Augmented Comparative Corpora and Monitoring Corpus in Chinese: LIVAC and Sketch Search Engine Compared

Benjamin K. Tsou
City University of Hong Kong,
The Chinese University of Hong Kong,
Hong Kong University of Science and Technology

The increasing availability of numerous corpora has significantly contributed to the understanding of words in terms of their underlying semantic structures and lexical networks (e.g. COBUILD, WordNet etc.). Through data mining and information retrieval, research in this area has vastly expanded our appreciation that what constitutes lexical knowledge goes beyond synonymy, hyponymy, metonymy, meronymy, grammatical and other collocations. Furthermore, they are fundamental to a universalistic conceptual base of ontologies and knowledge representation which are often enriched by deeper and newer analysis. In this context, each language foregrounds specific features or nodes within this knowledge base by usually non-uniform means.

At the same time, the arrival of the age of Big Data has attracted extensive studies on the actual and dynamic use of language as contextualized (ala. Jakobson 1960) within a given society, especially through the mass media. What are foregrounded in this medium tend to have graded cognitive saliency characterizing members of the common speech community, and such shared knowledge is usually at great variance with the thesaurus approach and show noticeable localized features. It is proposed here that the two kinds of knowledge (thesauric vs cognitive-cultural) complement each other in human cognition, and are integral to it.

We draw on two large Chinese media databases Sketch (2.1 billion character tokens[1]) and LIVAC (550 million character tokens[2]) for illustration and discussion. The Sketch Engine in Chinese shows how ***apple*** is, as expected, primarily related to ***orange***, ***peach***, ***fruit***, ***vegetable***, ***food*** etc. At the same time three sub-corpora of LIVAC we draw on show that ***apple*** has a different set of saliency linkage with ***computer***, ***iPhone***, ***Jobs***, ***roll out***, ***share price***, ***company*** etc. This linkage is related less to the universalistic semantic network for ***apple***, than to the foregrounded awareness of ***apple*** as a cultural artifact in actual human social interaction and encoded as social knowledge (Park 1955, Longino 1990). We also show and examine how the salient information associated with ***apple*** varies across the three major Chinese speech communities: Beijing, Hong Kong and Taipei, reflecting social and societal differences, and regional developments, as well as variations over time. Similarly ***free-freedom*** in Chinese varies in associated saliency linkage in the three speech communities in interesting ways but also contrasts with the Sketch Engine results.

The above comparison in LIVAC is made possible by rigorous improvement to the common and simplistic approach to the cultivation and use of databases. The augmentation efforts included the rigorous cultivation of 3 comparable (sub-) corpora for Beijing, Hong Kong and Taipei through geographical (*horizontal*), chronological (*vertical*) and domain (*topical*) partitioning of what is often assumed to be a common linguistic database. This partitioning required well-reasoned pre-conceived criteria to ensure adequate equivalency in comparability in terms of size, period and depth of analysis.

To facilitate comparison we propose a Cognitive-cultural Salience Index (CSI) which draws on comparable corpus data (e.g. LIVAC) to provide comparison of the relative saliency of target words in the relevant corpus and presented as word clouds. The results are viewed in the light of the Sketch Engine output

[1] As per Sketch Engine website.
[2] As per LIVAC website.

Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 1–2,
Beijing, China, July 30, 2015.

to explore how our appreciation of knowledge representation may be enhanced. It will also serve to echo the call to optimize our data collection efforts and to broaden our queries with data judiciously curated and cultivated.

References

K. E. Boulding. 1956. *The image: Knowledge in life and sociology.* University of Michigan Press, Ann Arbor, MI.

C. R. Huang, A. Kilgarriff, Y. Wu, C. M. Chiu, S. Smith, P. Rychly, and K. J. Chen. 2005. Chinese Sketch Engine and the extraction of grammatical collocations. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 48–55.

R Jakobson. 1960. Closing statement: Linguistics and poetics. In T. Sebeok, editor, *Style in Language*.

A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, Kovář V., J. Michelfeit, P. Rychlý, and Suchomel V. 2014. The Sketch Engine: Ten years on. *Lexicography: Journal of ASIALEX*, 1(1):7–36.

Livac. `http://www.livac.org`; `https://en.wikipedia.org/wiki/LIVAC_Synchronous_Corpus`.

H. E. Longino. 1990. *Science and social knowledge: Values and objectivity in scientific inquiry.* Princeton University Press, Princeton, NJ.

R. E. Park. 1955. *Society: Collective behavior, news and opinion, sociology and modern society.* The Free Press, Glencoe, IL.

Sketch engine. `https://the.sketchengine.co.uk`; `https://en.wikipedia.org/wiki/Sketch_Engine`.

B. Tsou and O. Kwong. 2015. LIVAC as a monitoring corpus for tracking trends beyond linguistics. In B. K. Tsou and O. K. Kwong, editors, *Linguistic Corpus and Corpus Linguistics in the Chinese Context*, number 25 in Journal of Chinese Linguistics Monograph Series, pages 447–471, Hong Kong. Hong Kong Chinese University Press.