used English corpus data for analyzing all three languages because historical data for Spanish and Chinese are more sparse and hence are not on an equal footing with English. In particular, they do not support many phrase-level queries in the stimulus set that are relevant to our analysis. For example, simple phrases such as "medicine bottle" ("yao4 ping2" in Mandarin Pinyin),"soy sauce bottle" ("jiang4 you2 ping2"), and other similar phrases rendered no successful retrievals from the Google Ngram database. Although object types may not have entered Chinese or Argentinean culture at exactly the same dates as they entered the U.S., the rank order is likely to be similar, e.g. aspirin bottles cannot pre-date the availability of aspirin as a medication, and children's juice boxes would be a recent entry for all cultures. We recorded the frequency of use of that phrase for each year. For each container phrase, we then applied the change-point detection algorithm described below to these historical frequency traces to determine the year in which each phrase experienced a substantial rise in frequency from a baseline of zero. We took that year to be the date of emergence of that object. Figure S1 illustrates the points of emergence for two example container items.

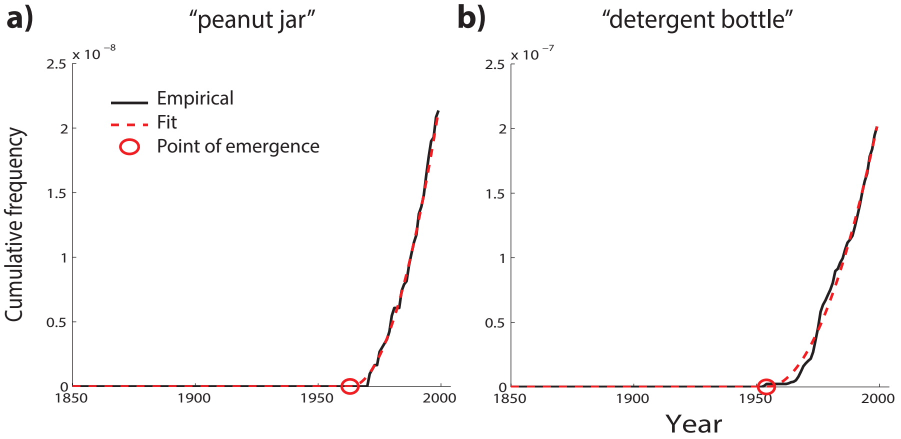

Figure S1: Illustration of the time-stamping procedure. a) Time-stamping the phrase "peanut jar". The change-point algorithm (Kass et al., 2014) fits the cumulative frequency of the given phrase and then finds a change point, which is then taken as the point of emergence. b) Time-stamping "detergent bottle".

Table S1: Phrases for 60 container stimuli used by Malt et al. (1999).

Index	Phrase	Index	Phrase	Index	Phrase
1	metal container	21	plastic bottle	41	baby bottle
2	film container	22	plastic container	42	glass bottle
3	eye drop bottle	23	spray can	43	plastic jug
4	bottle of vitamins	24	spray can	44	peanut jar
5	bottle of aspirin	25	lotion container	45	tupperware container
6	plastic container	26	bottle of cleaner	46	can of orange juice
7	glass jar	27	jelly jar	47	spray can
8	iodine bottle	28	squeeze bottle	48	plastic container
9	plastic container	29	plastic jar	49	squeeze bottle
10	spice jar	30	glass jar	50	glass bottle
11	olive jar	31	glass jar	51	squeeze bottle
12	baby food jar	32	juice box	52	lotion container
13	plastic jar	33	squeeze bottle	53	salt container
14	glass jar	34	glass jar	54	can of oil
15	applesauce jar	35	plastic container	55	squeeze bottle
16	plastic container	36	peanut butter jar	56	spray bottle
17	plastic jar	37	glass jar	57	detergent bottle
18	glass jar	38	glass jar	58	plastic container
19	glass jar	39	plastic container	59	plastic container
20	squeeze tube	40	baby bottle	60	jug of milk

Change-point detection

To time-stamp a container phrase, we applied a change point detection algorithm (Kass, Eden, & Brown, 2014, sections 14.2.1 and 14.2.2) to its frequency trace. This algorithm defines a change point in the cumulative frequency that optimally divides between a piecewise linear (plateau) region and a quadratic (rising) region. Formally, we search for an optimal changing point τ that minimizes the mean squared error between the empirical cumulative frequency F and the fitted one $\hat{F}$ specified as the following:

$$\hat{F} = \begin{cases} 0 & (\text{if } t < \tau) \\ k(t-\tau)^2 & (\text{if } t >= \tau) \end{cases}, \qquad (1)$$

where t indexes over time and k is a parameter we fit to data. The quadratic form offers a smooth transition in the fitted curve. This change-point detection algorithm yielded a mean correlation of 0.99 (standard deviation: 0.15) between empirical and fitted frequencies across 60 container phrases, which suggests that it is effective in capturing the emerging trends in the historical frequencies.

2. Analysis of model fit

We sought to determine whether the superiority of the chaining model over the clustering model was greater when assessed relative to our data than when assessed relative to hypothetical variants of our data. To test this, we applied a standard permutation test. For each round of permutation, we randomized the category labels of exemplars while preserving the dates (or historical order) and similarity relations of exemplars. We generated 10,000 such permuted sets and applied both models in the same predictive task as we had with the original set for each language. We then compared the between-model (chaining - clustering) difference in predictive accuracy in the original set to each of those in the permuted set. This test yields a significant overall result: the advantage of the chaining model over the clustering model is greater in the actual data than in the permuted datasets (combined $p < 0.005$ under Fisher's test); $p < 0.05$ for English, $p < 0.08$ for Spanish, and $p < 0.02$ for Chinese. We observe that chaining model is less superior for Spanish than for English and Chinese. The explanation for this difference may be that the chaining effect may be sensitive to the sizes of individual lexical categories in the sample (i.e. larger categories tend to allow more room for exemplars to exhibit chaining). Since Spanish has the fewest relatively large-size categories and the most small categories in the sample, it would necessarily allow fewer opportunities for chaining within categories. More specifically, Spanish partitions the 60 exemplars in our stimulus set into 15 modal lexical categories (compared with only 7 in English and 5 in Chinese), where one of these categories includes 28 exemplars, the next largest categories include only 6 exemplars (two each), and all remaining 12 categories contain no more than 3 exemplars. In comparison, English has three relatively large categories of sizes 19, 16 and 15 (covering 50 out of 60 exemplars in the set), and Chinese has two large categories of sizes 40 and 10 (covering equally 50 out of 60 in the set). Overall, this set of results supports the idea that historical chaining is involved in the formation of lexical categories, although the degree to which this can be demonstrated fully may be sensitive to category sizes (or how fine-grained lexical categories partition the space).

References

Kass, R. E., Eden, U. T., & Brown, E. N. (2014). *Analysis of neural data.* New York: Springer.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... Orwant, J. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.