Journal of Literature, Languages and Linguistics
ISSN 2422-8435 An International Peer-reviewed Journal
Vol.45, 2018

A Corpus-based Approach to the Quantitative Criterion for the Lexicographic Listing of Neologisms

Dr. Tao Ma
School of Foreign Studies, Shanghai Sanda University 2727 Jinhai Road, Shanghai, 201612, China
E-mail of the corresponding author: taoma@sandau.edu.cn

Abstract

This study answers whether there is a consistent quantitative criterion for the lexicographic listing of neologisms in English and Chinese. The quantitative patterns of neologisms to lexicographic listings are analyzed by using a comparative diachronic corpus-based approach to their pre-listing frequency patterns and the latencies between their initial occurrences in the WebCorp and the BCC Corpus and their final listings in the Oxford English Dictionary and the Contemporary Chinese Dictionary. It is found that Chinese and English neologisms display similar patterns of frequencies and latencies, based on which an implicit listing criterion is revealed.

Keywords: quantitative criteria, lexicographic listing, neologism, corpus-based

1. Introduction

A consistent criterion for listing a neologism is problematic for lexicographers due to the conflict between dictionary design and user profile, the inconsistency between needs and performance, and the divergence between exhaustiveness and effectiveness. Attention drawn to detecting and extracting neologisms in the field (Roche 1999, Cook 2011, Lau 2012, Torres del Rey 2014) is as much divided as the one paid to their listing rules for dictionary specification (Landau 2001, Breen 2004, Su 2003, Tao 2004), since there is few generally applicable frequency-based rules of neologism listing, which are hardly reliable when dictionaries of different languages are concerned.

Though stringent criteria of neologism listings is no more established than explicit agreements for neologism rejections in its uncharted semantic field among stakeholders of a dictionary, either rejection or acceptance constitutes one side of the same coin for neologism entries in published dictionaries with respect toothier quantitative properties before and after the listing. Therefore, corpus frequency patterns of new entries in dictionaries can inversely reveal lexicographic criteria for neologism listings, and the implicit rules of expert insights can be explicitly outlined in a more systematic and applicable way with reference to frequency data.

2. Criteria for neologism listings

There are two perspectives towards neologisms, namely lexicographic versus lexicological, based on which the listing criteria of neologisms vary considerably. From a lexicographic standpoint (Atkins & Rundell 2008), a neologism is the new combination of old lexical forms, which can be syntactic or morphological. At the syntactic level, a new combination is used but no morphological gap is filled, which attributes to the semantic expansion of lexis, or a semantic neologism. At the morphological level, a new lexical item is created and a morphological gap is filled, by which morphological expansion accompanies with semantic expansion of lexis, and consequently a lexical neologism appears.

The key task in the lexicographic analysis of neologisms is to determine whether the frequency is sufficiently high and necessarily consistent, in other words, the validity and reliability of corpus frequency before their listings. However, when the life-cycle of a neologism is tracked, the reoccurrence of most recorded neologisms is highly volatile (Renouf 2007, 2013). The simple reason behind this phenomenon of past-future conflict is that the baseline index of a neologism candidate reported in a corpus provides little diachronic information, so the prediction for its future use can be inaccurate and the window period of frequency confirmation is so limited that the relatively low probability of frequency consistency reduces its dictionary searchable value. In essence, a lexicographic perspective captures the supply-side of neologisms for formal purposes in observation, which is replicable and applicable in practice for dictionaries of different languages, but comparatively, it overlooks the demand-side of neologism listing from a functional approach, by which a so-called lexicological perspective avails.

The main difference between lexicographic and lexicological perspectives lies in the treatment of neologisms in terms of their formal and functional aspects. The formal observation of use based on frequency data can rarely indicate the functional observation of usage and usefulness. In other words, listing words is *lexicographic* but hardly *lexicological* (Considine & Iamartino 2007), as exhaustiveness is not equivalent to utility. A lexicological perspective focuses on the semantic ambiguity of a neologism for dictionary explanation while a lexicographic perspective emphasizes findability and occurrences for dictionary record. A sufficiently high frequent neologism of low ambiguity does not necessarily need explanation or, arguably, even listing in a dictionary, whereas a sufficiently ambiguous neologism of low frequency may not necessarily require a lexicographic listing, either, because sufficiency does not entail necessity. So, a neologism candidate for lexicographic listing has to satisfy two conditions at the same time: the degree of ambiguity and the level of frequency, based on which the selection of a neologism pertaining to the design of a dictionary varies. A lexicographic approach probes into the supply-side of neologism generation and the empirical basis of a neologism in the light of the frequency data from a corpus for the sufficiency of its listing, while a lexicological approach, on the contrary, deals with the demand-side of a neologism to be explained in a dictionary which constitutes the necessity of its listing.

Noticeably, in the editing process of the Contemporary Chinese Dictionary the 5th edition (2005), a monolingual dictionary of standard Chinese language, multi-layered criteria are adopted to select and listed neologisms in addition to the frequency-based criteria, namely event-based, meaning-based, structure-based and derivation-based (Tao 2004). The event-based criterion is a socio-linguistic measurement to determine the degree of social influence by which a neologism candidate related to a socially recognizable event is judged. The meaning-based criterion is used to decide whether a neologism candidate has constructed a new semantic feature in the lexicon of a language while the structure-based criterion is referred to determine whether a neologism candidate has been placed in a new syntactic combination of the existing lexemes. The derivation-based criterion is applied to judge whether a neologism candidate is derived from the existing morphological or syntactic structure. Two criteria out of four mentioned above have to be met at the same time for the neologism candidate to be listed in the dictionary.

In fact, the multi-layered criteria are the modified version of composite definitions for the classification of neologisms, that is: semantic versus lexical. The definition of a semantic neologism identifies the change in the existing lexemes' collocation and co-text while the perspective to a lexical neologism focuses on the initial latency of occurrence in a diachronic corpus (Renouf 2013). To certain extent, the data source for the classification and verification of neologisms to their definitions is largely built on the proper methods of extracting information from diachronic corpora while the valid measurable data of social influence from an event can hardly be collected other than the frequency of the event-related lexemes, so the multi-layered criteria testing returns to a frequency-based approach to neologisms again and the perspective to the same issue of diffusion and consistency is changed. Therefore, the legitimacy of a neologism listing and its proof is largely provided by the treatment of its frequency data by a quantitative analysis.

3. Exploring the borderline

Quantitative patterns of neologisms to their lexicographic listings will be analyzed to reveal whether there are consistent listing criteria for neologisms. Two variable, pre-listing frequency in diachronic corpora and latencies between their initial occurrences and final listing, will be observed to find the patterns. The two dictionaries referred to here are the Oxford English Dictionary and its Chinese counterpart, the CCD, for both of them being encyclopedia dictionaries to contain a substantial number of neologisms. Importantly, the year 1989 when the OED 2nd edition was published coincides with the year when 5th edition of the CCD Dictionary was also released, and this gives the starting point for observation. While the OED began to release its new word lists annually on the internet since 2000, the CCD published its 6th edition in 2009 which constitutes the cutting point of observation for lexicographic listing. Therefore, the window period for collecting pre-listing data of target lexical neologisms is from 1989 to 2009, and for the post-listing data from 2010 to the present.

The cumulative release of new words from the OED online will be the source of English lexical neologisms, while the entry difference between the two editions of the CCD dictionary over the time period will be the source of Chinese lexical neologisms. This study will randomly select 100 target candidates from each of the two dictionaries. The definition of a lexical neologism follows a technical criterion as follows: 'a lexical neologism is usefully often a lexical item which occurs for the first time in a diachronic corpus' (Renouf 2013:178). The two diachronic corpora used in this study are the WebCorp for English and the BCC for Chinese. WebCorp is a web-based English corpus being able to record and report diachronic frequency information (Renouf 2002), while the

BCC is a Chinese on-line corpus having the same functions (Xun *et al.* 2016).

4. A comparison between English and Chinese neologism listing patterns

It is calculated that there are more than twelve thousand new entries listed in the OED at the end of 2009 since their online release in 2000, and it is reported that, coincidentally, there are more than twelve thousand new entries in the CCD 6th edition, compared with its 5th edition (Cao2010), despite the fact that the number of entries deducted from the previous edition is around seven thousand, which makes the net increase of entries only around four thousand in record.

The one hundred neologism candidates are randomly selected in an alphabetic order with an interval of one hundred entries. Their corpus attestedness before 1989 is checked to confirm their lexical neologism status in accordance with the selection criteria of this study. There are five attested candidates in Chinese and one in English, which triggers a second round of selection.

Table 1. A least squares analysis of the diachronic frequencies.

Least squares analysis		Increasing trend		Decreasing trend	
		High Cst.	Low Cst.	High Cst.	Low Cst.
Pre- listing	English	33	48	11	8
	Chinese	31	33	16	20
Post- listing	English	51	36	4	9
	Chinese	42	40	6	12
Overall	English	48	41	4	7
	Chinese	47	43	4	6

A positive regression coefficient indicates the increasing frequencies of a lexical neologism over the time period of the diachronic corpus data, which suggests that the target lexical neologism has an increasing trend of use and coverage, while a negative regression coefficient denotes decreasing frequencies, which suggests the opposite trend. A high coefficient of determination indicates a high consistency of frequency development over the time period of the diachronic corpus data, while a low coefficient indicates the opposite. In this study, the coefficient of determination above 0.7 is marked as high consistency while below 0.7 as low consistency abbreviated as Cst. in the table 1.

In the pre-listing set, the least squares analysis indicates that 64 Chinese lexical neologisms have positive regression coefficients from 0.9 to 6.5 with the coefficients of determination from 0.44 to 0.84, while the rest have negative regression coefficients from -2.8 to -0.43 with a similar range of the coefficients of determination. 81 English lexical neologisms have positive regression coefficients from 0.69 to 5.8 with the coefficients of determination from 0.42 to 0.85 while the rest have negative coefficients from -3.2 to -0.33 with a similar range of the coefficients of determination.

In the post-listing set, the least squares analysis indicates that 82 Chinese lexical neologisms have positive regression coefficients from 0.8 to 3.5 with the coefficients of determination from 0.54 to 0.87, while the rest have negative coefficients from -2.8 to -0.43 with a similar range of the coefficients of determination. 87 English lexical neologisms have positive regression coefficients from 1.7 to 2.8 with the coefficients of determination from 0.53 to 0.88 while the rest have negative coefficients from -2.5 to -0.67 with a similar range of the coefficients of determination.

There is a generally increasing trend of diachronic corpus frequency data over the observation period from the initial corpus record to the present time, which suggests the constant diffusion of neologisms in terms of their use and coverage. A comparatively higher coefficient of determination in the post-listing period than the one in the pre-listing suggests a gradually consistent trend of frequency development after the lexicographic listing. That the annual mean frequencies are relatively similar across different categories of latencies and languages suggests the convergent patterns of lexical neologisms to lexicographic listings despite that they are not completely statistically significant.

Table 2. Mean and Standard Deviation of latencies (Years).

		Positive Coefficient		Negative Coefficient	
		Mean	SD	Mean	SD
English	High	7.7	3.3	9.3	6.1
	Low	9.3	3.9	11.3	7.1
	Overall	8.4	3.8	10.3	6.2
Chinese	High	8.4	3.2	9.5	6.5
	Low	9.5	3.8	10.8	6.9
	All	8.7	3.5	9.9	6.8

However, the fact that there are similar annual mean frequency data before and after lexicographic listing implies the developmental patterning of a lexical neologism to be a listed lexeme in the lexicon of a language. This also helps to answer when a lexical neologism starts to be frequently used with reference to its lexicographic listing, which suggests a sufficient condition for a frequency-based listing criterion of lexical neologisms from a dictionary. When the annual mean frequency of a neologism candidate is no more than 5 per million in a corpus, it is highly likely that the neologism candidate will be rejected from a dictionary, despite the fact that frequency is not a necessary condition for the lexicographic listing of a lexical neologism when the latencies between corpus record and lexicographic listing is referred to.

Noticeably, the latencies of high frequency are generally shorter than the ones of low frequency in both languages, which suggests that the more frequent lexical neologisms tend to have earlier lexicographic listings than the less frequent ones. The unbiased mean and standard deviation of the latencies before the lexicographic listing is 8.25 years and 3.77 years for English and 8.62 years and 3.43 for Chinese. Therefore, the sufficient condition of lexicographic listing in terms of latencies is 5 years by applying the same calculation method above. When a lexical neologism has no more than 5 years corpus presence, it is highly likely that the lexical neologism will be rejected from a dictionary.

The Pearson coefficient of the correlation between annual frequencies and latencies is 0.87 in English and 0.84 in Chinese, which suggests the connections between diffusiveness and consistency of lexical neologisms. The validity of the observation and the analysis method in this study is partially confirmed by the correlation test which indicates that the more frequently a lexical neologism is used, the less time it will consume before its lexicographic listing. Therefore, these findings reveal two sufficient conditions for the lexicographic listing of lexical neologisms: sufficient frequency and sufficient time. The similarities of frequency and latency between English and Chinese suggest the convergent patterns of lexical neologisms to their lexicographic listings.

5. Discussion

There are many debates over different criteria and principles pertaining to the dictionary listings of lexical neologisms when different approaches and perspectives to lexicography and neology are considered. However, there are convergent patterns found in this study for different lexical neologisms between English and Chinese from the same type of dictionaries, despite the fact that there are substantial variations in terms of diachronic corpus frequencies and latencies of individual lexical neologisms. The convergent minimum requirement that the lexical neologisms investigated have the mean annual frequency above 5 per million and the mean latency above 5 years suggests an implicit listing criterion of frequency and latency for the dictionary listing of lexical neologisms.

It should be acknowledged that the implicit criterion revealed here may not be universally applicable to other dictionaries but an internal consistent criterion for neologism listing should be consistently binding in the same dictionary. Although the register and genre variations of neologisms are not examined in this study, it is believed here that the single sufficient condition of frequencies over substantial length of time constitutes the demand for dictionary explanation even in a single register or genre, which means that the neologism candidate with a frequency above certain level over certain amount of time should be listed in the dictionary regardless of other factors. So, it may be problematic for the OED or the CCD not to list a potential neologism candidate which

satisfies the minimum criterion of frequency and latency. This uncritical opinion echoes the idea from a Chinese linguistic expert Lv(1984:14) who once commented on the dictionary listing of neologisms that'[i]t will more acceptable to good 'mistake' in the dictionary listing of neologisms by loose criteria than by stringent criteria'. Therefore, the 5 per million over 5 year listing criterion of frequency over latency should not be regarded as a loose criterion when the decision of lexical neologism listing is to be made.

6. Conclusion

Listing patterns of neologisms from the OED and the CCD are examined in terms of their diachronic frequencies and latencies data from the WebCorp and the BCC. Listed neologisms display a minimum annual mean frequency of 5 per million and a minimum mean latency of 5 years between the initial corpus occurrence and final dictionary listing. It is proposed that an explicit quantitative criterion, the frequency-latency ratio, to list a neologism entry, besides the intuitive expert suggestions of specific dictionary design, is a good 'standard practice' (Lv 1984) to initiate and facilitate the uncritical listing of neologisms.

References

Atkins, B.T.S. & Rundell, M. (2008), *The Oxford Guide to Practical Lexicography*, Oxford: Oxford University Press.

Cao, F. 2010 "Summary of compiling and publishing works of CCD Sixth Edition", *Dictionary research* (3),112-134.

Considine, J. & Iamartino, G. (2007), *Words and Dictionaries from the British Isles in Historical Perspective*, Newcastle. Cambridge Scholars.

Cook, P. & Graeme, H. (2011), "Automatic identification of words with novel but infrequent senses", In *PACLIC* 265–274.

Landau, I. S. (2001), *Dictionaries: The Art and Craft of Lexicography*, Cambridge, Cambridge University Press.

Lau, J. H., Cook, P., McCarthy, D., Newman, D. & Baldwin, T. (2012), "Word sense induction for novel sense detection", *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* 591–601.

Lv, S.X.(1984), "Everyone is concerned about new words and new meanings", *Dictionary Research* (1).8-14.

Renouf, A. (2003),"WebCorp:providing a renewable data source for corpus linguists", in Granger, S. & Petch-Tyson, S.(eds.)*Extending the scope of corpus-based research:new applications, new challenges*,Amsterdam/Atlanta GA:Rodopi,39-58.

Renouf, A. (2013), "A finer definition of neology in English: The life-cycle of a word. In: Corpus Perspectives on Patterns of Lexis", *Studies in Corpus Linguistics* 57. John Benjamins Publishing Company, 177-208.

Roche, S. & Bowker, L. (1999), "Cenit: Système de Détection Semi-Automatique des Néologismes. Terminologies Nouvelles", vol. 20. Sablayrolles, Jean-François. *Neologieet Dictionnaire(s) comme Corpus d'Exclusion*. Actes de la Journée des dictionnaires de l'Université de Cergy. CergyPontoise, France

Su, X. C. & Huang, Q. (2003), "The maturity of the 2003. new words and the criteria for the selection of the standard dictionary -- on the "appendix new words in the modern Chinese dictionary", *The study of dictionaries* (3): 106-113.

Tao, L. (2004), "Suggestions on the revision of Xinhua Dictionary of new words ",. *Dictionary Research* (6): 27-36.

Torres del Rey, J. & Maroto, N. (2014), "Building the interface between experts and linguists in the detection and characterisation of neology in the field of the neurosciences", *Proceedings of the 4th International Workshop on Computational Terminology,* Dublin.

Xun E.D., Rao, G.Q., Xiao,X.Y. & Zhang, J.J. (2016), "The construction of the BCC Corpus in the age of Big Data", *Corpus Linguistics* **3**(1),93-118.