Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Co-Varying Collexeme Analysis of Chinese Classifiers 棵 *kē* and 株 *zhū*

Aneta Dosedlová
Masaryk University, Brno, Czechia

Wei-lun Lu
Masaryk University, Brno, Czechia

Abstract The numeral classifier is a grammatical category in plenty of East Asian languages, with Chinese being one of the most widely reported. In Chinese, there are many classifiers that are near-synonymous, meaning that certain classifiers may be interchangeable in certain contexts. However, these classifiers are used with semantically similar nouns and, as a result, the distinction between the various usages is not always clear. In view of this issue, we propose to study near-synonymous classifiers using the co-varying collexeme method and the Euclidean distance, by exploring the case of the classifiers 棵 *kē* and 株 *zhū*. We report results that not only partially confirm but also complement what has been found in previous raw-frequency-based research.

Keywords Categorization. Collostructional analysis. Co-varying collexeme analysis. Eluclidean distance. Near-synonymy. Prototype.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access
Submitted 2020-04-17 | Accepted 2020-10-14 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/007

1 Near-Synonymy. What It Is and the State of the Art[1]

The linguistic issue of near-synonymy is never an easy one. For decades, there have been different approaches trying to discuss and settle how different words have similar meanings and in what situations they do, based on conceptual semantic discussions, usage dictionaries, or a scrutiny of a body of linguistic samples. Among the numerous types of efforts, recent decades have witnessed the rise of corpus linguistics, which offers a methodological opportunity to approach linguistic phenomena in a way that can be faithful to how a word is actually used in real-world context. Based on the principle that one should "know a word by the company it keeps" (Firth 1957, 11), there have been numerous studies applying such rubric in the study of lexical semantics, generalising the contextual information over a number of usages of a particular word, in order to understand the lexical and grammatical company kept by the word at issue.

In corpus linguistics, there are several methods used to study similar and potentially confusing words, with the one most relevant to the present study being *collostructional analysis* (Stefanowitsch, Gries 2003; Schmid 2010; Schmid, Küchenhoff 2013), which is a family of corpus-based quantitative methods that helps measure mutual attraction between lexemes and constructions. Collostructional methods do not simply rely on numbers of lexical frequencies, but also measure the degree of probability that the patterns of analysed frequencies are due to chance. Such analyses work under the rubrics of *construction grammar* (Goldberg 1995), which claims that lexical and grammatical constructions are symbolic form-meaning pairings.[2] Collostructional analyses compare the strength of association between the analysed constructions and the chosen lexical elements in the actual use found in linguistic corpora.

In the present study, we employ the collostructional method called *co-varying collexeme analysis* (Stefanowitsch, Gries 2005;

[1] The completion of this paper was partially supported by the grant "The influence of socio-cultural factors and writing system on perception and cognition of complex visual stimuli" (GC19-09265J), of which the second Author is a member. The analysis of this paper is based on the raw data obtained from the first Author's master's thesis research. We especially thank Dr. Alvin Cheng-hsien Chen for his kind advice on the statistical methods used in this paper. Thanks also go to the editors of the volume and the anonymous reviewers. All correspondences and requests for reprints should be addressed to the second Author at wllu@med.muni.cz.
Author contributions: both Authors conceptualised the study (main responsibility being with the first Author). The data collection and annotation were done by the first Author. All the sections were jointly written by both Authors.

[2] Interested readers are referred to an overview of the position of synonymy research within Cognitive Linguistics in Glynn 2014.

Tang 2016), due to the nature of the linguistic phenomenon that we investigate. We will return to this point in § 3.

2 Classifier Constructions in Chinese and Their Near-Synonymy

Classifiers are linguistic devices that help humans categorise objects in the world. In language, classifiers are words that encode "salient perceived or imputed characteristic of the entity to which the associated noun refers" (Allan 1977, 285). Tai (1994) takes a similar stance and argues that Chinese classifiers are used to denote a group of perceptually- or functionally- based attributes associated with a given noun. Among all the systems of classifiers, the numeral classifier system is one of the most commonly recognised type (Aikvenhald 2003; Saalbach, Imai 2012). The usage of numeral classifiers is mostly compulsory with counting objects in a classifier language, which is also the case for Chinese. In a classifier language, a typical classifier construction consists of a numeral, a classifier, and a noun (Allan 1977, 288). In Chinese, the grammatical schema of such construction is [QUAN]-[CLF]-[N], exemplified by (1) below.[3]

1. 一只狗
yī *zhī* *gǒu*
one CLF dog
'one dog'

The choice of a numeral classifier is never random but is based on the perceived properties of the head noun (Tai 1994; Jiang 2017). For the choice of a classifier in a usage like (1), when a speaker of Chinese (or a learner of Chinese as a second language) expresses the quantity of a noun such as 狗 *gǒu*, the noun needs to take a suitable classifier from the conceptual category of ANIMACY[4] that captures the imputed characteristics associated with DOG. As there are multiple classifiers in each linguistic category and as some of them overlap in meaning, by using a classifier, the speaker *profiles* (Langacker 2008, 66) a perceptual or a functional aspect of the noun. For instance, the classifiers for PLANT 棵 *kē* and 株 *zhū* are near-synonymous and interchangeable in certain contexts, as exemplified by (2a) and (2b) (cited from Dosedlová, Lu 2019, 115).

[3] The glosses in this paper follow the general guidelines of the Leipzig Glossing Rules, with the addition of LK = 'linker'. Further in-text abbreviations include: N = 'noun'; QUAN = 'quantifier'.

[4] We follow the typographic convention in Cognitive Linguistics, which uses lower caps to represent a concept.

2. a. 爸爸买了两棵巨大的圣诞树
bàba *mǎi-le* *liǎng-kē* *jùdà-de* *shèngdàn-shù*
father buy-PFV two-CLF big-LK Christmas-tree
‘Father bought two huge Christmas trees’.

b. 爸爸买了两株巨大的圣诞树
bàba *mǎi-le* *liang-zhū* *jùdà-de* *shèngdàn-shù*
father buy-PFV two-CLF big-LK Christmas-tree
‘Father bought two huge Christmas trees’. (constructed from (2a))

In their study, Dosedlová and Lu argue that 棵 *kē* and 株 *zhū* conceptually profile slightly different aspects of PLANT – by observing the span of nouns the classifiers co-occur with, the authors report that 株 *zhū* occasionally co-occurs with nouns of PLANT that invoke SMALL and VULNERABLE, such as 苗 *miáo* ‘seedling’ and 花 *huā* ‘flower’, and nouns of MICRO-ORGANISM, such as 霉 *méi* ‘mold’, 细菌 *xìjùn* ‘bacterium’, 病毒 *bìngdú* ‘virus’, and so on, but that pattern is not seen among the nouns that co-occur with 棵 *kē* as a classifier. However, a methodological insufficiency of that paper is that the observations are based merely on separate raw frequency counts of *each* of the slots in the classifier construction, while no attention is paid to how the multiple slots in the construction interact.[5] Therefore, to investigate the interaction between different slots within a construction, an alternative must be sought.

From an onomasiological point of view, it will be useful to find out the interaction and the detailed relationship between the classifier and the noun within [QUAN]-[CLF]-[N]. Therefore, we would like to focus on how the two slots in that particular construction (and *only* in that particular construction, *not elsewhere* in the language/corpus) co-vary. After all, a word with classifier as part of its syntactic function may occur in various grammatical constructions in Chinese, which is the case for 只 (also as an adverb when pronounced as *zhǐ* or as a noun when pronounced as *zhī*), 棵 *kē* (also as a noun), and 株 *zhū* (also as a noun or a verb), among numerous others, but that is something we would certainly like to exclude in order to achieve a more statistically-precise result. For this purpose, we consider it suitable to conduct the so-called co-varying collexeme analysis. Such an analysis always *begins with a construction* and studies which lexemes tend to be attracted to that particular construction and which do not. A typical collostructional analysis relies on frequency measures of tokens of different types of lexemes extracted from a corpus. Once obtained from the language sample, the frequencies are

[5] A similar general observation from studies done in cognitive semantics is made in Stefanowitsch, Gries 2005, 1.

used for calculating the *p*-values of the list of collexemes (lexemes that may be attracted to a particular construction), which show the degree of association between the collexemes and the construction. Each lexeme analysed has its own *p*-value, which indicates its collocational strength with the construction. The calculation is done via the Fisher-Yates Exact test.

3 Co-Varying Collexeme Analysis and Euclidean Distance

In a co-varying collexeme analysis, it is important to identify the association strength between pairs of lexical items appearing in two different slots of the same construction. In our study, the lexical slots to examine are the CLF and the N within the [QUAN]-[CLF]-[N] construction. To conduct such an analysis, we first need to find out the span of lexemes that may occur in each of the slots investigated. We also need the frequency of the construction (C) investigated (which is the total number of concordance lines included in the sample), the frequency of the first target word (L1) in a particular slot (S1) in C in the sample, and the frequency of the second target word (M1) in the other slot (S2) in C in the sample. A template is shown in table 1 below.

Table 1 A schematic distribution table for a co-varying collexeme analysis (adapted from Stefanowitsch, Gries 2005, 9)

	M1 in S2 of C	**Other words (M2, M3…) in S2 in C**	**Total**
L1 in S1 in C	frequency of S2(M1) and S1(L1) in C	frequency of S2(⌐M1) and S1(L1) in C	total frequency of S1(L1) in C
other words (L2, L3…) in S1 in C	frequency of S2(M1) and S1(⌐L1) in C	frequency of S2(⌐M1) and S1(⌐L1) in C	total frequency of S1(⌐L1) in C
total	total frequency of S2(M1) in C	total frequency of S2(⌐M1) in C	total frequency of C

We illustrate such a template with the case study of the distribution of the causing event and the resulting event in the English *into* causative (Stefanowitsch, Gries 2005), as in *we must not fool ourselves into thinking there is no longer any problem*. To determine the extent of the correlation between *fool* (as the causing event) and *think* (as the resulting event) in *fool into thinking*, a distribution table for this pair of lexemes is given in table 2.

Table 2 Information needed for studying the correlation between *fool* and *think* in *fool into thinking* (Stefanowitsch, Gries 2005, 10)

	think	**Other verbs**	**Total**
fool	46 (7)	31 (70)	77
Other verbs	101 (140)	1,408 (1,369)	1,509
Total	147	1,439	1,586

Such a table is submitted to a contingency test and the whole procedure is done for *each* word pair appearing in the construction in question. The data of the tables is submitted to Fisher-Yates Exact test. The result of this test is a *p*-value that indicates the association strength between the lexeme and the construction. The strongest mutual association between a lexeme and a construction is the one with the smallest *p*-value (Desagulier 2014, 157). Co-varying collexemes are those pairs of words that co-occur more frequently than by pure chance (Stefanowitsch, Gries 2003, 2005). The final result can be submitted to further analysis, such as *cluster analysis* (Divjak 2010; Divjak, Fieller 2014), for a more detailed understanding of the results. Table 3 shows the information needed for studying the correlation between a classifier and the noun in [QUAN]-[CLF]-[N].

Table 3 Information needed for studying the correlation between CLF and N in [QUAN]-[CLF]-[N]

	CLF1 in s1 in [QUAN]-[CLF]-[N]	**Other words (CLF2, CLF3…) in s1 in [QUAN]-[CLF]-[N]**	**Total**
N1 in s2 in [QUAN]-[CLF]-[N]	frequency of s1(CLF1) and s2(N1) in [QUAN]-[CLF]-[N]	frequency of s1(⌐CLF1) and s2(N1) in [QUAN]-[CLF]-[N]	total frequency of s2(N1) in [QUAN]-[CLF]-[N]
other words (N2, N3…) in s2 of [QUAN]-[CLF]-[N]	frequency of s1(CLF1) and s2(⌐N1) in [QUAN]-[CLF]-[N]	frequency of s1(⌐CLF1) and s2(⌐N1) in [QUAN]-[CLF]-[N]	total frequency of s2(⌐N1) in [QUAN]-[CLF]-[N]
total	total frequency of s1(CLF1) in [QUAN]-[CLF]-[N]	total frequency of s1(⌐CLF1) in [QUAN]-[CLF]-[N]	total frequency of [QUAN]-[CLF]-[N]

Cluster analysis is a family of statistical methods used for deciding the distance and similarities between entities, which may be applied to the study of language to measure the internal structure of a set of synonymous lexical constructions. Divjak and Gries (2006), for instance, study nine Russian verbs that all share the tentative meaning of TRY. The paper examines 1,585 concordance lines by tagging the individual usages using morphosyntatic cues that may influence the behavioural profile of the nine verbs. The authors find that the nine verbs form three groups and that each group exhibits similar internal behaviours, which means that the members in a group have smaller conceptual semantic distances with each other than with members outside the group.

The first step in conducting a cluster analysis is to choose the variables. There are several kinds of variables to choose from, which can be numerical, categorical, or ordinal.[6] We illustrate this with a simplified example below. Let us suppose we have four constructions (C1, C2, C3 and C4) to analyse. We also assume there are four possible variables that may factor in learning about the conceptual semantic distance between the four words, including: frequency in the corpus, co-occurrence with Word *x*, co-occurrence with Word *y*, and co-occurrence with an adjective. The hypothetical situation is put forth in table 4.

Table 4 A possible scenario with four constructions and four variables for a cluster analysis

	C1	**C2**	**C3**	**C4**
frequency in corpus	379	254	468	342
co-occurrence with *x*	257	159	374	285
co-occurrence with *y*	53	49	85	62
co-occurrence with adjective	81	37	103	64

The next step is to decide on a method for calculating the similarities among the words involved. In a cluster analysis, one of the most common methods for calculating distances (similarities) is *Euclidean distance*. The result of such method is a dissimilarity matrix table, which shows the distances among all the entities within a dataset.

The Euclidean distance between two objects is gained by summing the squared differences between the pairs of corresponding values for the two individuals and taking the square root of the sum (Divjak, Fieller 2014, 417). The formula for the calculation of Euclidean distance is as follows:

$$d_{y} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^{2}}$$

[6] Interested readers are referred to Divjak, Fieller 2014 for a detailed discussion on how to choose the variables.

Following the hypothetical situation outlined in table 4, a Euclidean distance analysis can be conducted using the above formula for the set of the target words. For instance, the similarity distance between C1 and C2 can be figured out as follows:

$$d_{c1c2} = \sqrt{(379-254)^2 + (257-159)^2 + ¿\,(53-49)^2 + (81-37)^2} = 164.9¿$$

The same can be done between each two of the four: the results are summarised in table 5. The lowest number in each column in bold indicates the smallest distance (or the highest degree of similarity) between words. As the table shows, the closest items are C1 and C4, with a distance of 50.23 (underlined, in bold), and the most dissimilar items are C2 and C3, with a distance of 312.5 (underlined only).

Table 5 Summarised result of the Euclidean distances based on table 4

	C1	**C2**	**C3**	**C4**
C1	0	164.9	152.0	**50.23**
C2	164.9	0	312.5	156.6
C3	152.0	312.5	0	160.8
C4	**50.23**	156.6	160.8	0

Having introduced the related statistical algorithms, now we move on to a detailed description of the research issues and the research steps.

4 Research Issue, Scope, and Steps

In this paper, we address the following issues: first of all, what can we learn about the relationships between a pair of synonymous classifiers using a co-varying collexeme analysis? In what way does the Euclidean distance help? We believe that the relationships between the synonymous classifiers can be made available based on the nouns that collocate with each of these classifiers and that a co-varying collexeme analysis will provide useful data related to the behaviour of the classifiers involved, including the collocational strength and certain association measures. Such results are what we may further submit for a cluster analysis in order to explicate the internal structure of the synonymous set. Secondly, does the co-varying collexeme analysis and an analysis based on the Euclidean distance tell us anything beyond an analysis informed only by a raw frequency count of the lexical items in question?

To answer the questions above, we chose to investigate the classifiers 棵 *kē* and 株 *zhū*, which had already been examined based on a raw frequency approach in Dosedlová and Lu (2019). In that paper, the authors used data extracted from Sketch Engine[7] and observed the types of nouns that occurred in their language sample, and the token frequencies of each of the nouns, which allowed the authors to come up with the conceptual similarities and differences between the two classifiers. In order to see how a different methodological approach may shed alternative light on the same linguistic phenomenon, we extracted the collocating nouns and analysed the data to calculate their T-score, MI score and logDice. After that, we calculated the Euclidean distance between the nouns in the dataset. The steps are outlined below.

In order to properly sample the usages of each of the classifiers investigated, we built a corpus for each of the classifiers by extracting random concordance lines from a large representative body of authentic linguistic data. To this end, we used the function 'sample' of Sketch Engine, which created a random collection of concordances that involved the two target classifiers. We set the size of each subcorpus five hundred lines, which was more than sufficient to investigate the semantics of a common word.[8] After we input the extracted data to Excel, we went through the data manually to look for the collocating nouns and their frequencies in the sub-corpora. In addition, we looked up the frequencies of each of the collocating nouns in each of the sub-corpora. All the information acquired from the above steps was used to calculate the association measures and collocational strengths in the co-varying collexeme analysis. These association measures included: 1) T-score, which indicates the level of certainty with which one can argue for a clear association between the linguistic units analysed. A T-score higher than 2 is seen as statistically significant, which means that the co-occurrence of the two linguistic units is more than mere chance. 2) logDice, which is a measure of the typicality of the co-occurrence of the classifier and its collocating noun. The maximum logDice value is 14, which means the exclusive collocation between the linguistic units investigated (that all occurrences of X co-occur with Y and vice versa). A negative value means that the XY collocation is not statistically significant. 3) MI score, which stands for the extent to which words co-occur compared to the frequency of their separate appearance. An MI score higher than 3 is an indicator of a statistically significant collocation. The lower the MI score, the more likely the linguistic units co-occur only by chance.

[7] https://www.sketchengine.eu.

[8] Sinclair (2005) claims that it takes around 20 tokens to determine the meaning of a not particularly complicated lexeme and around 50 tokens for an average lexeme.

The three association measures may or may not converge, as we will show in the body of the analysis.

5 Results

In this section, we report the findings based on the data retrieved from Sketch Engine following the steps outlined above.

5.1 Nouns in [QUAN]-[*kē*]-[N]: Their T-Score and logDice

In the sub-corpus of 棵 *kē*, we found 38 different nouns that co-occurred with the classifier. Below, we discuss the association measures of T-score and logDice.

It is important to bear in mind that each of these measures takes a different approach in measuring the strength of the collocation. If we look at the most frequent noun collocating with 棵 *kē*, i.e 树 *shù* 'tree', its T-score and logDice are the highest among all collocating nouns, but its MI score is not. The reason is that the MI score is strongly influenced by the size of the corpus, hence it is usually considered subsidiary if compared to the T-score. As for the T-score, it promotes pairings that are frequently observed but does not concern the total frequencies of each of the linguistic units, hence the size of the corpus is irrelevant. For instance, if we look at the noun 木棉树 *mùmiánshù* 'cotton tree', the T-score is relatively low because there are only three tokens of its collocation with 棵 *kē*, but the MI score is quite high, as the MI score takes into account all the other occurrences of both of the words. As for the logDice, it is an important indicator of the typicality of a collocation.

Therefore, in this study, T-score and logDice are our main foci. Table 6 lists the first five nouns with the highest T-score and the highest logDice in the sub-corpus of 棵 *kē*.

Table 6 Top five collocations with 棵 *kē* in terms of T-score and logDice

Noun	T-score	Noun	LogDice
树 *shù* 'tree'	16.3200	树 *shù* 'tree'	5.7562
树木 *shùmù* 'tree-wood'	3.9987	杨树 *yángshù* 'poplar'	3.4276
杨树 *yángshù* 'poplar'	3.2991	树木 *shùmù* 'tree-wood'	3.2250
树苗 *shùmiáo* 'tree seedling'	3.1353	树苗 *shùmiáo* 'tree seedling'	3.2247
果树 *guǒshù* 'fruit tree'	3.0853	果树 *guǒshù* 'fruit tree'	2.8927

As we see in table 6, the two association measures largely overlap and jointly confirm the status of 树 *shù*, 杨树 *yángshù*, 树木 *shùmù*, 树

苗 *shùmiáo*, and 果树 *guǒshù* being statistically significant collocates of 棵 *kē*. 树 *shù* is the most significant lexeme attracted to [QUAN]-[*kē*]-[N], based on the T-score and the logDice.

5.2 Nouns in [QUAN]-[*zhū*]-[N]: Their T-Score and logDice

The same analysis was done with the nouns that co-occurred with 株 *zhū*. In the sub-corpus, there are 75 different nouns found to co-occur with 株 *zhū*. We also calculated the T-score and the logDice for each of the nouns, now listing the top five in terms of the T-score and the logDice in table 7.

Table 7 Top five collocations with 株 *zhū* in terms of T-score and logDice

Noun	T-score	Noun	LogDice
树 *shù* 'tree'	13.4313	苗 *miáo* 'seedling'	6.0427
苗 *miáo* 'seedling'	10.9546	树 *shù* 'tree'	5.1596
花 *huā* 'flower'	9.2243	植树 *zhíshù* 'plant-tree'	4.4780
植树 *zhíshù* 'plant-tree'	6.3984	菌 *jùn* 'bacteria'	4.4602
苗木 *miáomù* 'seedling'	6.0901	苗木 *miáomù* 'seedling'	4.4198

As we can see in table 7, the top five collocates in terms of each of the association measures still largely overlap, which confirms the status of 树 *shù*, 苗 *miáo*, 植树 *zhíshù*, and 苗木 *miáomù* as the most statistically significant lexemes that are attracted to [QUAN]-[*zhū*]-[N].

However, if we compare all the five most significant collocates between the two classifiers in the corpora, we see that 棵 *kē* generally collocates with nouns that contain 树 *shù* as part of it, whereas the significant collocates of 株 *zhū* are more diversified (that is, do not necessarily involve 树 *shù* as part of the lexeme). In addition, 株 *zhū* has collocates that invoke SMALL and VULNERABLE, such as 苗 *miáo*, 花 *huā*, and 菌 *jùn*. We will return to this point when we compare the results from this co-varying collexeme analysis with the results in Dosedlová and Lu (2019).

A comparison of tables 6 and 7 allows us to identify 树 *shù* as the lexeme that appears in both tables, meaning that it is the lexeme that has the highest T-score and logDice in both [QUAN]-[*kē/zhū*]-[N], indicating the strongest attraction between 树 *shù* and the two classifier constructions. Based on this fact, we may say that 树 *shù* is the prototypical lexical instantiation of PLANT that collocates with both 棵 *kē* and 株 *zhū* (but only within the particular construction of [QUAN]-[CLF]-[N] and only when it co-varies with 棵 *kē* and 株 *zhū*, rather than in Chinese in general). In addition to 树 *shù*, 苗 *miáo* is also a lexeme that has a very high T-score and logDice in [QUAN]-[*zhū*]-

[N], so is another prototypical lexical instantiation of PLANT in that classifier construction. We will return to this point in our discussion.

5.3 A Cluster Analysis of Nouns within [QUAN]-[*kē*/*zhū*]-[N]

After we obtained the association measures, we further submitted the numbers to a cluster analysis based on the Euclidean distance. In the analysis we used the same corpora, where we first identified the nouns that collocated with both of the classifiers. There are fourteen of such nouns, which includes 树 *shù* 'tree', 槐树 *huáishù* 'Chinese scholar tree', 果树 *guǒshù* 'fruit tree', 杨树 *yángshù* 'poplar tree', 植树 *zhíshù* 'plant-tree', 松树 *sōngshù* 'pine tree', 柳树 *liǔshù* 'willow', 树木 *shùmù* 'tree-wood', 林木 *línmù* 'forest', 银杏 *yínxìng* 'ginkgo', 柳杉 *liǔshān* 'Japanese cedar', 核桃 *hétáo* 'walnut', 樱花 *yīnghuā* 'cherry blossom', 玉米 *yùmǐ* 'corn', and 桂花 *guìhuā* 'osmanthus'.

Secondly, we calculated the Euclidean distance between the fourteen nouns that co-occurred with 棵 *kē* and 株 *zhū* within the construction [QUAN]-[CLF]-[N], following the formula introduced in § 3 and using the raw frequency, T-score, MI value and logDice of the fourteen lexemes as the possible variables. A summary of the Euclidean distances is given as table 9.

Table 9 Euclidean distances between pairs of the fourteen nouns co-occurring with 棵 *kē* and 株 *zhū* within [QUAN]-[CLF]-[N]

	shù	***huáishù***	***guǒshù***	***yángshù***	***zhíshù***	***sōngshù***	***liǔshù***	***shùmù***	***línmù***	***yínxìng***	***liǔshān***	***hétáo***	***yīngtáo***	***yùmǐ***	***guìhuā***
shù	0.0000	10.4612	6.4445	5.4078	3.3374	4.4432	11.7994	2.5257	8.3046	5.3465	8.7752	**14.0385**	**12.8982**	6.3567	8.4151
huáishù	10.4612	0.0000	5.3431	6.8035	9.5656	6.8840	3.0736	8.9606	6.9774	7.8378	10.7145	6.0012	2.4650	10.2923	9.9332
guǒshù	6.4445	5.3431	0.0000	1.4954	4.4740	2.0294	5.5865	4.2356	2.8454	2.4955	5.9007	7.6001	7.6979	4.9964	5.1410
yángshù	5.4078	6.8035	1.4954	0.0000	3.0277	1.1295	7.0205	2.9924	2.9934	1.0824	5.0565	8.8309	9.1825	3.6956	4.3665
zhíshù	3.3374	9.5656	4.4740	3.0277	0.0000	2.6831	10.0446	1.2045	5.4170	2.4293	5.4401	11.7935	12.0071	3.0495	5.0853
sōngshù	4.4432	6.8840	2.0294	1.1295	2.6831	0.0000	7.5671	2.2245	4.1106	1.8019	5.9873	9.6246	9.3303	4.2890	5.3452
liǔshù	11.7994	3.0736	5.5865	7.0205	10.0446	7.5671	0.0000	9.7916	5.8479	7.8344	9.5297	2.9601	3.6330	9.8310	8.7971
shùmù	2.5257	8.9606	4.2356	2.9924	1.2045	2.2245	9.7916	0.0000	5.7990	2.8214	6.4452	11.7939	11.4239	4.1497	6.0053
línmù	8.3046	6.9774	2.8454	2.9934	5.4170	4.1106	5.8479	5.7990	0.0000	3.0166	3.7724	6.7290	8.9456	4.1286	3.0044
yínxìng	5.3465	7.8378	2.4955	1.0824	2.4293	1.8019	7.8344	2.8214	3.0166	0.0000	4.1896	9.4007	10.1869	2.6200	3.5685
liǔshān	8.7752	10.7145	5.9007	5.0565	5.4401	5.9873	9.5297	6.4452	3.7724	4.1896	0.0000	9.8562	12.7176	2.4658	0.7830
hétáo	**14.0385**	6.0012	7.6001	8.8309	11.7935	9.6246	2.9601	11.7939	6.7290	9.4007	9.8562	0.0000	5.8824	10.8428	9.2455
yīnghuā	**12.8982**	2.4650	7.6979	9.1825	12.0071	9.3303	3.6330	11.4239	8.9456	10.1869	12.7176	5.8824	0.0000	12.5538	11.9492
yùmǐ	6.3567	10.2923	4.9964	3.6956	3.0495	4.2890	9.8310	4.1497	4.1286	2.6200	2.4658	10.8428	12.5538	0.0000	2.3161
guìhuā	8.4151	9.9332	5.1410	4.3665	5.0853	5.3452	8.7971	6.0053	3.0044	3.5685	0.7830	9.2455	11.9492	2.3161	0.0000

The summary in table 9 allows us to compare the Euclidean distance between all the nouns involved and the prototypical PLANT within the two particular grammatical constructions. Remember that 树 *shù* is the lexical prototype in both constructions. In table 9, we can see that among the fourteen lexemes shared by the two classifier constructions, 核桃 *hétáo* and 樱花 *yīnghuā* are the two lexemes that have the highest Euclidean distance from 树 *shù*, with a Euclidean distance value of 14.0385 and 12.8982 (in bold), respectively. This means that the behaviours of these two lexemes are the most different from the prototype in the corpora. On the other hand, the two lexemes that have the smallest Euclidean distance with 树 *shù* are 树木 *shùmù* and 植树 *zhíshù*, having a Euclidean distance value of 2.5257 and 3.3374 (underlined), respectively, meaning that the two lexemes have the most similar behaviour with 树 *shù* in the corpora. Note that the two lexemes are also conceptually closer to 树 *shù* than the other lexemes, as they do not refer to any particular type of tree, so are at the same level with 树 *shù* in terms of taxonomy. Therefore, the similar behaviour between 树 *shù*, 树木 *shùmù* and 植树 *zhíshù* is natural.

6 Discussion and Concluding Remarks

The statistically informed analysis in the present paper largely confirms the results in Dosedlová and Lu's (2019) study based on raw lexical frequencies, but it also turns up meaningful patterns that were not reported in the previous study.

In particular, based on the T-score and the logDice, we firstly confirm that 树 *shù* is the lexeme that has the strongest association measures with both [QUAN]-[*kē*]-[N] and [QUAN]-[*zhū*]-[N]. This matches the fact that 树 *shù* is the most frequent noun that co-occurs both with 棵 *kē* and with 株 *zhū* (Dosedlová, Lu 2019, 123). Following on from that, we see that the raw frequency, T-score and logDice constitute pieces of converging evidence that jointly support the claim that 树 *shù* is the prototypical lexical instantiation of PLANT in [QUAN]-[*kē*/*zhū*]-[N]. Secondly, the statistically informed analysis allows us to confirm that [QUAN]-[*zhū*]-[N] does attract nouns that invoke SMALL and VULNERABLE, such as 苗 *miáo*, 花 *huā*, and 菌 *jùn* (Dosedlová, Lu 2019, 122). In the above two respects, the results obtained via a co-varying collexeme approach echo the findings based on raw lexical frequency.

However, a co-varying collexeme analysis can build on the previous analysis and can allow us to see patterns beyond an exclusively raw-frequency-based approach – first of all, it allows us to identify 苗 *miáo* as another lexeme that is strongly associated with [QUAN]-[*zhū*]-[N]. According to the list of token frequencies in Dosedlová and Lu (2019, 123), 苗 *miáo* accounts for 14.3% of the total usages in [QUAN]-[*zhū*]-[N], but that is only less than one third of the percentage of 树 *shù* (which is 47.3% in their table). Accordingly, a study merely based on the token frequency may not give the collocation between 苗 *miáo* and 株 *zhū* too much weight. But once the T-score and the logDice are included, that brings the lexeme back to our attention. Secondly, another linguistic fact that is uncovered through the Euclidean distance is the similarity between each of the fourteen shared lexemes with the prototype 树 *shù*. For instance, the Euclidean distance analysis indicates 树木 *shùmù* and 植树 *zhíshù* to be the lexemes that are most similar to 树 *shù* in terms of the behavioural profile, which cannot be captured by a simple frequency count – that would only identify 木 *mù* and 植 *zhí* being infrequent lexical types in the corpus, about one eighth of 树 *shù* in [QUAN]-[*kē*]-[N] (Dosedlová, Lu 2019, 121) and one fourth of 树 *shù* in [QUAN]-[*zhū*]-[N] (Dosedlová, Lu 2019, 123). In addition, the cluster analysis has found the behavioural profiles of 核桃 *hétáo* and 樱花 *yīnghuā* to be the most distant from the prototype among the fourteen shared lexemes, meaning that the two lexemes behave most differently from 树 *shù* in [QUAN]-[*kē*/*zhū*]-[N], which is an observation that can be made only through a Euclidean distance analysis.

Despite of the advantages of a co-varying collexeme analysis and a cluster analysis mentioned above, we maintain and emphasise that an analysis based on type and token frequencies is still capable of uncovering linguistic facts about near-synonymy that cannot be seen through a collostructional analysis, and that the two approaches should be considered *complementary* to each other. An interesting part of the conceptual semantic difference between 棵 *kē* and 株 *zhū*, for instance, lies in the fact that [QUAN]-[*zhū*]-[N] has an extended group of usages that covers entities that do not invoke PLANT, such as MOLD, BACTERIUM, BIOLOGICAL SUBSTANCE and CHEMICAL SUBSTANCE (Dosedlová, Lu 2019, 122-3). These usages are peripheral members of the linguistic category (defined by the categorising structure [QUAN]-[*zhū*]-[N]) and are very low in lexical frequency. Such periphery of a linguistic category is typically difficult to observe given its low frequency, but may contain important conceptual information that helps define the linguistic category. Such information may become available only through an extensive type frequency analysis of the language sample.

Finally, we would like to conclude by proposing a synergy between different quantitative methods for analysing the near-synonymy of classifiers, similar to the advocacy for a methodological synthesis in Janda, Kudrnáčová and Lu (2019). As we have shown in this paper, each research method has its strengths and its limitations, so we consider it always advisable to try to obtain converging and consolidating evidence from different angles, or to try to obtain comprehensive results from complementary methodological approaches.

Bibliography

Aikvenhald, A.Y. (2003). *Classifiers. A Typology of Noun Categorization Devices*. Oxford: Oxford University Press.

Allan, K. (1977). “Classifiers”. *Language*, 53(2), 285-311.

Desagulier, G. (2014). “Visualizing Distances in a Set of Near Synonyms: Rather, Quite, Fairly, and Pretty”. Robinson, Glynn 2014, 145-78. `https://doi.org/10.1075/hcp.43.06des`.

Divjak, D. (2010). *Structuring the Lexicon. A Clustered Model for Near-Synonymy*. Berlin: De Gruyter Mouton.

Divjak, D.; Fieller, N. (2014). “Cluster Analysis. Finding Structure in Linguistic Data”. Robinson, Glynn 2014, 405-41.

Divjak, D.; Gries, S.T. (2006). “Ways of Trying in Russian. Clustering Behavioral Profiles”. *Corpus Linguistics and Linguistic Theory*, 2(1), 23-60. `https://doi.org/10.1515/cllt.2006.002`.

Dosedlová, A.; Lu, W. (2019). “The Near-Synonymy of Classifiers and Construal Operation. A Corpus-Based Study of 棵 *kē* and 株 *zhū* in Chinese”. *Review of Cognitive Linguistics*, 17(1), 113-30. `https://doi.org/10.1075/rcl.00028.dos`.

Firth, J.R. (1957). “A Synopsis of Linguistic Theory 1930-1955”. *Studies in Linguistic Analysis*. Oxford: Blackwell, 1-32.

Goldberg, A.E. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.

Glynn, D. (2014). “Polysemy and Synonymy. Cognitive Theory and Corpus Method”. Robinson, Glynn 2014, 7-38.

Janda, L.A.; Kudrnáčová, N.; Lu, W. (2019). “Deep Dives into Big Data. Best Practices for Synthesis of Quantitative and Qualitative Analysis in Cognitive Linguistics”. *Review of Cognitive Linguistics*, 17(1), 1-6. `https://doi.org/10.1075/rcl.00023.jan`.

Jiang, S. (2017). *The Semantics of Chinese Classifiers and Linguistic Relativity*. London: Routledge.

Langacker, R.W. (2008). *Cognitive Grammar. A Basic Introduction*. Oxford: Oxford University Press.

Robinson, J.A.; Glynn, D. (eds) (2014). *Corpus Methods for Semantics. Quantitative Studies in Polysemy and Synonymy*. Amsterdam: John Benjamins.

Saalbach, H.; Imai, M. (2012). “The Relation between Linguistic Categories and Cognition. The Case of Numeral Classifiers”. *Language and Cognition Processes*, 27(3), 381-428. `https://doi.org/10.1080/01690965.2010.546585`.

Schmid, H.-J. (2010). “Does Frequency in Text Instantiate Entrenchment in the Cognitive System?”. Fischer, K.; Glynn, D. (eds), *Quantitative Methods in Cognitive Semantics. Corpus-Driven Approaches*. Berlin: De Gruyter Mouton, 101-32. `https://doi.org/10.1515/9783110226423.101`.

Schmid, H.-J.; Küchenhoff, H. (2013). “Collostructional Analysis and Other Ways of Measuring Lexico-Grammatical Attraction. Theoretical Premises, Practical Problems and Cognitive Underpinnings”. *Cognitive Linguistics*, 24(3), 531-78. `https://doi.org/10.1515/cog-2013-0018`.

Sinclair, J. (2005). “Corpus and Text. Basic Principles”. Wynne, M. (ed.), *Developing Linguistic Corpora. A Guide to Good Practice*. Oxford: Oxbow Books, 1-16.

Stefanowitsch, A.; Gries, S.T. (2003). "Collostructions. Investigating the Interaction of Words and Constructions". *International Journal of Corpus Linguistics*, 8(2), 209-43. `https://doi.org/10.1075/ijcl.8.2.03ste`.

Stefanowitsch, A.; Gries, S.T. (2005). "Covarying Collexemes". *Corpus Linguistics and Linguistic Theory*, 1(1), 1-43. `https://doi.org/10.1515/cllt.2005.1.1.1`.

Tai, J.H-Y. (1994). "Chinese Classifier Systems and Human Categorization". Chen, M.Y.; Tzeng, O.J. (eds), *Interdisciplinary Studies on Language and Language Change*. Taipei: Pyramid, 479-94.

Tang, X. (2016). "Lexeme-based Collexeme Analysis with DepCluster". *Corpus Lingustics and Linguistic Theory*, 13(1), 165-202. `https://doi.org/10.1515/cllt-2015-0007`.