From *Frying* to *Speculating*: Google Ngram evidence to the meaning development of '炒' in Mandarin Chinese

Jing Chen, Chu-Ren Huang
The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
Yuk Choi Road 11, Hung Hom, Kowloon, Hong Kong
jing95.chen@connect.polyu.hk, churen.huang@polyu.edu.hk

Abstract

This paper explores semantic change simply using the Ngram information, with the intuition that Ngrams as the direct neighbors offer collocation cues that further signal meaning changes. We specifically investigate the case of '炒' in Mandarin Chinese on the basis of unigrams extracted from *Google Books Ngrams Corpus* and reconstruct the meaning development of '炒' with specific stages. The results indicated that the major meaning changes that occurred to '炒' is from 'frying' to 'speculating', which roughly started in the 1970s. The attested word types related to the latter sense denote economic events, such as stocks, foreign currencies, and speculators. It further reflects that social context plays an essential role in the process of semantic change.

1 Introduction

The availability of large historical digital corpora and the recent advance in natural language processing have greatly facilitated empirical studies on semantic change nowadays (Michel et al., 2011; Hamilton et al., 2016a,b; Tahmasebi et al., 2019). Beyond the world of historical linguistics, the computational community has shown a growing interest in exploiting statistical and computational models to automatically detect semantic change over the past two decades, from monitoring the fluctuation in the frequency of target words in historical texts (Michel et al., 2011; Hilpert and Gries, 2009; Kulkarni et al., 2014) to measuring their context differences using the state-of-art distributional models(Kim et al., 2014; Hamilton et al., 2016b; Giulianelli et al., 2020).

Building off of the distributional hypothesis in linguistics, '*You shall know a word by its company*' (Firth, 1957), the distribution-based approach has taken up the predominant position in the recent lexical semantic change detection task (Schlechtweg et al., 2020; Tahmasebi et al., 2019). In general, distributional differences over a period of time could be quantified by constructing and evaluating historical word embeddings (Hamilton et al., 2016a,b). However, this novel trend is still in its youth. Most existing studies focus on evaluating how much the overall distribution of a word form deviated over two or three intervals but tell little about how the meaning of a word developed as a continuous process (Kutuzov and Pivovarova, 2021; Rodina and Kutuzov, 2020; Tahmasebi et al., 2019; Kutuzov et al., 2018). More importantly, the popular word-type models, also known as static word embeddings, may not be sensitive to less contrastive usage drifts (Tahmasebi et al., 2019; Schlechtweg et al., 2020).

In this paper, we present a simple and interpretable way to track the meaning development by collecting evidence directly from N-grams data extracted from *Google Books Ngram* dataset(the Chinese subpart) and manually checking the direct neighbours of target words over times. The intuition here is that N-grams provide collocation information and its changes may signal semantic changes. Besides, the collocations are also more linguistically informative in the sense of interpreting the specific stages of meaning development. Based on this intuition, we discuss the meaning development of '炒' in Mandarin Chinese as a case study and find that it develops a sense of 'speculating' from the original 'frying' sense over the observed period, which coincides but supplements existing observations with empirical data (Diao, 1995; Shen, 2009).

The remainder of this paper is organized as follows. Section 2 summarizes related work and then situates our study. In Section 3, we introduce how

the Ngrams data for this study have been collected. Section 4 presents the distribution information of '炒' in the collected data, and Section 5 discusses the possible changing path of '炒' and its possible reasons.

2 Related work

Semantic change generally refers to 'a form historically acquires a new function to replace or augment its old ones' (Sweetser, 1990). Studies anchored in the distributional hypothesis where 'difference of meaning correlates with difference of distribution' (Harris, 1954) assume that meaning change could be quantified by its neighboring information over time. For example, broadening and narrowing, regarded as two fundamental categories of meaning change, could be further interpreted as extending or contracting context varieties for target words (Campbell, 2013).

Following this working hypothesis, most recent studies use computational methods to monitor the change of context varieties to detect meaning change. Sagi et al. (2009, 2011) performed semantic density analysis by measuring the average cosine similarities of context vectors to identify the increase and decrease of context dispersions. The density information would be further interpreted as the general gain or loss of senses.

Tang et al. (2013, 2016) and Schlechtweg et al. (2017) exploited the *entropy* concept in Information theory to measure the gain or loss of information for target words. For example, Schlechtweg et al. (2017) specifically detected the metaphorical changes in German, an analogical mapping process from a more 'concrete' source domain onto a more 'abstract' target domain (Traugott and Dasher, 2001). They made use of entropy as an indicator to quantify the semantic generality of target words, which further calculate the meaning change.

With the very recent development in computational science, contextualized word embeddings have also been exploited in the detection task (Devlin et al., 2019; Hu et al., 2019; Giulianelli et al., 2020). For example, Giulianelli et al. (2020) first used K-Means clustering to group word token representations derived from BERT models into different word usage types and then applied the Entropy difference and the Jensen-Shannon divergence metrics to measure variations in the relative prominence of coexisting usage types.

The above proposals showed inspiring future directions for automatically detecting semantic change. However, training historical word embeddings, and especially clustering token embeddings have requirements on the computing capacity. The cumulative nature of meaning development and sophisticated relations among senses pose huge challenges for both statistical and neural language models. More crucially, distributional representations have a notorious bias on the frequency of target words. For example, a lower frequency of a novel sense may not be salient enough to be detected.

In this paper, we investigate the meaning development of '炒' by checking its direct neighbors in the Google Ngram corpus. N-grams refer to an n-word sequence (Jurafsky and Martin, 2009), which provides neighboring information for the target words. To some extent, N-grams also reflect collocation preferences. It is clear that the change of collocation preferences would be a more precise and interpretable indicator for meaning development.

3 Method

Google Books Ngram Corpus is a collection of digitized books with over 500 billion words in 7 languages, which is publicly accessible [1]. The data in the corpus is stored in the n-grams format(n is with a maximum to 5) in order to protect intellectual property, containing ngrams with its frequency information in each specific year (Michel et al., 2011; Lin et al., 2012).

For our investigation, we used the Chinese (simplified) subcorpus(version 2), with texts spanning from 1990 to 2012. We first exploit python to crawl all 1-gram and 2-grams (together with their frequencies in each specific year) that occurred larger than two times in the corpus and then filter in words containing the '炒' character in 1-gram and 2-grams for further analysis.

After manually checking all 1-gram data and 2-grams data containing '炒', we found that unigram data actually provide sufficient cues for depicting the meaning development of the case '炒'. Highly frequent collocations with the sense of 'speculating', such as '炒股'(speculate in stocks) and '炒汇'(speculate in foreign currencies) and have been segmented as one unit, that is, 1-gram, over the

[1]Google Books Ngram dataset could be accessed via: https://storage.googleapis.com/books/ngrams/books/datasetsv3.html

whole observed period. Likely, highly frequent expressions with the sense of 'frying', such as '炒米' (fried rice), '炒面'(fried noodles), are also segmented as unigrams.

According to Modern Chinese Dictionary(the 7th edition) (Department of Chinese Lexicography, 2019)，'炒' has four senses: 1) to fry; 2-3) used as '炒作', to speculate or hype; 4) get fired. As mentioned above, highly frequent collocations associated with senses 2-4 have been conventionalized as one unit, such as '炒作', '炒鱿鱼'. We thus assume that 1-gram for '炒' containing collocation and temporal information provides possibilities for meaning reconstruction.

4 The meaning distribution of '炒'

In the 1-gram data, there are 10 word types containing the character '炒'(frying, speculating): '炒米' (fried rice), '炒鱿鱼' (get fired), '炒家'(speculator), '炒汇'(to speculate in foreign currency), '炒面'(fried noodles), '炒货'(fried snacks), '炒股'(to speculate in stocks), '炒冷饭'(to heat leftover, to repeat without any new content), '炒作'(speculating, hype).

Among these words, '炒作' refers to 'speculate' and 'hype', which could be further regarded as less specific action compared with '炒股' etc. We first surveyed '炒' and '炒作' in the Google Ngram Viewer. As Figure 1 below shows, '炒' appeared at least before 1900, while '炒作' was first attested in 1979. A sharp increase of '炒作' occurred in the 1990s.

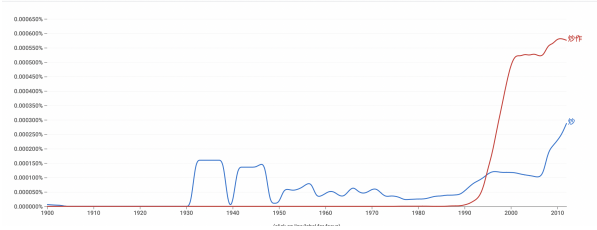

Figure 1: '炒' and '炒作' in *Google Ngram Viewer*

We plotted the remaining 8 word types using the raw data extracted from the corpus to discuss their distributions over time (see Figure 2 below). Among these words, '炒鱿鱼', '炒家', and '炒股' demonstrated a significant rise after the 1980s approximately. In contrast, '炒面', '炒米', '炒冷饭', and '炒货' stay relatively stable during the period from 1900 to 2012.

We also acquired their distributions in *Google Books Ngram Viewer*, which are plotted after normalization and smoothing. As illustrated in Figure 3, frequencies of '炒股', '炒家', '炒汇' have surged around the 1990s, while '炒货', '炒面', '炒冷饭' are much stable [2].

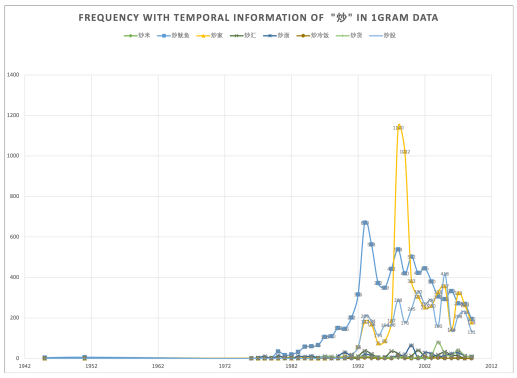

Figure 2: '炒' in 1gram: word type, raw frequency, and temporal information

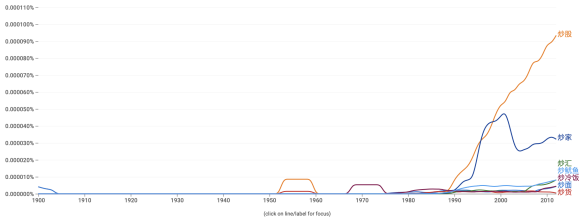

Figure 3: '炒面', '炒货', '炒冷饭', '炒股', '炒汇', '炒家', '炒鱿鱼' in *Google Books Ngram Viewer*

5 Reconstructing the meaning development of '炒'

The frequency and its temporal information regarding the distributions over time provide evidence to trace the meaning development. Data derived from Google Books Ngram indicates the general path of meaning development of '炒', from frying to speculating. As seen from the above figures, '炒' denoting cooking-related action appeared much earlier than the one representing abstract action (such as speculating in stocks and foreign currencies). The latter sense was first attested in the late 19th century. These two senses are closely related, such as having common characteristics of' fast stirring'.

One interesting case '炒作' have senses of 'to speculate' and 'to hype', frequently collocated as '炒作股票'(to speculate in stocks) and '炒作新闻'(to hype). We search these collocations in

[2]The google Ngram viewer uses normalized data for plotting. The low raw frequency of '炒米' made it invisible in the plot.

Google Ngram Viewer (see Figure 4). ‘炒作’ was first attested in the late 1970s, ‘炒作股票’ was first attested in the late 1980s, and ‘炒作新闻’ or ‘新闻炒作’ was in the early 1990s.

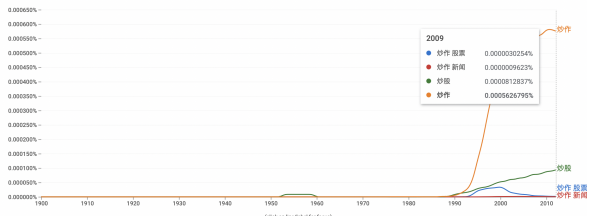

Figure 4: ‘炒作’, ‘炒股’, ‘炒作新闻’, ‘炒作股票’ in *Google Books Ngram Viewer*

Compound verbs such as ‘炒米’, ‘炒面’, ‘炒冷饭’, and ‘炒鱿鱼’ all have a basic sense of ‘frying’. ‘炒鱿鱼’ and ‘炒冷饭’ later acquired extra extended meanings, respectively. For ‘炒冷饭’, ‘to repeat without new content’ was metaphorically developed based on the specific event ‘to heat leftovers’. Similarly, ‘炒鱿鱼’, primarily denoting ‘get fired’ now, also got developed based on the original sense ‘fried squid’. In contrast, ‘炒股’, ‘炒汇’, and ‘炒家’ are closely related to ‘to speculate’, much deviated from the basic sense of ‘frying’.

According to the above analysis, we roughly reconstruct the meaning evolution of ‘炒’ (see Figure 5). The main changing path of ‘炒’ is from frying to speculating, and this change approximately started in the 1970s as witnessed that related compound words with a sense of speculating had a surge in terms of frequencies. The rising tendency even becomes more salient around the 1990s. The latter sense is closely related to economic events, such as stocks, foreign currencies, and speculators, in our data. The development of the ‘speculating’ sense roughly coincides with the timeline of the Reform and Opening-up, one of the most influential milestones in the recent history of modern China. It is generally assumed that this remarkable social change brought significant changes to the lexicon of Modern Chinese (Diao, 1995; Lin, 2021). In this case, we would assume that profound social change is one of the most fundamental driving forces behind the meaning development of ‘炒’.

6 Conclusion

In this paper, we reconstructed the meaning development for ‘炒’ in Mandarin Chinese, using

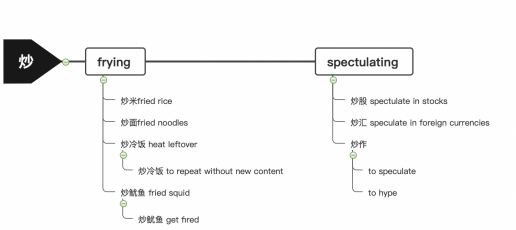

Figure 5: A possible changing path for the meaning of ‘炒’

Google Ngrams data, with the intuition that collocation information in Ngrams helps detect meaning change. Our results also provided more specific time information for the stages when ‘炒’ acquired the sense of ‘to speculate’. The meaning development of ‘炒’, from denoting the concrete cooking-related action to a relatively abstract economics-related action, further reflected that the social context plays an essential role in semantic change.

This paper showcased that Ngrams provide possibilities to depict the changing path of meaning directly. However, there are also some limitations in this study. For example, given the nature of Ngrams, a sliding window for a given sentence or sequence, there are ngrams that are less informative in terms of collocation. A question here is about how to evaluate ngrams in terms of its collocational weights. Another related question is about which types of target words would be suitable for such detection. These questions are served as research directions in the near future.

References

Lyle Campbell. 2013. *Historical linguistics*. Edinburgh University Press.

Chinese Academy of Social Science Department of Chinese Lexicography, Institute of Linguistics. 2019. *Contemporary Chinese Dictionary (Xiandai Hanyu Cidian)*, the 7th edition. Commercial Press, Peking.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Yanbin Diao. 1995. *The Development and Reform of Mainland Chinese in the New Era*. Hung Yeh Publishing, Taibei.

J.R. Firth. 1957. *A Synopsis of Linguistic Theory, 1930-1955*.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of ACL*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of EMNLP*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of ACL*.

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

Martin Hilpert and Stefan Gries. 2009. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24:385–401.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. *arXiv preprint arXiv:1405.3515*.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Statistically Significant Detection of Linguistic Change.

Andrey Kutuzov and Lidia Pivovarova. 2021. Threepart Diachronic Semantic Change Dataset for Russian. In *Proceedings of the ACL International Workshop on Computational Approaches to Historical Language Change*.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic Word Embeddings and Semantic Shifts: A Survey. In *Proceedings of COLING*.

Sheng Lin. 2021. *30 Words that have experienced "expressional" changes in recent years*, pages 369–380. De Gruyter Mouton, Berlin, Boston.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, and Peter Norvig. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: A Dataset of Historical Lexical Semantic Change in Russian. In *Proceedings of COLING*.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. *Proceedings of the EACL Workshop on GEMS: Geometrical Models of Natural Language Semantics*.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. *Tracing semantic change with Latent Semantic Analysis*, pages 161–183.

Dominik Schlechtweg, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde, and Daniel Hole. 2017. German in flux: Detecting metaphoric change via word entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 354–367, Vancouver, Canada. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of SemEval*.

Mengying Shen. 2009. *New Words and New Expressions in Chinese New Era (1949-2009)*. Sichuan Lexicographical Press, Chengdu.

E. Sweetser. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge Studies in Linguistics. Cambridge University Press.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2019. Survey of Computational Approaches to Lexical Semantic Change. *arXiv preprint arXiv:1811.06278*.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2013. Semantic Change Computation: A Successive Approach. In *Behavior and Social Computing*, pages 68–81, Cham. Springer International Publishing.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2016. Semantic Change Computation: A Successive Approach. *World Wide Web*, 19.

E.C. Traugott and R.B. Dasher. 2001. *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge University Press.