What Factors can Facilitate Efficient Propagation of Chinese Neologisms -- a Corpus-Driven Study with Internet Usage Data

Menghan Jiang[1], Kathleen Ahrens[2], and Chu-Ren Huang[3]

[1]Chinese Language Center, Shenzhen MSU-BIT University, Shenzhen, China
[2]Department of English and Communication, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
[3]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
menghan.jiang@connect.polyu.hk;
kathleen.ahrens@polyu.edu.hk;
churen.huang@polyu.edu.hk

Abstract. With the development of information technologies, our world currently faces such an overwhelming mass of neologisms. Therefore, the study of neologisms has become an important research topic in recent years [1]. In this research, we investigate the factors that facilitate the efficient propagation of Chinese neologisms, based on Internet usage data extracted from Google Trends. We collected 342 neologisms from the published authoritative lists and annotated them with eight factors that potentially contribute to their popularity. The results demonstrate that the development speed can to some extent be successfully predicted by certain factors, such as the topic, syntactic type, length, and semantic polarity of the neologisms. We also calculated weights for each factor and found that the syntactic type and semantic polarity of the neologisms played more significant roles in their development.

Keywords: Chinese Neologisms, Internet Usage Data, Corpus-Driven Approach

1 Introduction

With the development of information technologies, our way of life and communication have changed significantly. Instead of traditional face to face communication, people nowadays communicate with others, seek information, and provide opinions online on social media. Ubiquitous and immediate access to information with the potential for real-time responses has provided a strong impetus for creating ideas and has allowed such ideas to be spread more broadly and quickly than ever before. This has allowed the emergence and spread of neologisms,

especially, those that occur on social media and other websites, which are known as Internet neologisms. The neologisms used online have proliferated worldwide.

Moreover, due to different sociocultural interests, Internet neologisms that originated from different regions are oftentimes distinct, allow for an examination of social preferences. To investigate Internet neologisms not only helps us to know aspects of different cultures, it also sheds light on sociology, human behavior, and especially linguistic trends of novel word usage [2][3].

Hence, Internet neologisms have attracted the interest of language researchers, especially in investigating the driving factors involved in the development of neologisms, and in predicting the survival chance of neologisms in the language system (e.g., [4][5]). A variety of studies focusing on conditions or factors that could maximally explain lexical establishment, i.e., in determining whether a neologism will disappear or survive (e.g., [6]-[8]). For example, word frequency is considered to play a very important role in explaining the success story of words, life stages, and the prediction force of whether a word may survive after being coined (e.g., [9][10]). However, many studies have been conducted under a qualitative paradigm, and only a few exceptions adopt statistical tools and large-scale data (e.g., [9-12]). The reason for lacking quantitative studies is that it is almost impossible to have reliable documentation for the process of fast changes [13]. Since it has been observed that neologisms nowadays evolve and fade much faster than ever before, traditional linguistic approaches to date have lacked a way to deal with either the scale or the speed of the fast-development neologisms.

To address this problem, Jiang et al. [12][14] provides a new method to obtain statistical data for neologisms from an Internet statistical tool, Google Trends, to measure the popularity of the fast-changing neologisms. Google Trends is an online search tool that can provide the searched frequency of the selected word within a flexible range of time intervals. They argue that the Internet search frequency is a reasonable approximation of its popularity and how the word spreads over the Internet. Their study adopts internet-based data from Google Trends to model the life cycle of neologisms, and the results indicate that the propagation of neologisms is similar to the propagation of diseases. However, they have not investigated which factors (such as word length, semantic polarity or syntactic type) might influence the life cycle of neologisms. In this study, we follow the method of Jiang et al. [12][14] also using Google Trends. Based on the Internet usage data, we aim to examine the factors which can determine the development speed and facilitate the efficient propagation of Chinese neologisms.

2 Method

2.1 Data source

We collected the neologisms following the published authoritative lists from three platforms: 咬文嚼字, National Language Resources Monitoring and Research Center (joint with other platforms), and 上海语言文字周报. At the end of each year or the beginning of the next year, these three organizations respectively release the "annual neologisms" of the previous year based on the development of the language in the past year. Although the platforms utilize different criteria in selecting the words, they all have strict and scientific identification standards. The results are authoritative and recognized by the public. We have collected the annual neologisms from each year between 2008-2021, for a total of about 342 words. We have not taken the most recent neologisms from the past year to ensure that there is enough longitudinal data for the development of each new word.

2.2 Popularity Measurement

As mentioned in the introduction, we follow Jiang et al. [12][14] and measure the word's popularity by its internet search frequency to investigate the variation of the rapid change of neologisms. The Internet search frequency is a reasonable approximation of how the words spread over the Internet and the frequency of their use. Fortunately, this data can be easily accessible and downloaded from Google Trends for free. Google Trends is an online tool that provides access to a large sample of actual search requests. The daily search frequency of a specified neologism reported on Google Trends is normalized to a range from zero to one hundred, which is adopted as a reflection of each word's development, from its coinage (0) to the most popular time (100).

In this study, we accessed Google Trends to generate a data pool of the search frequencies of the 342 selected neologisms as our source data. We calculate the development duration based on the searched frequency, and the duration is defined as how long (how many days) a neologism takes to develop from 0 (or a very low-frequency level) to 100. For example, Figure 1 illustrates the development of 不差钱 shown in Google Trends. The word emerged and started to be popular on Jan 18th 2009, reaching its peak at the normalized value of 100 on Feb 1st 2009.

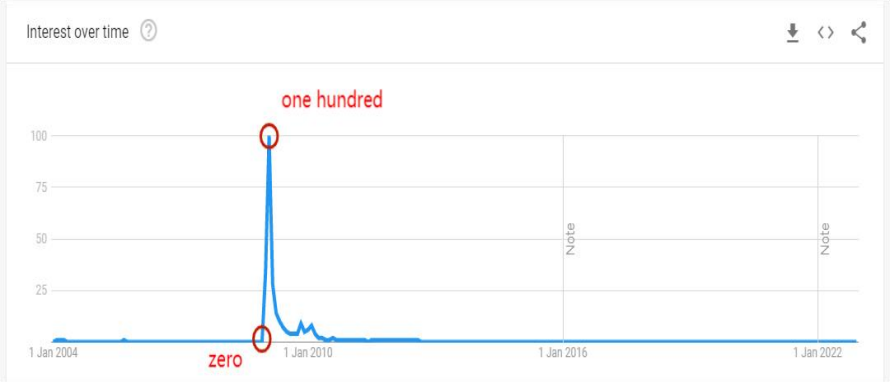

Figure 1. The Popularity of 不差钱

2.3 Data Annotation

We manually annotated each neologism with 8 factors which could influence the popularity of the neologisms. These factors were selected based on previous research and our observations (e.g., [15]). The annotation schema is shown below in Table 1. Note there are three topics listed under "Topic", Economic, Social, and Political. We hand coded all the categories in each of the eight factors and will report inter-rater reliability in the main study.

Table 1. The Annotation Schema

Factors	Explanations	Examples
Length	How many characters a neologism contains	给力：**2** 挖掘机技术哪家强：**8**
Time	In which year the neologism was coined	囧：**2008** 柠檬精：**2019**
Platform	On which platform was the neologism published	**语言监测中心**：脱贫攻坚（2020） **咬文嚼字**：直播带货（2020） **语言文字周报**：甩锅（2020）
Topic	The topic the neologism refers to	**Economic**：口红效应 **Political**：反腐 **Social**：蜗居
POS	The syntactic type of the neologism	**N or NP**：高铁 **V or VP**：打 call **ADJ or ADJP**：中国式 **Clause**：我太南了
Polarity	The semantic polarity	**Negative**：坑爹

	of the neologism	**Neutral**：微电影 **Positive**：一带一路
Metaphor	Whether the neologism contains conceptual metaphors or not	**Yes**：油腻 **No**：不差钱
Coinage	Whether the neologism is a new form or an existing form with new interpretations	**A new word:** 高铁 **A new meaning is assigned to an existing form**: 山寨

3 Data Analysis

We further utilize a linear regression to examine the relationship between the development duration and the potential factors. The results can be summarized in Table 2. We transfer the results into signs. The "+" can be interpreted as the duration is positively correlated with the factor, while the "-" displays that the duration has a significant negative correlation with the factor, and "0" indicates that no statistically significant results have been detected.

Table 2. The Results of Linear Regression

	Duration
Time	0
Platform	0
Coinage	0
length	-
Topic-social	-
Topic-economic	+
Topic-political	-
POS-V	+
POS-N	-
POS-ADJ	0
Polarity-negative	-
Polarity-neutral	+
Polarity-positive	+
Metaphor-Yes	0

As can be seen from the results, the overall regression is statistically significant ($R2 = 0.73$, $F(2, 17) = 23.46$, $p = < .000$). Particularly, the development duration of neologism is significantly increased by the occurrence of the economic topic, while is significantly decreased by the political and social topic. The duration is significantly increased by the verbal neologism, while it is significantly decreased by the nominal

neologism. The duration is positively correlated with a neutral and positive context, while it is negatively correlated with a negative context. The duration cannot be predicted by the time being coined, platform, coinage, and metaphorization. In sum, the regression model predicts that a nominal neologism of longer length, with negative semantic polarity, refering to social or political topics, is more likely to spread faster (i.e., it takes a shorter time to reach the peak).

Table 3. The Results of Linear Regression with weight

	coef	std err	t	P>|t|	[0.025	0.975]
const	130.0245	24.890	5.224	0.000	81.061	178.988
length	-8.0138	4.955	-1.617	0.107	-17.761	1.733
topic_political	-20.2836	19.793	-1.025	0.306	-59.220	18.652
topic_social	-29.6053	15.912	-1.861	0.064	-60.907	1.697
syntactic type_m	85.9967	51.083	1.683	0.093	-14.494	186.488
syntactic type_p	-0.3953	16.202	-0.024	0.981	-32.268	31.477
syntactic type_w	47.7278	21.982	2.171	0.031	4.484	90.972
POS_ADJP	10.1767	21.119	0.482	0.630	-31.369	51.722
POS_N	-53.5797	31.915	-1.679	0.094	-116.362	9.203
POS_NP	-10.4983	12.367	-0.849	0.397	-34.827	13.831
POS_V	-22.2108	37.363	-0.594	0.553	-95.712	51.290
POS_VP	-0.0737	12.677	-0.006	0.995	-25.012	24.864
POS_clause	-3.3047	29.127	-0.113	0.910	-60.603	53.993
polarity_neutral	69.2954	24.707	2.805	0.005	20.692	117.899
polarity_p	25.1777	12.116	2.078	0.038	1.343	49.012

More importantly, instead of investigating the influence of each factor independently, in the next step, we also include the calculation of weight for the significant factors, since the development of neologisms is affected by various factors simultaneously. We perform a multivariate regression analysis with wight and present the results in Table 3. The coefficients in the table illustrate the relative contribution of each factor to the duration of neologism development. Notably, we find that the “syntactic type -- morpheme” factor has the most significant positive effect on the duration, followed by neutral polarity. Additionally, “syntactic type -- word” and positive polarity also contribute positively to the duration, while social and political topics have a negative impact. The factor “Noun” has the most pronounced negative influence on the duration. In summary, our findings suggest that the syntactic type and semantic polarity of neologisms are crucial factors in their spread.

4 Summary

This study examines the potential factors that may facilitate the efficient propagation of neologisms. The results show that the popularity and development speed can, to some extent, be predicted by certain features such as the topic, syntactic type, length, and semantic polarity of the neologisms. In the future study, more data with more features annotated will be included, to further explore how neologisms are created and developed.

References

1. Jing-Schmidt Z, Hsieh. Chinese neologisms. In: Huang CR, Jing Schmidt Z, and Meisterernst B, editors. The Routledge Handbook of Chinese Applied Linguistics. Routledge (2019) 514-534

2. Sonnad N: How brand-new words are spreading across America. Quartz. 2015 July 30. Available from: https://qz.com/465820/how-brand-new-words-are-spreading-across-america/

3. Castellví, Maria T. Cabré, Rosa E. Bagot, and Chelo V. Sierra: Neology in specialized communication. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication. 2012 18(1): 1-8

4. Schmid, Hans-Jörg: New words in the mind: Concept-formation and entrenchment of neologisms. Anglia. 2008 126(1): 1-36

5. Renouf, Antoinette: A finer definition of neology in English: The life-cycle of a word. Corpus perspectives on patterns of lexis. 2013 57: 177

6. Klosa-Kückelhaus, Annette, and Sascha Wolfe: Considerations on the Acceptance of German neologisms from the 1990s. International Journal of Lexicography. 2020. 33(2): 150–167. https://doi.org/10.1093/ijl/ecz033.

7. Edmonds, Bruce: Three Challenges for the Survival of Memetics. Journal of Memetics - Evolutionary Models of Information Transmission6. 2002 45-50

8. Metcalf, Allan: Predicting New Words: The Secrets of their Success. Boston, New York: Houghton Mifflin Harcourt 2004

9. Altmann, Eduardo G., Janet B. Pierrehumbert, and Adilson E. Motter: Niche as a determinant of word fate in online groups. PloS one. 2011 6(5): e19009

10. Altmann, Eduardo G., Zakary L. Whichard, and Adilson E. Motter: Identifying trends in word frequency dynamics. Journal of Statistical Physics. 2013 151(1-2): 277-288

11. Heylighen, Francis, and Klass Chielens: Cultural Evolution and Memetics. In Encyclopedia of Complexity and Systems Science, ed. by Robert A. Meyers. Berlin, Germany: Springer 2009 3205-3220.

12. Jiang M., X. Shen, K. Ahrens, and C. R. Huang: Neologisms are Epidemic: Modeling the Life Cycle of neologisms in China 2008-2016. PloS one. 2021 16(2): e0245984

13. Lei, S., R. Yang, and C. R. Huang: Emergent neologism: A study of an emerging meaning with competing forms based on the first six months of COVID-19. Lingu. 2021 103095

14. Jiang M., K. Ahrens, X. Shen, Sophia Y. M. Lee, C. R. Huang. Do New Words Propagate Like Memes? An Internet Usage Based Two-Stage Model of the Life Cycle of Neologisms. Accepted by Journal of Chinese Linguistics.

15. Tsur, O., and Rappoport, A. Don't let me be# misunderstood: Linguistically motivated algorithm for predicting the popularity of textual memes. In Proceedings of the International AAAI Conference on Web and Social Media. 2015 9(1): 426-435