# When Frequency Data Meet Dispersion Data in the Extraction of Multi-word Units from a Corpus: A Study of Trigrams in Chinese

**Chan-Chia Hsu**
Graduate Institute of Linguistics, National Taiwan University
No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan (R.O.C)
chanchiah@gmail.com

## Abstract

One of the main approaches to extract multi-word units is the frequency threshold approach, but the way this approach considers dispersion data still leaves a lot to be desired. This study adopts Gries's (2008) dispersion measure to extract trigrams from a Chinese corpus, and the results are compared with those of the frequency threshold approach. It is found that the overlap between the two approaches is not very large. This demonstrates the necessity of taking dispersion data more seriously and the dynamic nature of lexical representations. Moreover, the trigrams extracted in the present study can be used in a wide range of language resources in Chinese.

## 1 Introduction

In the past decades, multi-word units have been of great interest not only to corpus linguists but also to cognitive linguists and psycholinguists. It has been empirically demonstrated that multi-word units are pervasive in our languages (cf. Wray and Perkins, 2000), and they are considered psychologically real when it is found that a language learner starts with formulaic phrases that serve specific functions (e.g., Ellis, 2003; Tomasello, 2003). One of the current approaches to extract multi-word units is the frequency threshold approach (cf. Wei and Li, 2013).

The frequency threshold approach reflects the argument that frequently used items are more entrenched in our mind. While many have recognized that frequency data are more useful when complemented with dispersion data (e.g., Juilland et al., 1970), the way the frequency threshold approach considers the dispersion of a multi-word unit still leaves a lot to be desired. For example, when automatically extracting multi-word units in English, Gray and Biber (2013) simply set a dispersion threshold, i.e., occurring in at least five corpus texts.

Therefore, the present study aims to probe more deeply into the interaction between the frequency data and the dispersion data of multi-word units. Both a frequency-based set of multi-word units and a dispersion-based set are automatically extracted from a Chinese corpus, and the two sets are compared. The present study adopts a more scientific method to compute the dispersion of a multi-word unit, i.e., the DP value (Gries, 2008). This method is argued to be more flexible, simple, extendable, and sensitive than previous methods (Gries, 2008:425-426). Note that given the limited resources, the present study focuses on three-word units (trigrams for short, hereafter).

The paper is organized as follows. Section 2 introduces the method of the present study. Section 3 presents the results. Section 4 is a general discussion of the findings and the implications. Section 5 highlights the contributions of the findings.

## 2 Method

The corpus for the present study is the Academia Sinica Balanced Corpus of Modern Chinese (the Sinica Corpus, hereafter).[1] The fourth edition contains 11,245,853 tokens.

This work is licenced under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: http://creativecommons.org/licenses/by/4.0/



*Zock/Rapp/Huang (eds.): Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon*, pages 99–104,
Dublin, Ireland, August 23, 2014.

This study adopted a fully inductive approach to identify trigrams in Chinese. A computer program run in R automatically retrieved trigrams in the Sinica Corpus. Each trigram did not cross a punctuation boundary in a written text or a turn boundary in a spoken text. Then, the raw occurrence of each trigram was counted, and the raw occurrence was also normalized to the relative frequency in one million words. A frequency threshold was set to be 5 occurrences in one million words, and 1,279 trigrams passed the threshold. For each of them, the dispersion value was calculated.[2]

Regarding the dispersion value, the present study adopted Gries's (2008) measure. First, the corpus was roughly evenly divided into ten parts. Next, the raw occurrence of each trigram in each part was counted. Then, the dispersion value was calculated as shown in Table 1. Take the trigram *shi yi ge* 'be a CLASSIFIER', for example. Given that the first part of the corpus (1,081,955 tokens altogether) accounts for 9.6% of all the corpus data, the raw occurrences of *shi yi ge* in the first corpus part should also account for 9.6% of its overall occurrences. However, the observed frequency of *shi yi ge* in the first part (405/2,931 = 13.8%) was found to be slightly higher than its expected frequency. The absolute difference for each corpus part (shown in the third column) was summed up (shown in the fourth column), and the sum was divided by 2. The figure in the fifth column was the dispersion value for the trigram *shi yi ge*. The dispersion value always falls between 0 and 1: the lower the value is, the more evenly dispersed the trigram is in the corpus.

| Expected Percentage **(A)** | Observed Percentage **(B)** | Absolute Difference **(C) = (A)-(B)** | Sum of Absolute Differences **(D)** | Divided by 2 **(E) = (D)/2** |
|---|---|---|---|---|
| 1,081,955/11,245,853 = 0.096 (9.6%) | 405/2,931 = 0.138 (13.8%) | \|0.096 - 0.138\| = 0.042 | 0.257 | 0.1285 |
| 1,018,642/11,245,853 = 0.091 | 202/2,931 = 0.069 | \|0.091 - 0.069\| = 0.022 | | |
| 1,163,099/11,245,853 = 0.103 | 283/2,931 = 0.097 | \|0.103 - 0.097\| = 0.006 | | |
| 1,023,536/11,245,853 = 0.091 | 388/2,931 = 0.132 | \|0.091 - 0.132\| = 0.041 | | |
| 1,050,833/11,245,853 = 0.093 | 408/2,931 = 0.139 | \|0.093 - 0.139\| = 0.046 | | |
| 1,214,233/11,245,853 = 0.108 | 224/2,931 = 0.076 | \|0.108 - 0.076\| = 0.032 | | |
| 1,132,756/11,245,853 = 0.101 | 287/2,931 = 0.098 | \|0.101 - 0.098\| = 0.003 | | |
| 1,185,658/11,245,853 = 0.105 | 164/2,931 = 0.056 | \|0.105 - 0.056\| = 0.049 | | |
| 1,200,826/11,245,853 = 0.107 | 313/2,931 = 0.107 | \|0.107 - 0.107\| = 0 | | |
| 1,174,315/11,245,853 = 0.104 | 257/2,931 = 0.088 | \|0.104 - 0.088\| = 0.016 | | |

Table 1. Computation of the dispersion value of the trigram *shi yi ge* 'be a CLASSIFIER'.

After the dispersion value for each of the 1,279 trigrams was calculated, the top 300 trigrams in the frequency-based list and the top 300 trigrams in the dispersion-based list were further analyzed

[1] The fourth edition of the Sinica Corpus is currently available at http://asbc.iis.sinica.edu.tw/. For more information about the Sinica Corpus, refer to http://app.sinica.edu.tw/kiwi/mkiwi/98-04.pdf.

[2] The present study aims to compare frequency-based and dispersion-based trigrams, and the best way would be to compute the dispersion value for *all* the trigrams automatically extracted from the corpus. This, however, seems to be too difficult because this approach could be resource-intensive. Therefore, the present study set a frequency threshold to obtain a computationally reasonable number of trigrams, and computed the dispersion value for each trigram that passed that frequency threshold. Actually, the frequency threshold of the present study is relatively low.

manually.[3] Each of them were then manually coded based on the form. There are five categories, as shown in Table 2.

| Category | Definition | Example |
|---|---|---|
| verb trigrams | trigrams that contain at least one verb | *you ren shuo* 'have person said' |
| finite trigrams | trigrams that contain a copula (e.g., *shi* 'be') and/or a modal verb (e.g., *hui* 'can'), but not a verb | *zhe ye shi* 'this is also' |
| content word trigrams | trigrams that contain at least one content word (i.e., a noun, an adjective, and an adverb), but not a verb or a finite | *shi nian qian* 'ten years ago' |
| function word trigrams | trigrams that contain only function words | *ling yi ge* 'another one CLASSIFIER' |
| incomprehensibly incomplete trigrams | trigrams that are structurally and/or semantically incomplete and incomprehensible | *bu yi bu* 'step one step' |

Table 2. Categories for trigrams.

## 3 Results

The total numbers of trigram types at different frequency thresholds (per one million words) are presented in Table 3. The following discussions will center around trigrams that occur five or more times per one million words.

Table 3. The total numbers of trigram types at different frequency thresholds (per one million words).

| Frequency Threshold | Trigram Types |
|---|---|
| > 1 time per one million words | 15,655 |
| **> 5 times per one million words** | **1,279** |
| > 10 times per one million words | 422 |
| > 40 times per one million words | 35 |

Figure 1 presents the frequency distribution (per one million words) of the 1,279 trigrams, which occur five or more times. Among all the trigrams here, the most frequent one is *shi yi ge* 'be one CLASSIFIER' (260.62 times per one million words), and the least frequent one is *zhongyao de shi* 'important DE thing' (5.07 times).

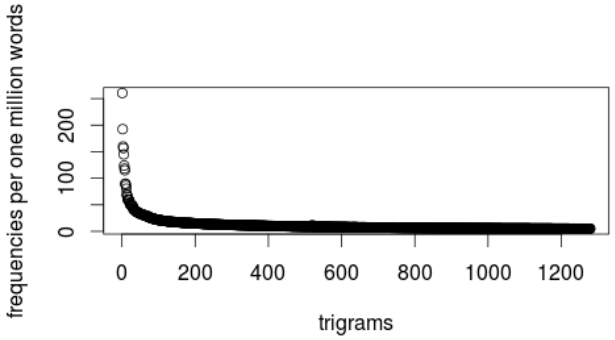


Figure 1. The frequency distribution (per one million words) of the 1,279 trigrams.

[3] The number of trigrams for further analysis was determined for convenience, with a view to yielding a manageable set of trigrams to be hand-coded.

Figure 2 presents the distribution of the dispersion values of the 1,279 trigrams, which occur five or more times. Among all the trigrams here, the best-dispersed one is *zhe ye shi* 'this also be' (0.0375), and the one with the highest dispersion value is *kaifang kongjian zhi* 'open space ZHI' (0.9085). As Figure 2 shows, the majority of trigrams are quite well-dispersed across the corpus (i.e., most of the dispersion values are lower than 0.4).

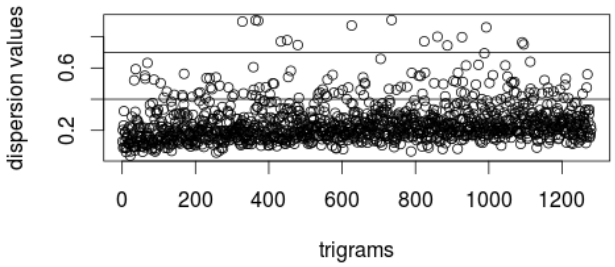


Figure 2. The distribution of the dispersion values of the 1,279 trigrams.

When zooming in to examine the top 300 trigrams in the frequency-based list and the top 300 trigrams in the dispersion-based list, we can find that there is an overlap of 126 trigrams (only 42%) between the two list. Table 4 summarizes the category distributions of the top 300 trigrams in the frequency-based list, the top 300 trigrams in the dispersion-based list, and the 126 trigrams in the overlap.

| **Category** | **Frequency-based** | | **Dispersion-based** | | **Overlapping** | |
|---|---|---|---|---|---|---|
| content word trigrams | 132 | 44.0% | 125 | 41.7% | 55 | 43.7% |
| finite trigrams | 58 | 19.3% | 62 | 20.7% | 31 | 24.6% |
| verb trigrams | 38 | 12.7% | 42 | 14.0% | 13 | 10.3% |
| function word trigrams | **42** | **14.0%** | **23** | **7.7%** | 12 | 9.5% |
| incomprehensibly incomplete trigrams | **30** | **10.0%** | **48** | **16.0%** | 15 | 11.9% |
| TOTAL | 300 | 100% | 300 | 100% | **126** | 100% |

Table 4. The category distributions of the top 300 trigrams in the frequency-based list, the top 300 trigrams in the dispersion-based list, and the 126 trigrams in the overlap.

## 4 Discussion

Overall, whether from the frequency-based perspective or from the dispersion-based perspective, content word trigrams are the most dominant. This is not too surprising, for this category covers a wide range of word classes (i.e., nouns, adjectives, and adverbs). In Chinese, finite trigrams are also frequent and well-dispersed, perhaps because the finite serves many interpersonal metafunctions (i.e., expressing the polarity of a sentence/utterance) (Thompson, 1996). In this category, *shi* 'be' is the most frequent. The main difference between the frequency-based approach and the dispersion-based approach is that the former extracts more function word trigrams, while the latter extracts more incomprehensibly incomplete trigrams.

Additionally, as shown in Table 4, the overlap between the two approaches is not very large (i.e., 126/300 = 42%). Now, consider Table 5.

| Top *n* trigrams in the two lists | Overlap | |
|---|---|---|
| 300 | 126/300 = | 42% |
| 500 | 262/500 = | 52.4% |
| 700 | 438/700 = | 62.6% |
| 1,000 | 798/1,000 = | 79.8% |
| 1,279 | 1,279/1,279 = | 100% |

Table 5. The overlap between the frequency-based approach and the dispersion-based approach.

Since the trigrams in the two lists are the same, the overlap should be getting larger as *n* is getting larger. However, even when *n* reaches 700, the overlap between the two approaches is only slightly more than half. This suggests that when a certain type number is set (e.g., 300, 500, or 700), the frequency-based approach and the dispersion-based approach can extract quite different sets of trigrams.

Some may argue that the frequency-based approach is more useful because it extracts fewer incomprehensibly incomplete trigrams (cf. Table 4). On the other hand, we can also see the dispersion value as an ancillary measure to the relative frequency, just as the standard deviation is usually presented whenever a mean is presented. Frequencies can be regarded as an important dimension of the sum of one's linguistic experience (cf. Bybee, 2006), and dispersion data should also be considered so. Items, whether single words or multi-word units, that achieve a high frequency *and* are well-dispersed across the corpus should be much more entrenched in the mental lexicon, for their frequent occurrences are ubiquitous, not just in certain text types. These items should deserve more attention from linguists and may be more useful in a language resource.

Moreover, the findings of the present study have demonstrated the dynamic nature of lexical representations. When different measures (e.g., the relative frequency, the dispersion value) are used, the ranking of a trigram may change dramatically. If the association measure is also taken into account or different measures are integrated in a certain way, another picture of trigrams in Chinese may emerge. This echoes Biber's (2009) suggestion that in the extraction of multi-word units, no methodology should be considered to be correct. That is to say, different sets of multi-word units extracted by different approaches can all be useful in one way or another and reflective of some aspects of our cognition. However, those ranking high in all the approaches may be at the core of our mental lexicon.

The implications of a list of trigrams (or other multi-word units) in Chinese can be pinpointed as follows. First, most dictionaries in Chinese compile words, but a dictionary of multi-word units in Chinese can also be of great use. For example, the trigram *you yi ge* 'have one CLASSIFER' is usually used to introduce a new topic in discourse, and this needs to be included in a dictionary in Chinese. Second, such useful sequences as *you yi ge* can also be included in teaching materials for language learners. Third, we can try to use automatically extracted sequences to build a language/lexical resource like WordNet (Miller et al., 1990). In such a resource (i.e., perhaps something like the Net of Multi-word Units), multi-word units in Chinese can be organized according to words or characters contained in them or even according to their discourse functions (and perhaps in some other creative ways), and lexical relations between multi-word units can be coded.

## 5 Conclusion

The contribution of the present study is twofold. Methodologically speaking, this study adopts a more sensitive dispersion measure (i.e., Gries, 2008) instead of setting an arbitrary dispersion threshold (e.g., occurring in at least five corpus files), and demonstrates that dispersion data are needed in the automatic extraction of multi-word units since those ranking high in a frequency-based list are not necessarily at the top of a dispersion-based list. It is argued that the dispersion of a multi-word unit, together with its frequency, can contribute to the entrenchment of the item in the mental lexicon because the dispersion measure reveals where a language user is confronted with the item. Practically speaking, trigrams in the overlap between the frequency-based approach and the dispersion-based approach may be at the core of the Chinese lexicon and can serve as a point of departure for future linguistic studies and resources in Chinese.

The present study can be extended in the following directions. First, some evaluations from psycholinguistic experiments are needed to further examine the role of frequency data and dispersion data in the mental lexicon. Second, the same method can be adopted to automatically extract multi-word units in different genres, and the results will be helpful for genre studies.

## References


Alison Wray and Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication*, 20(1), 1-28.

Alphonse Juilland, Dorothy Brodin, and Catherine Davidovitch. *Frequency dictionary of French words*. The Hague: Mouton de Gruyter.

Bethany Gray and Douglas Biber. 2013. Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, 18(1):109-135.

Douglas Biber. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3):275-311.

Geoff Thompson. 1996. *Introducing Functional Grammar*. London; New York: Arnold.

George A. Miller, Richard Beckwith, Christiane Fellbuam, Derek Gross, and Katherine Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235-244.

Joan Bybee. 2006. From usage to grammar: The mind's response to repetition. *Language*, 82(4):711-733.

Michael Tomasello. 2003. *Constructing a Language: A Usage-based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Naixing Wei and Jingjie Li. 2013. A new computing method for extracting contiguous phraseological sequences from academic text corpora. *International Journal of Corpus Linguistics*, 18(4):506-365.

Nick C. Ellis. 2003. Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty and M. H. Long (Eds.), *Handbook of Second Language Acquisition*. Oxford: Blackwell. (pp. 33-68)

Stefan Th. Gries. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4):403-437.