On the general classifiers *ge* and *zàg* in Hakka

A corpus-based collostructional analysis
[論客語泛用分類詞「個」與「隻」： 語料庫為本的搭配結構分析]

Han-Chun Huang [黃漢君]
National Tsing Hua University [國立清華大學]

This paper investigates the distribution and properties of the Hakka general classifiers *ge* and *zàg*. We focus on the [determiner/numeral + classifier + noun] construction where we observe the relations between the general classifiers and their following nouns, chosen based on their frequency in this construction. We adopt a corpus-based collostructional analysis which calculates the collocational strength values of *ge* and *zàg* with following nouns. A Hakka corpus was compiled for the study. The three-way distinction in the collostructional analysis (attractive, neutral, and repulsive) is directly mapped to acceptability of various degrees. The results show that *ge* is highly correlated with human-denoting nouns, whereas *zàg* is highly correlated with animal-denoting nouns. Nouns denoting abstract entities or concrete objects without physical properties like size or shape usually lack specific classifiers, and both *ge* and *zàg* can collocate with them, albeit with varying degrees of preference. We argue that both *ge* and *zàg* are general classifiers because both are more frequently used than specific classifiers and both exhibit disjointed semantic distribution and allow abstract nouns. While they show preferences for different nouns, requirements to qualify as general classifiers are equally met.

Keywords: general classifier, collocational strength, collostructional analysis, Hakka

關鍵詞： 泛用分類詞、搭配強度、搭配結構分析、客語

https://doi.org/10.1075/consl.23003.hua
Concentric 49:2 (2023), pp. 261–294. ISSN 1810-7478 | E-ISSN 2589-5230

1. Introduction

Like other Sinitic languages, Hakka is a classifier language in which intervening classifiers are obligatory between determiners/numerals and nouns. Unlike Mandarin Chinese which has only one general classifier *ge*, Hakka has two: *ge* and *zàg*. Hakka *ge* and *zàg* and Mandarin *ge* and *zhī* are etymologically related.[1]

Questions arise when there are two general classifiers in a language: What nouns collocate with the two general classifiers *ge* and *zàg*? What semantic properties exist for those nouns? Are *ge* and *zàg* qualified to be general classifiers? These three questions are interrelated, and we believe that a quantitative approach based on corpus data would be capable of answering them.

To answer these questions, one may try to count the number of nouns that are compatible with each classifier. However, this leads to controversial issues regarding the representativeness of the nouns and definitions of compatibility. Since it is widely acknowledged that the acceptability of linguistic expressions is gradient rather than absolute, there is no point in relying on subjective judgments to solve this issue.

Therefore, we resort to a quantitative, particularly corpus-based approach in which the representativeness of nouns is determined by token frequencies, and acceptability is measured in terms of collocational strength values in the collostructional analysis (Stefanowitsch & Gries 2003, 2005, Gries & Stefanowitsch 2004a, 2004b, Gries, Hampe & Schönefeld 2005). The construction in question is [determiner/numeral + classifier + noun] (henceforth [Det/Num-Cl-N] for short).

In this paper, we aim to clarify the properties of *ge* and *zàg* via statistics and semantic descriptions of nouns that are attracted to them and by comparing them with specific classifiers. The results should benefit linguistic studies *per se* and language teaching.

This paper is organized as follows: Section 2 provides brief sketches of previous work on classifiers of Hakka and related languages. Section 3 presents the theoretical background appropriate to our study. Section 4 describes how the corpus in our analysis was compiled. Section 5 recounts details of the procedures of retrieving relevant data and calculating collocational strength values. Section 6 presents the results, both in tabulated and graphic form. Section 7 presents the discussion. Section 8 concludes this paper.

1. Romanization of Hakka in this paper reflects the Sixian variety and is based on the Taiwanese Hakka Romanization System published by the Ministry of Education, Taiwan. Tone marks, however, are shown as diacritics rather than appearing after syllables.

2. Literature review

Language can be regarded as a means of human cognition. Different languages employ different grammatical mechanisms to implement nominal categorization. Aikhenvald (2003: 1–4) argues that the classification of nouns can be based on semantic features such as animacy, gender, and humanness. Some languages have a bipartite distinction (e.g., Portuguese), some have as many as 10 categories (e.g., Bantu), and others have several dozen (e.g., some South American languages). Nouns in many European languages typically have two or three grammatical genders, which are not always consistent with the biological genders (if any) of the nouns. Many Asian languages (e.g., Chinese, Vietnamese, Korean, and Japanese) use numerical classifiers to categorize nouns. Between a numeral and a noun, a classifier can be either obligatory e.g., *sān *(běn) shū* 'three books' in Mandarin Chinese, or optional, e.g., *dua (buah) buku* 'two books' in Malay.

Chao (1968: 595–631) views classifiers (termed *individual measures*) as a subtype of measures, which also includes group measures like *qún* 'group', partitive measures like *piàn* 'piece', container measures like *bēi* 'cup', and standard measures like *bàng* 'pound'. Functionally speaking, measure words are used to quantify nouns, and sometimes sort them (Tai & Wang 1990, Tai 1994). Therefore, it makes sense to separate classifiers from ordinary measure words.

Her & Hsieh (2010) claim that classifiers carry *essential features* as in *yī wěi yú* 'one fish' whereas ordinary measure words carry *accidental features* as in *yī tǒng yú* 'one bucket of fish'. They also notice that the structural differences between [Num-Adj-M-N] and [Num-M-Adj-N] lead to a semantic distinction for ordinary measure words as in *yī dà xiāng píngguǒ* 'one big box of apples' vs. *yī xiāng dà píngguǒ* 'one box of big apples' but not classifiers as in *yī dà kē píngguǒ* 'one big apple' vs. *yī kē dà píngguǒ* 'one big apple'. Moreover, classifiers imply semantic redundancy and express the quantity *one* mathematically. They therefore can be omitted in circumstances such as in *wǔ bǐng èr yú wèi bǎo wǔ qiān rén* 'five loaves and two fish can feed 5,000 people'. Contrastively, measure words lack semantic redundancy and usually express quantities other than one, and thus cannot be omitted.

They also suggest structural similarities between classifiers and ordinary measure words. First, classifiers and measure words are mutually exclusive and do not appear together. Second, both allow NP-ellipses. Third, both allow omission of the numeral *one*. Forth, both can be followed by *bàn* 'half' and *duō* 'and some'.

It is well observed that a prototype effect exists between classifiers and their following nouns (Tai 2006, Tai & Wu 2006). The classifiers *bué* in Southern Min and *mí* in Hakka are etymological cognates and select fish and snakes. It is also recognized that classifiers are in competition and overlap in distribution, e.g., *bué* and *tsiah* in Southern Min and *mí, tiǎu*, and *zàg* in Hakka. They exhibit variations among native speakers as well as language learners.

The most frequently used classifier in Mandarin is *ge*, called the *general classifier* while all the other individual classifiers are called *specific* (or *special*) *classifiers* (Li & Thompson 1981: 112, Myers 2000: 192–193).

Zubin & Shimojo (1993) suggest that general classifiers can be characterized by three distinct functions, i.e., the complement function, the default function, and the unspecified referent function. Mandarin *ge* is subsumed under the default function (p. 491), though it also exhibits the other two functions. General classifiers with a complement function are mutually exclusive in terms of distribution. Additionally, general classifiers with a default function can replace specific classifiers. Finally, general classifiers with an unspecified referent function are used when information regarding the referent is unavailable.

Myers (2000) argues that the general classifier *ge* is selected by a default rule rather than by analogy (as in specific classifiers) and has no lexical semantics in its own right.[2] It is the 'last resort' classifier should strategies of analogy fail. Zhang (2013: 46–47) suggests that *ge* can alternate with other individual classifiers, though this alternation is not always possible.

Chiu (2007: 199) observes that universal (i.e., general) classifiers exist in Sinitic languages, e.g., *ge* in Mandarin, *ê* in Southern Min, and *zàg* in Hakka. Universal classifiers can replace all other classifiers, at least for the most part. She argues that in Hakka *zàg* collocates with nouns denoting birds, beasts, and human beings, as well as fruit, chairs, watches, boats, money, and words. It can be implied that a prototypical scale exists for *zàg*: birds > beasts > human beings > inanimate things.

Huang (2021) presents a collostructional analysis on four human-denoting classifiers in Hakka. The covarying collexeme analysis is used to measure the collocational strength of classifiers and nouns in the structure [Det/Num-Cl-N]. The results show that *ge* and *zàg* are general classifiers for human beings, though *zàg* sometimes carries a derogatory overtone; *vi* is relatively low in productivity and is usually used to show respect; *sǎ* has the lowest productivity and only combines with the noun *ngǐn* 'human being'.

2. In contrast, Frankowsky & Ke (2016) suggest examining acceptance of *ge* based on an *anthropocentric continuum* (an animacy scale on which all living beings can be placed). They present a six-level scale for animals in terms of humanness: monkeys / predators > mammals > birds / fish > reptiles / snakes / amphibians > insects > mollusks. They found that the acceptance rate of *ge* collocating with different animals exhibits a U-shaped distribution, showing high acceptance rates for animals at both ends and the lowest for birds/fish. They attribute this distribution to two factors: (a) *ge* is for animals distant from humans; (b) *ge* is also the sortal classifier for humans.

Based on previous discussions, we argue that the following properties can distinguish general classifiers and specific classifiers: Structurally speaking, general classifiers are the most frequently used individual classifiers; semantically speaking, general classifiers have disjointed meanings among member nouns and have the ability to categorize abstract nouns. These properties will be referred to in our discussion of *ge* and *zàg* as general classifiers in Hakka.

3. Theoretical framework

We briefly describe the constructional approach and the collostructional analysis in this section. Although both are self-explanatory in their own right, understanding the basic underpinnings of the former benefits understanding the mechanisms involved in the latter.

3.1 The constructional approach

Traditionally, grammar and lexicon have been regarded as distinctive components of language. Grammar, expressed by a set of phrase structure rules, combines with words of the syntactic categories designated by those rules to generate grammatical sentences. The meaning of a grammatical sentence is also compositionally derived from the meaning of the component words in the sentence.

This approach to language was successful, though issues remained of idiomaticity, collocation, and semantic compositionality. Back in the eighties and nineties of the last century, linguists began to observe and study idiosyncrasies of lexical as well as phrasal expressions, such as *let alone* (Fillmore, Kay & O'Connor 1988) which is fully substantive (i.e., lexical), the ditransitive construction (Goldberg 1995) which is fully schematic, and the 'time-*away*' construction (Jackendoff 1997) and the [*What's* X *doing* Y?] construction (Kay & Fillmore 1999) which are partially substantive and partially schematic. They all noticed that structural and semantic inconsistencies in these expressions could not be explained using the traditional approach, and started to believe that a constructional approach might be better.

The basic tenets of a constructional framework, represented by Construction Grammar among others, as in Goldberg (1995, 2003), are that constructions are the building blocks of grammar and that the traditional lexicon-grammar dichotomy is best replaced with constructions of different scales and substances.

Scale-wise, constructions can be a morpheme, a word, a phrase, or a sentence. Substance-wise, constructions can be substantive (equivalent to lexicon in the traditional approach), schematic (equivalent to grammar in the traditional approach), or partially substantive/schematic (no equivalents in the traditional approach). Goldberg (1995) gives the following definition of a construction.

(1) C is a CONSTRUCTION iff$_{def}$ C is a form-meaning pair $<F_i, S_i>$ such that some aspect of F_i or some aspect of S_i is not strictly predictable from C's component parts or from other previously established constructions. (Goldberg 1995: 4)

Goldberg (1995) argues that in argument structure constructions (ditransitive, caused-motion, and resultative), idiosyncrasies can be attributed to constructions *per se* instead of verbs. For example, the verb *bake* would require different argument structures in (2a) and (2b) if a constructional approach were not adopted. One would need to stipulate that *bake* in (2a) is a two-argument verb and that in (2b) a three-argument verb. There would be a proliferation of verbal senses here.

(2) a. Sally baked a cake.
    b. Sally baked her sister a cake. (Goldberg 1995: 141)

Goldberg (1995) suggests that the syntactic pattern [NP1 V NP2 NP3] in (2b), a ditransitive construction, has its own constructional argument roles that must be 'fused' with the verbal participant roles. The constructional argument roles of the ditransitive construction is <agt, rec, pat>.[3] In this way, we may retain a simple two-argument analysis of the verb *bake* for both (2a) and (2b) whose participant roles must be linked to the constructional argument roles. The 'fusion' (or linking) of participant roles and argument roles must observe the Semantic Coherence Principle and the Correspondence Principle as described in Goldberg (1995: 50).

It therefore seems that (schematic) constructions deserve more attention than they have received thus far. This idea also influenced the way collocations were previously treated. In the next subsection, we will show how word-to-word relations in collocations can be extended to word-to-construction relations and as far as word-to-word relations in a certain construction.

3.2 The collostructional analysis

Collocations are common word combinations in which constituent words co-occur more often than may be expected. What counts as 'more often', however, may vary according to subjective judgment. One typical example of a collocation is *strong tea* versus the unlikely *powerful tea*, which can be compared to *powerful computer* versus the unlikely *strong computer*.

3. Here 'agt', 'rec', and 'pat' are short forms of Agent, Recipient, and Patient, respectively.

The concept of collocation lies somewhere between the lexicon and the grammar. Collocations have to be not only syntactically correct, but also lexically consistent. It may be difficult to draw a clear line between (good) collocations and (bad) non-collocations, but many plausible quantitative measures exist that can tell them apart.

The simplest way of measurement is to calculate the raw frequency of a given combination of words. The validity of this method, however, is heavily influenced by the raw frequencies of the constituent words. For example, *of the* may rank top in bigram measurement though it is far from being a good example of collocation. An improvement is to measure the *mutual information*, i.e., the raw frequency of the word combination divided by the multiplication of raw frequencies of constituent words.

A method called *hypothesis testing* can be used to measure collocation, as described in Manning & Schütze (1999: 162–163). In statistics, we can calculate the probability of two events being independent of each other. We formulate a *null hypothesis* H_0 stating that there is no association between the two events beyond mere chance. Then we calculate the probability p of H_0 being true. We reject H_0 if the probability p is too low (typically beneath a significant level of, say, .05), and accept it if otherwise.

A good implementation of hypothesis testing in measuring collocation is the *collostructional analysis* (Stefanowitsch & Gries 2003, 2005, Gries & Stefanowitsch 2004a, 2004b, Gries, Hampe & Schönefeld 2005) which employs the Fisher Exact Test (a small-sample version of the Chi-Squared Test) to calculate numbers in contingency tables. There are three versions of collostructional analysis: the *collexeme analysis* investigates the relations between a lexical item W and a construction C; the *distinctive collexeme analysis* investigates the relations of a lexical item W with respect to two constructions C1 and C2; the *covarying collexeme analysis* investigates the relations of two lexical items W1 and W2 within a construction C. Since only the covarying collexeme analysis is used in this study, we briefly describe how it works using examples from Stefanowitsch & Gries (2005: 11). The English *into*-causative construction can be characterized by [VP+NP+*into*+V-*ing*], exemplified below:

(3) a. … most customers are misled into believing that those guarantees and warranties cover far more than they do
b. … he was forced into making a reluctant announcement
c. Newley had been tricked into revealing his hiding place

The question here is what combination of VP and V-*ing* has the largest collocational strength value in the [VP+NP+*into*+V-*ing*] construction.

Before starting, a contingency table is created as shown in Table 1. The boldface numbers must be calculated before it can be useful.

Table 1. Covarying collexeme analysis

	Word W2	**¬Word W2**	**Row totals**
Word W1	***a***	*b*	***a+b***
¬Word W1	*c*	*d*	*c+d*
Column totals	***a+c***	*b+d*	***(a+b)+(c+d)=N***

First, we calculate ***a***, the frequency of co-occurrence of the words W1 and W2 in construction C, or f((W1,W2)|C) for short. Second, we calculate ***a+b***, the frequency of occurrence of the word W1 in the construction C, or f(W1|C) for short, disregarding whether W2 is present or absent. The difference of the two numbers is *b*, which is the frequency of co-occurrence of the word W1 and any word other than W2 in the slot. Third, we calculate ***a+c***, the frequency of occurrence of the word W2 in the construction C, or f(W2|C) for short, disregarding whether W1 is present or absent. The difference between the two numbers is *c*, which is the frequency of co-occurrence of the word W2 and any word other than W1 in the slot. Last, we also need to know ***N***, the frequency of occurrence of the construction.

We compare the actual value of ***a*** with the expected value of ***a***, which by proportion would be ***(a+c)*(a+b)/N***. If ***a*** is larger than this value, we have an attraction of collexemes. Otherwise, we have a repulsion of collexemes. The results in Table 2 show that [*fool* NP *into thinking*] is a good collocational unit since it has the highest collocational strength value in terms of attraction, whereas [*force* NP *into thinking*] is not since it has the highest collocational strength value in terms of repulsion. We will return to details of calculation when we deal with the Hakka data.

Table 2. Top five in the ranking, from Stefanowitsch & Gries (2005: 13)

Attracted covarying-collexeme pairs in the *into*-causative		**Repelled covarying-collexeme pairs in the *into*-causative**	
fool into thinking	30.06	*force into thinking*	2.554
mislead into thinking	12.755	*coerce into thinking*	1.421
mislead into believing	8.355	*trick into making*	0.945
deceive into thinking	5.651	*push into thinking*	0.794
trick into parting	5.248	*trick into accepting*	0.717

4. The compilation of the Hakka corpus

The Taiwan Hakka Corpus (Hakka Affairs Council 2022) is currently the most updated, balanced corpus of Hakka. From the description on its website, it contains 6 million characters of written data and 0.4 million characters of oral data, covering the six officially recognized varieties of Taiwanese Hakka.[4] It provides functions like online search of keywords and collocations, as well as annotation (including segmentation of characters and part-of-speech labelling) of user-uploaded data. It also authorizes academic use of part of its data (about 1 million characters) by written consent. Sometimes, however, data appear to be repetitive, given that six varieties of the same content are included in the corpus.

To compile a corpus for analysis, we selected the Sixian variety of recurring parts and other nonrecurring parts from the authorized data. We also incorporated data collected previously by the researcher.[5] We then uploaded the collected data to the Taiwan Hakka Corpus for annotation. Currently (at the time of writing), this online service processes at most 5,000 characters in a batch, so we had to limit the size of the uploaded data each time. The annotated data were downloaded and saved.

The results were satisfactory and contained only a few errors. We adapted the annotated data to suit our needs, as described below.

First, all numerals were originally labeled as determiners (DETs). We fixed this issue by finding all numeral tokens and replacing their parts of speech with numerals (NUMs).

Second, the Taiwan Hakka Corpus has no label for classifiers, but only measure words (Ms). Most classifiers were correctly labeled as Ms, though some were labeled as nouns (Ns), e.g., *vi* and *mí*. We changed their parts of speech to Ms if they appeared after DET or NUM and before N. We did not, however, change other non-classifier measure words if they were labeled as Ns, e.g., *zùng, iong, bí*, and *gon*, since they are irrelevant to our study.

The compiled corpus used by this study contains 908,846 characters (before annotation), equivalent to 666,757 word tokens (after annotation), which belong to 29,489 word types.

4. The six varieties are Sixian, Hailu, Dabu, Raoping, Zhao'an, and Southern Sixian.

5. Authorized data from Taiwan Hakka Corpus include *Hakka Certificate Vocabulary Database* (Sixian variety), *Collected Works of the Tung Flower Literary Award* (in the years of 2015 and 2016), *On Hakka Settlements Past and Present and Cyber Settlements*. None of them overlap with the data collected previously by the researcher, including collections of Hakka folk tales, articles for reciting in National Language Contests, and other publicly released data.

To facilitate human labor, we wrote various Python programs to extract data and provide statistics. The results were manually checked and filtered, as sometimes anomalies occur due to incorrect annotation and/or non-standard, unconventional characters. We then modified the programs accordingly to minimize repetitive manual checking.

5. The procedures

We focus on two patterns only: [Det-Cl-N] and [Num-Cl-N], merged as [Det/Num-Cl-N] henceforth. Classifiers and nouns not in the two patterns are not considered.

All matches appearing in the [Det/Num-Cl-N] construction in the corpus were automatically filtered out by the Python programs. We manually removed nouns that did not make sense in the construction.[6] Also, for representativeness we ignored nouns with token frequencies less than 6 in the construction. This left 116 nouns for analysis. In the Appendix we list the 116 nouns used in our calculation, along with f(N) (their own token frequencies), f(*ge*) and f(*zàg*) (their co-occurrence frequencies with *ge* and *zàg*), CS(*ge*) and CS(*zàg*) (their collocational strength values for *ge* and *zàg*).

In the corpus, the pattern [Det/Num-Cl-N] appears 3957 times. This number is ***(a+b)+(c+d)***, or ***N***, in the contingency table. We also calculated the frequency of occurrence of each classifier in question in the construction [Det/Num-Cl-N]. The frequencies of *ge* and *zàg* in the construction are 1102 and 1090, respectively. This number is ***(a+b)*** in the contingency table.

Take the noun *lai-è* 'son' for example. This noun appears 66 times in the construction, disregarding the classifier. Of these 66 times, the classifier *ge* appears 55 times and the classifier *zàg* 11 times. This number is ***a*** in the contingency table.

Then we could calculate all the missing numbers in the contingency table. Table 3 shows the contingency table for *ge* and *lai-è*.

We also calculated the expected value of ***a*** on the assumption that the classifier and the noun are mutually independent. In other words, if we assume a cer-

6. Some removed examples are parts of larger compound nouns, usually modifiers of head nouns, e.g., *hàg-gá* 'Hakka' in *hàg-gá-ngien-gí* 'Hakka proverb' or non-constituent fragments due to incorrect annotation, e.g., *tai-sag* 'big stone' in *tai-sag-těu* 'big stone'. Some removed examples appear in the construction by chance, due to non-standard, unconventional characters used in the data. For example, sii 'to be', incorrectly annotated as a noun, is used to represent the adverb *sii* 'then'. The researcher has tried to minimize unqualified examples by manually checking high-frequency words in the construction, though some errors would remain.

Table 3. Contingency table for *ge* and *lai-è*

	lai-è	**¬*lai-è***	**Row totals**
ge	**55 (exp ≈ 18.4)**	1047	**1102**
¬*ge*	11	2844	2855
Column totals	**66**	3891	**3957**

Note. 'exp' stands for the expected value if *ge* and *lai-è* are mutually independent.

tain classifier and a certain noun are independent of each other, the ratio of their co-occurrence over the occurrence of that noun alone should be the same as the ratio of the occurrence of that classifier alone over the occurrence of all classifiers and nouns in the construction.

In Table 3, the expected value is 66*1102/3957 ≈ 18.4. Since the actual value 55 is larger than the expected value, we therefore know that *ge* and *lai-è* do not co-occur by chance but are attracted to each other.

We then calculated the *p*-value of Table 3 by passing the four numbers in the grids as arguments to the Python function *scipy.stats.fisher_exact* ([[55,1047],[11,2844]]), which is approximately 4.0×10^{-21}. The Fisher Exact Test (two-tailed) shows that *ge* and *lai-è* are highly unlikely to be independent to each other (p-value $= 4.0 \times 10^{-21}$). Since this value is far below the significant level .05, the two words are strongly attracted to each other. To better appreciate the degree of attraction/repulsion, we applied the logarithmic function with base 10 to the *p*-value to get a value of about −20.4 (rounded to the first decimal place). Since the original *p*-value is a measure of probability, which always leads to a negative value after the logarithmic conversion, the negative sign was removed to get a positive collocational strength (henceforth CS) value of about 20.4. The larger this value is, the more unlikely it is that the two words (classifier and noun) in the construction are mutually independent, or, in other words, the more likely it is that they are attracted to each other.

Likewise, we repeated the steps for the classifier *zàg*. Table 4 shows the contingency table for *zàg* and *lai-è*.

Table 4. Contingency table for *zàg* and *lai-è*

	lai-è	**¬*lai-è***	**Row totals**
zàg	**11 (exp ≈ 18.2)**	1079	**1090**
¬*zàg*	55	2812	2867
Column totals	**66**	3891	**3957**

Note. 'exp' stands for the expected value if *zàg* and *lai-è* are mutually independent.

In Table 4, the expected value is 66*1090/3957 ≈ 18.2. Since the actual value 11 is smaller than the expected value, *zàg* and *lai-è* are repulsive to each other, as their co-occurrence is not preferred, with a frequency lower than would occur by chance. The same calculation yields a *p*-value of approximately .051. The Fisher Exact Test (two-tailed) shows that *zàg* and *lai-è* could be independent of each other (*p*-value = .051). Since this value is a little above the significant level .05, we believe that *zàg* and *lai-è* are more or less independent of each other. We also applied the logarithmic conversion to the *p*-value to get a value of approximately −1.3 (rounded to the first decimal place). If the minus sign were removed, we would have a positive CS value of 1.3 for the repulsion of *zàg* and *lai-è*, the same as for *ge* and *lai-è* where a positive CS value of 20.4 expresses attraction.

As there is no way to distinguish repulsion and attraction from the *p*-value only (either the original version or the negative logarithmic version), we add a minus sign on the negatively logarithmically converted *p*-value if the actual value is smaller than the expected value. In this way, positive CS values signal attraction and negative ones signal repulsion. In our example, the corresponding CS value was therefore −1.3 (also rounded to the first decimal place). A linguistic interpretation of the data indicates that the noun *lai-è* 'son' favors the classifier *ge* (with the CS value being 20.4) and is neutral to the classifier *zàg* (with the CS value being −1.3).

6. The results

In this section we present results acquired from the procedures covered in the previous section.

6.1 The general distribution

A two-dimensional Cartesian coordinate system was drawn for all 116 nouns, with the values of the x-axis and the y-axis being the collocational strength values of *ge* and *zàg*, respectively. The distribution is shown in Figure 1.

In Figure 1, most data appear in the vicinity of the origin, whereas a few are located distantly. A coordinate with linear scales is not ideal for visualizing this type of uneven distribution as data points too close to each other cannot be distinguished clearly. The solution is to employ a coordinate with logarithmic scales for both the x- and the y-axes. Since ordinary logarithmic scales deal with positive values only, we chose symmetrical logarithmic scales which allow negative values as well (they are still linear near the origin). Figure 2 is based on Figure 1,

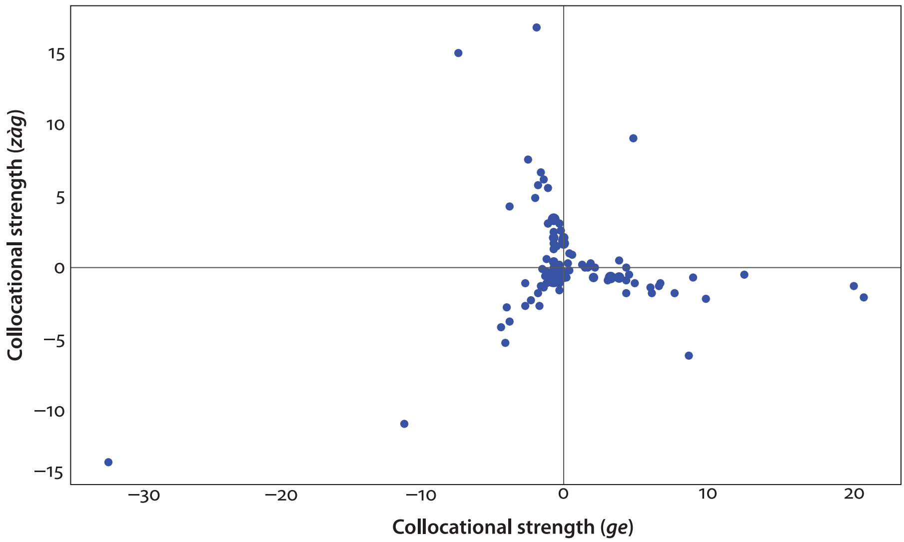

Figure 1. The distribution of CS values of *ge* and *zàg* (in linear scales)

with symmetrical logarithmic scales instead of linear scales, and with data also labeled.[7] The data points are now more evenly distributed.

We divided the distribution into four zones for further discussion: Zone A covers nouns which are attracted to *ge*; Zone B covers nouns attracted to *zàg*; Zone C covers nouns which are neutral with respect to both *ge* and *zàg*; Zone D covers nouns which are repelled by either *ge* or *zàg*, or both.

6.2 Zone A: Nouns attracted to *ge*

We assume the correlation of a classifier and a noun in the [Det/Num-Cl-N] construction is significant if the *p*-value is beneath the level of .05, as is generally practiced. The negative logarithmic conversion of .05 is approximately 1.3. Therefore, Figure 2 uses a value of 1.3 for both the x- and the y-axes as the cut-off value in dividing the zones. Speaker intuition also confirms that combinations with either an x or y value above 1.3 are linguistically acceptable.

Mathematically, nouns attracted to *ge* and *zàg* have CS values of *ge* and *zàg* above 1.3, respectively. Therefore, Zone A and Zone B contain nouns having CS values (x, y) with $x \geq 1.3$ and $x > y$ and those with $y \geq 1.3$ and $y > x$, respectively.[8]

7. Some nouns have the same or approximate values and thus are grouped into a bigger dot, with a label of a noun as a representative followed by the total number of the member nouns.

8. We take the liberty of including 'border nouns' having CS values (x, y) with $x = 1.3$ in Zone A and those with $y = 1.3$ in Zone B, respectively. Moreover, since the noun *ngied* 'month' with CS

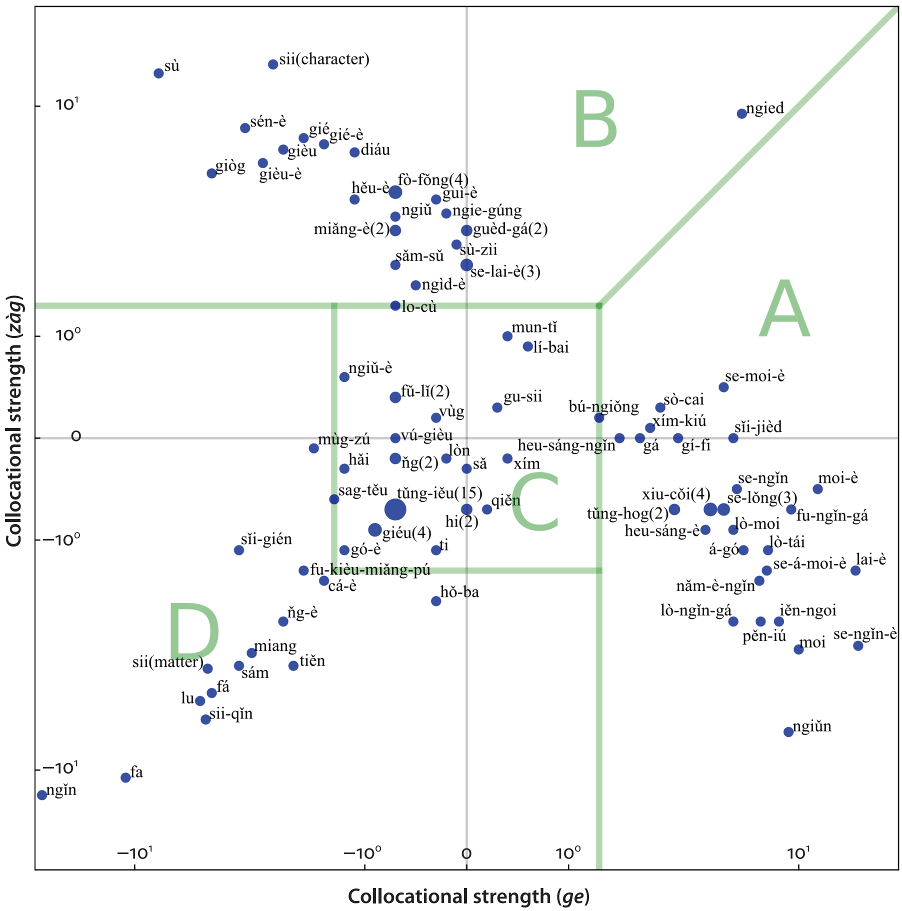

Figure 2. The distribution of CS values of *ge* and *zàg* (in symmetrical logarithmic scales)

Twenty-seven of the 33 nouns in Zone A denote human beings. Therefore, the correlation between the classifier *ge* and human-denoting nouns is high. The CS values for related classifiers, sorted in descending order of the CS values of *ge*, are summarized in Table 5.

Most nouns in the table only collocate with *ge*, with a few exceptions: *ngiǔn* 'money' also accepts *gòg* as its classifier (with an even higher CS value than that of *ge*). Some human-denoting nouns here also accept *vi* as their classifiers (though with lower CS values than that of *ge*).

values (4.9, 9.1) qualifies for inclusion in both zones, further constraints are set for Zone A (x > y) and Zone B (y > x). Therefore, the noun *ngied* 'month' falls in Zone B.

Table 5. CS values of nouns in Zone A

Noun	Gloss	CS (*ge*, *zàg*)	Other CS values
se-ngǐn-è (細人仔)	child	(21.1, −2.1)	(N/A)
lai-è (俫仔)	son	(20.4, −1.3)	(N/A)
moi-è (妹仔)	daughter	(12.7, −0.5)	(N/A)
moi (妹)	daughter	(10.0, −2.2)	(N/A)
fu-ngǐn-gá (婦人家)	woman	(9.1, −0.7)	(N/A)
ngiǔn (銀)	money	(8.8, −6.2)	*gòg* (角) (24.1)
iěn-ngoi (員外)	landlord	(7.8, −1.8)	(N/A)
lò-tái (老弟)	younger brother	(6.8, −1.1)	(N/A)
se-á-moi-è (細阿妹仔)	young lady	(6.7, −1.3)	(N/A)
pěn-iú (朋友)	friend	(6.2, −1.8)	*vi* (位) (1.0)
nǎm-è-ngǐn (男仔人)	adult male	(6.1, −1.4)	(N/A)
á-gó (阿哥)	elder brother	(5.0, −1.1)	(N/A)
se-ngǐn (細人)	child	(4.6, −0.5)	(N/A)
sǐi-jièd (時節)	time	(4.4, 0.0)	(N/A)
lò-moi (老妹)	younger sister	(4.4, −0.9)	(N/A)
lò-ngǐn-gá (老人家)	the elderly	(4.4, −1.8)	*vi* (位) (3.8)
se-moi-è (細妹仔)	young lady	(3.9, 0.5)	(N/A)
se-lǒng (婿郎)	son-in-law	(3.9, −0.7)	(N/A)
sún (孫)	grandchild	(3.9, −0.7)	(N/A)
nǎm-ngǐn (男人)	adult male	(3.9, −0.7)	(N/A)
xiu-cǒi (秀才)	scholar	(3.3, −0.7)	(N/A)
hiúng-ti (兄弟)	brother	(3.3, −0.7)	(N/A)
lò-fo-è (老貨仔)	the elderly (derogatory)	(3.3, −0.7)	(N/A)
sii-gie (世界)	world	(3.3, −0.7)	(N/A)
heu-sáng-è (後生仔)	the youth	(3.1, −0.9)	*vi* (位) (2.3), *sǎ* (儕) (0.2)
gí-fi (機會)	chance	(2.2, 0.0)	(N/A)
tǔng-hog (同學)	classmate	(2.1, −0.7)	*vi* (位) (1.4)
xín-sáng (先生)	teacher	(2.1, −0.7)	*vi* (位) (1.4)
sò-cai (所在)	place	(1.9, 0.3)	(N/A)
xím-kiú (心臼)	daughter-in-law	(1.8, 0.1)	(N/A)
gá (家)	home	(1.7, 0.0)	(N/A)
heu-sáng-ngǐn (後生人)	the youth	(1.5, 0.0)	(N/A)
bú-ngiǒng (餔娘)	wife	(1.3, 0.2)	(N/A)

6.3 Zone B: Nouns attracted to *zàg*

In Zone B, 12 of the 29 nouns denote animals. Others denote body parts, small objects (including celestial bodies, which are conceptually small), time, location, and other abstract entities. The CS values for related classifiers, sorted in descending order of the CS values of *zàg*, are summarized in Table 6.

It is worth noting that other classifiers also collocate with nouns here. Body parts like *sù* 'hand' and *giòg* 'foot' collocate with *gí*. Beasts like *gièu* 'dog' and *gièu-è* 'dog' collocate with *tiău*. Celestial bodies like *sén-è* 'star' collocate with *liab* since they are conceptually small.

Table 6. CS values of nouns in Zone B

Noun	Gloss	CS (*ge*, *zàg*)	Other CS values
sii (字)	Chinese character	(−1.9, 16.9)	*hăng-è* (行仔) (1.9), *zàg-è* (隻仔) (1.3)
sù (手)	hand	(−7.4, 15.1)	*gí* (支) (3.9)
ngied (月)	month	(4.9, 9.1)	(N/A)
sén-è (星仔)	star	(−2.5, 7.6)	*liab* (粒) (1.2)
gié (雞)	chicken	(−1.6, 6.7)	(N/A)
gié-è (雞仔)	chicken	(−1.4, 6.2)	(N/A)
gièu (狗)	dog	(−1.8, 5.8)	*tiău* (條) (0.9)
diáu (鳥)	bird	(−1.1, 5.6)	(N/A)
gièu-è (狗仔)	dog	(−2.0, 4.9)	*tiău* (條) (0.8)
giòg (腳)	foot	(−3.8, 4.3)	*gí* (支) (7.4)
fò-fŏng (伙房)	aggregated homestead	(−0.7, 3.4)	(N/A)
fŭng-báu (紅包)	red packet	(−0.7, 3.4)	(N/A)
gié-mă (雞嫲)	hen	(−0.7, 3.4)	(N/A)
tai-sii (大字)	big word; calligraphy	(−0.7, 3.4)	(N/A)
guì-è (鬼仔)	ghost	(−0.3, 3.1)	(N/A)
hĕu-è (猴仔)	monkey	(−1.1, 3.1)	*tiău* (條) (0.7)
ngie-gúng (蟻公)	ant	(−0.2, 2.6)	(N/A)
ngiŭ (牛)	cattle	(−0.7, 2.5)	*tiău* (條) (1.8)
guèd-gá (國家)	country	(0.0, 2.1)	(N/A)
sĭi-toi (時代)	era	(0.0, 2.1)	(N/A)
miăng-è (名仔)	name	(−0.7, 2.1)	(N/A)
diáu-è (鳥仔)	bird	(−0.7, 2.1)	(N/A)
sù-zìi (手指)	finger	(−0.1, 1.9)	*gí* (支) (1.2)
se-lai-è (細倈仔)	boy	(0.0, 1.7)	(N/A)
vi-sò (位所)	location	(0.0, 1.7)	(N/A)

Table 6. *(continued)*

Noun	Gloss	CS (*ge, zàg*)	Other CS values
gúng-iěn (公園)	park	(0.0, 1.7)	(N/A)
sǎm-sǔ (蟾蜍)	toad	(−0.7, 1.7)	(N/A)
ngìd-è (日仔)	day	(−0.5, 1.5)	*ton* (段) (4.0)
lo-cù (老鼠)	mouse	(−0.7, 1.3)	*tiǎu* (條) (1.1)

6.4 Zone C: Nouns neutral to both *ge* and *zàg*

Mathematically, nouns neutral to both *ge* and *zàg* have CS values of *ge* and *zàg* between −1.3 and 1.3. Therefore, Zone C contains nouns having CS values (x, y) with $-1.3 < x < 1.3$ and $-1.3 < y < 1.3$. Since both classifiers are not in attractive relations with the nouns in this zone, it is expected that other specific classifiers attract them.

This is true for all nouns in this zone except two: *lí-bai* 'week' and *mun-tǐ* 'problem' do not have other specific classifiers. They appear in the upper-right corner and have relatively higher CS values for both *ge* and *zàg* than other nouns. If a noun in this zone has its own specific classifier, it is usually favored. The higher the CS value for the specific classifier is, the lower the CS values for *ge* and *zàg* become.

We observe that collocations of various degrees of acceptability fall in this zone. Although all are categorized as 'neutral' in the collostructional analysis, there is no denying that higher CS values map to higher degrees of acceptability. While expressions like *ìd ge mun-tǐ* 'one problem' and *ìd zàg mun-tǐ* 'one problem' are both acceptable, those like *ìd ge bìd* and *ìd zàg bìd* (with the intended reading 'one pen') are not. We therefore draw a line between positive CS values and non-positive ones. Therefore, except for nouns that do not have specific classifiers (e.g., *lí-bai* 'week' and *mun-tǐ* 'problem'), nouns having CS values of between 0 and 1.3 can be considered marginally acceptable, and those having non-positive CS values can be considered unacceptable. Accordingly, the data in Zone C can all be considered either marginally acceptable or unacceptable with respect to *ge* or *zàg*, or both. This is more consistent with native speaker intuition.

Table 7. CS values of nouns in Zone C

Noun	Gloss	CS (*ge, zàg*)	Other CS values
lí-bai (禮拜)	week	(0.6, 0.9)	(N/A)
mun-tǐ (問題)	problem	(0.4, 1.0)	(N/A)
gu-sii (故事)	story	(0.3, 0.3)	*ton* (段) (1.8), *tiǎu* (條) (1.1)

Table 7. *(continued)*

Noun	Gloss	CS (*ge*, *zàg*)	Other CS values
xím (心)	heart	(0.4, −0.2)	*liab* (粒) (2.0), *tiǎu* (條) (0.3)
vùg (屋)	house	(−0.3, 0.2)	*co* (座) (4.6), *súng* (雙) (1.4)
sǎ (蛇)	snake	(0.0, −0.3)	*mí* (尾) (8.1)
fǔ-lǐ (狐狸)	fox	(−0.7, 0.4)	*tiǎu* (條) (2.8)
biàng-è (餅仔)	cake; pie	(−0.7, 0.4)	*de* (垤) (4.2)
lòn (卵)	egg	(−0.2, −0.2)	*liab* (粒) (6.9)
qiěn (錢)	money	(0.2, −0.7)	*lǐ* (厘) (2.8), *bìd* (筆) (2.1)
ngiǔ-è (牛仔)	cattle	(−1.2, 0.6)	*tiǎu* (條) (2.3), *těu* (頭) (1.4)
hi (戲)	drama	(0.0, −0.7)	*cùd* (齣) (12.0)
gùg (穀)	grain	(0.0, −0.7)	*gín* (斤) (9.0), *liab* (粒) (0.9)
vú-gièu (烏狗)	black dog	(−0.7, 0.0)	*tiǎu* (條) (4.1)
ňg (魚)	fish	(−0.7, −0.2)	*mí* (尾) (11.0)
gòg (角)	horn	(−0.7, −0.2)	*gí* (支) (7.7)
ti (地)	land	(−0.3, −1.1)	*de* (垤) (7.1), *kuai* (塊) (2.9), *xióng* (廂) (1.9)
tǔng-iěu (童謠)	nursery rhyme	(−0.7, −0.7)	*sù* (首) (10.5), *tiǎu* (條) (0.9)
xien (線)	line; thread	(−0.7, −0.7)	*tiǎu* (條) (7.5)
bìd (筆)	pen	(−0.7, −0.7)	*gí* (支) (9.9)
bu-è (布仔)	cloth	(−0.7, −0.7)	*de* (垤) (7.8), *pìd* (匹) (2.8), *tiǎu* (條) (0.3)
san-è (扇仔)	fan	(−0.7, −0.7)	*gí* (支) (8.5)
su-è (樹仔)	tree	(−0.7, −0.7)	*těu* (頭) (8.2), *cǔng* (叢) (3.5)
xin-è (信仔)	letter	(−0.7, −0.7)	*fúng* (封) (17.3)
bu (布)	cloth	(−0.7, −0.7)	*de* (垤) (8.4), *kuai* (塊) (1.4)
síi (詩)	poem	(−0.7, −0.7)	*sù* (首) (8.2), *gi* (句) (0.9), *tiǎu* (條) (0.4)
pǐ (皮)	skin; leather	(−0.7, −0.7)	*cěn* (層) (9.8), *de* (垤) (1.1)
sò-sìi (鎖匙)	key	(−0.7, −0.7)	*gí* (支) (8.5)
lug-è (鹿仔)	deer	(−0.7, −0.7)	*tiǎu* (條) (6.4)
biàg (壁)	wall	(−0.7, −0.7)	*san* (扇) (16.4)
cò (草)	grass	(−0.7, −0.7)	*gí* (枝) (7.3), *těu* (頭) (1.5)
ién (菸)	cigarette	(−0.7, −0.7)	*gí* (支) (6.3), *hèu-è* (口仔) (2.5)
hǎi (鞋)	shoe	(−1.2, −0.3)	*súng* (雙) (17.7)
giéu (溝)	ditch	(−0.9, −0.9)	*tiǎu* (條) (8.5)
sǎ-gó (蛇哥)	snake	(−0.9, −0.9)	*mí* (尾) (10.4), *tiǎu* (條) (0.8)

Table 7. *(continued)*

Noun	Gloss	CS (*ge, zàg*)	Other CS values
miĕn-iŏng (綿羊)	sheep	(−0.9, −0.9)	*tiău* (條) (8.5)
tŭng-fá-su (桐花樹)	tung tree	(−0.9, −0.9)	*cŭng* (叢) (19.5)
gó-è (歌仔)	song	(−1.2, −1.1)	*tiău* (條) (7.6), *sù* (首) (1.4)

6.5 Zone D: Nouns repelled by either *ge* or *zàg*, or both

Mathematically, nouns repelled by either *ge* or *zàg* have at least one of the CS values for *ge* and *zàg* beneath –1.3. Therefore, Zone D contains nouns having CS values (x, y) with $x \leq -1.3$ or $y \leq -1.3$. None of the data here collocate with either *ge* or *zàg*. Note that the noun in the distant corner, *ngĭn* 'human being' is peculiar in that it attracts the highly dedicated classifier *să* while repelling both *ge* and *zàg*. Other nouns also have their own specific classifiers.

Table 8. CS values of nouns in Zone D

Noun	Gloss	CS (*ge, zàg*)	Other CS values
mùg-zú (目珠)	eye	(−1.5, −0.1)	*liab* (粒) (15.7), *luí* (蕊) (0.7)
hŏ-ba (河壩)	river dam	(−0.3, −1.6)	*tiău* (條) (8.3)
sag-tĕu (石頭)	stone	(−1.3, −0.6)	*liab* (粒) (18.9)
cá-è (車仔)	car	(−1.4, −1.4)	*tŏi* (臺) (19.6), *bióng* (枋) (1.7), *liŏng* (輛) (1.5)
fu-kièu-miăng-pú (戶口名簿)	household register	(−1.6, −1.3)	*bùn* (本) (29.4)
ňg-è (魚仔)	fish	(−1.8, −1.8)	*mí* (尾) (24.2), *tiău* (條) (0.4)
sĭi-gién (時間)	time	(−2.7, −1.1)	*ton* (段) (37.8)
tiĕn (田)	farmland	(−1.7, −2.7)	*kiú* (坵) (36.2), *fun* (份) (2.7), *de* (埞) (0.6)
miang (命)	life; fortune	(−2.3, −2.3)	*tiău* (條) (18.3)
sám (衫)	clothes	(−2.7, −2.7)	*liáng* (領) (40.3), *tĕu* (頭) (1.0)
sii (事)	matter	(−4.0, −2.8)	*kien* (件) (48.1)
fá (花)	flower	(−3.8, −3.8)	*luí* (蕊) (49.6), *gí* (枝) (0.7)
lu (路)	road	(−4.4, −4.2)	*tiău* (條) (29.8), *ton* (段) (1.4)
sii-qĭn (事情)	matter	(−4.1, −5.3)	*kien* (件) (63.8)
fa (話)	spoken word	(−11.2, −11.0)	*gi* (句) (149.7), *xid* (席) (1.4)
ngĭn (人)	human being	(−32.0, −13.7)	*să* (儕) (265.9)

7. Discussion

From the data distribution in the previous section, we see that the two general classifiers *ge* and *zàg* in Hakka exhibit different attraction patterns on the following nouns. While *ge* is highly correlated with human-denoting nouns, *zàg* is correlated to a wide range of nouns denoting animals, body parts, small things, time, location, and abstract entities. We also notice that specific classifiers compete with the general classifiers. In this section, we discuss the general classifiers *ge* and *zàg*, as well as some specific classifiers found in the previous section, and then compare them with each other.

7.1 General classifiers

In this subsection, we present a brief discussion on the properties of the general classifiers found in our data.

7.1.1 *Ge*

From Section 6.2, we see that human-denoting nouns are prominent in Zone A: 27 out of the 33 nouns are human-denoting. The top five nouns with high CS values for *ge* are all human-denoting: *se-ngǐn-è* 'child' (21.1), *lai-è* 'son' (20.4), *moi-è* 'daughter' (12.7), *moi* 'daughter' (10.0), and *fu-ngǐn-gá* 'woman' (9.1).

7.1.2 *Zàg*

From Section 6.3, we see that although animal-denoting nouns are prominent in Zone B, they amount to only 12 out of the 29 nouns. What is more interesting is that the four nouns with the highest CS values for *zàg* are not animal-denoting: *sii* 'Chinese character' (16.9), *sù* 'hand' (15.1), *ngied* 'month' (9.1), and *sén-è* 'star' (7.6). It is difficult, if not impossible, to find a common property among these four nouns.

7.1.3 Distribution of human-denoting and animal-denoting nouns

To better understand the correlations between *ge/zàg* and human/animal-denoting nouns, we redrew Figure 2 using green triangles to express human-denoting nouns and red squares to express animal-denoting nouns, as shown in Figure 3. For simplicity, overlapping data are categorized based on the first member nouns only.

Although *ge* and *zàg* favor human-denoting and animal-denoting nouns respectively, they also select a wide range of nouns unrelated to either humans or animals. Non-human-denoting nouns with CS values of *ge* above or equal to 1.3 include *ngiǔn* 'money', *sǐi-jièd* 'time', *sii-gie* 'world', *gí-fi* 'chance', *sò-cai* 'place', and

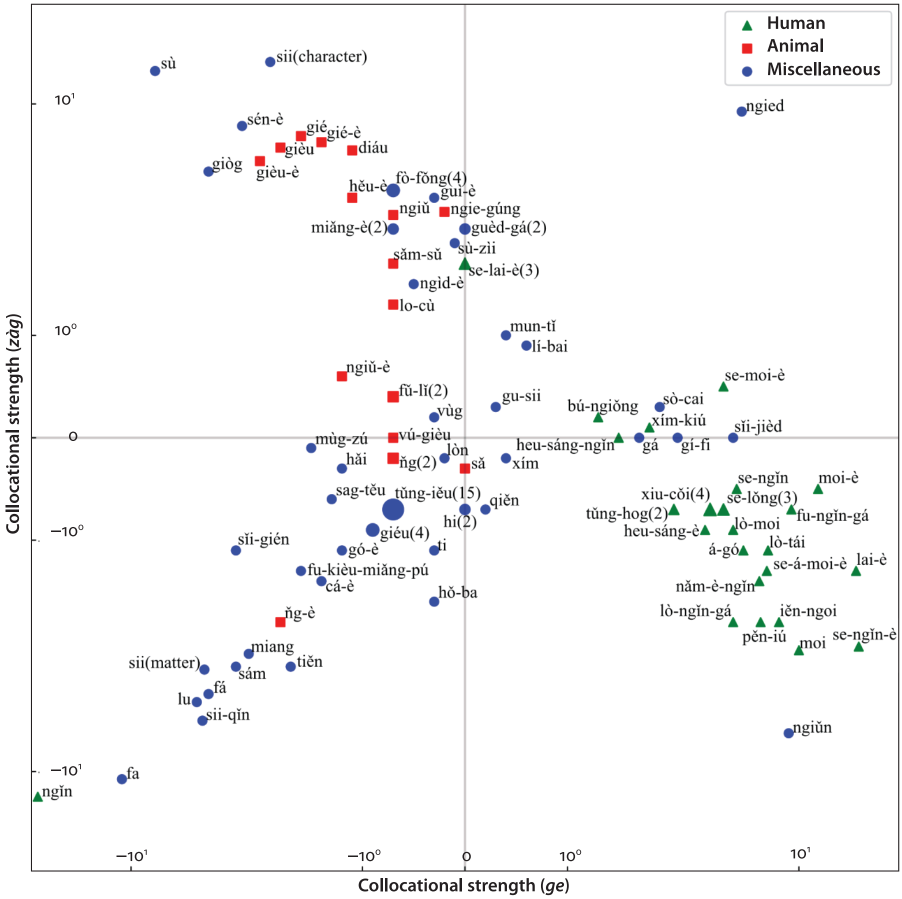

Figure 3. The distribution of human-denoting and animal-denoting nouns

gá 'home'. They do not possess an obvious common semantic feature except for maybe abstractness.

Non-animal-denoting nouns with CS values of *zàg* above or equal to 1.3 include *sii* 'Chinese character', *sù* 'hand', *ngied* 'month' (also selected by *ge*), *sén-è* 'star', *giòg* 'foot', *guì-è* 'ghost', *fò-fŏng* 'aggregated homestead', *fŭng-báu* 'red packet', *tai-sii* 'big word; calligraphy', *guèd-gá* 'country', *sĭi-toi* 'era', *miăng-è* 'name', *sù-zìi* 'finger', *se-lai-è* 'boy', *vi-sò* 'location', *gúng-iĕn* 'park', and *ngìd-è* 'day'. Since the Chinese character 隻 (*zhī* in Mandarin and *zàg* in Hakka) was originally used to express birds, semantic features like 'smallness', 'derogatoriness', and 'body parts' can be postulated to be extensions from the original meaning. This covers data like *sù* 'hand', *sén-è* 'star' (being visually small as seen through human eyes), *giòg* 'foot', *guì-è* 'ghost', *fŭng-báu* 'red packet', *tai-sii* 'big word; calligraphy', *sù-zìi* 'finger', and *se-lai-è* 'boy'. However, other abstract nouns cannot be explained at all.

Therefore, it seems that although semantic features like 'human' and 'animal' are characteristic of the majority of nouns with high CS values of *ge* and *zàg* respectively, there is no way to find any common features among all the nouns with high CS values of either *ge* or *zàg*.

7.2 Specific classifiers

Classifiers other than *ge* and *zàg* are specific since they pose semantic restrictions on following nouns. The CS values of these specific classifiers are usually high, indicating their mutual attraction to each other. In this subsection, we present a brief discussion on the properties of specific classifiers found in our data.

7.2.1 *Tiău*

The classifier *tiău* collocates with nouns denoting linear objects (either concrete or abstract) or animals. Below is a list of nouns that were found to follow *tiău* in our data, sorted in descending order by their CS values of *tiău*:

Of the 23 nouns that collocate with *tiău*, 12 denote animals that can be conceptualized as linear objects, 5 denote linear objects, 4 denote linguistic and/or musical contents, and 2 are not in any of the categories above. Since linguistic and/or musical contents are temporal, and thus conceptually linear, it is clear from the data that the prototypical meaning of nouns collocating with *tiău* is linearity.

Table 9. Nouns that collocate with *tiău*

Noun	Gloss	CS (*tiău*)	Semantic category
lu (路)	road	29.8	linear object
miang (命)	life	18.3	miscellaneous
miĕn-iŏng (綿羊)	sheep	8.5	animal
giéu (溝)	ditch	8.5	linear object
hŏ-ba (河壩)	river dam	8.3	linear object
gó-è (歌仔)	song	7.6	content (linguistic, musical)
xien (線)	line	7.5	linear object
lug-è (鹿仔)	deer	6.4	animal
vú-gièu (烏狗)	black dog	4.1	animal
fŭ-lĭ (狐狸)	fox	2.8	animal
ngiŭ-è (牛仔)	cattle	2.3	animal

Table 9. *(continued)*

Noun	Gloss	CS (*tiǎu*)	Semantic category
ngiǔ (牛)	cattle	1.8	animal
lo-cù (老鼠)	mouse	1.1	animal
gu-sii (故事)	story	1.1	content (linguistic, musical)
gièu (狗)	dog	0.9	animal
tǔng-iěu (童謠	nursery rhyme	0.9	content (linguistic, musical)
gièu-è (狗仔)	dog	0.8	animal
sǎ-gó (蛇哥)	snake	0.8	animal
hěu-è (猴仔)	monkey	0.7	animal
síi (詩)	poem	0.4	content (linguistic, musical)
ňg-è (魚仔)	fish	0.4	animal
bu-è (布仔)	cloth	0.3	linear object
xím (心)	heart	0.3	miscellaneous

7.2.2 *Gí*

The classifier *gí* collocates with nouns denoting straight objects or body parts. Below is a list of nouns that follow *gí* in our data, sorted in descending order by their CS values of *gí*:

Table 10. Nouns that collocate with *gí*

Noun	Gloss	CS (*gí*)	Semantic category
bìd (筆)	pen	9.9	straight object
san-è (扇仔)	fan	8.5	straight object
sò-sìi (鎖匙)	key	8.5	straight object
gòg (角)	horn	7.7	body part
giòg (腳)	foot	7.4	body part
ién (菸)	cigarette	6.3	straight object
sù (手)	hand	3.9	body part
sù-zìi (手指)	finger	1.2	body part

Of the 8 nouns found to collocate with *gí*, 4 denote straight objects and 4 denote body parts (of human beings or animals). Since body parts, especially limbs, are also conceptualized as straight objects, we suggest that the prototypical meaning of nouns that collocate with *gí* is straightness.

7.2.3 *Liab*

The classifier *liab* collocates with nouns denoting tiny, round, and spherical objects. Below is a list of nouns found to follow *liab* in our data, sorted in descending order by their CS values of *liab*:

The nouns that collocate with *liab* are tiny, round, and spherical objects. Stars, though geometrically huge, are visually small and thus count as such objects.

Table 11. Nouns that collocate with *liab*

Noun	Gloss	CS (*liab*)	Semantic category
sag-těu (石頭)	stone	18.9	tiny, round, and spherical object
mùg-zú (目珠)	eye	15.7	tiny, round, and spherical object
lòn (卵)	egg	6.9	tiny, round, and spherical object
xím (心)	heart	2.0	tiny, round, and spherical object
sén-è (星仔)	star	1.2	tiny, round, and spherical object
gùg (穀)	grain	0.9	tiny, round, and spherical object

7.2.4 *De*

The classifier *de* collocates with nouns denoting planar objects. Below is a list of nouns found to follow *de* in our data, sorted in descending order by their CS values of *de*:

Table 12. Nouns that collocate with *de*

Noun	Gloss	CS (*de*)	Semantic category
bu (布)	cloth	8.4	planar object
bu-è (布仔)	cloth	7.8	planar object
ti (地)	land	7.1	planar object
biàng-è (餅仔)	cake; pie	4.2	planar object
pǐ (皮)	skin; leather	1.1	planar object
tiěn (田)	farmland	0.6	planar object

If we compare *ti* 'land' with *tiěn* 'farmland', we can see a huge difference with respect to their CS values with *de* (7.1 vs. 0.6). A classifier being able to collocate with a certain noun does not mean it is the best choice. We see from Table 8 that the classifier *kiú* has a relatively high CS value with *tiěn* (36.2). This also supports the advantage of using a collostructional analysis to deal with Hakka classifiers. It is not terribly wrong to say something like *ìd de tiěn* to express 'a piece of farmland' in Hakka, but saying *ìd kiú tiěn* is much more natural and generally favored.

7.2.5 *Vi*

The classifier *vi* collocates with nouns denoting respectable human beings. Below is a list of nouns found to follow *vi* in our data, sorted in descending order by their CS values of *vi*:

Table 13. Nouns that collocate with *vi*

Noun	Gloss	CS (*vi*)	Semantic category
lò-ngǐn-gá (老人家)	the elderly	3.8	respectable human
heu-sáng-è (後生仔)	the youth	2.3	respectable human
tǔng-hog (同學)	classmate	1.4	respectable human
xín-sáng (先生)	teacher	1.4	respectable human
pěn-iú (朋友)	friend	1.0	respectable human

7.2.6 *Mí*

The classifier *mí* collocates with nouns denoting fish or snakes. Below is a list of nouns found to follow *mí* in our data, sorted in descending order by their CS values of *mí*:

Table 14. Nouns that collocate with *mí*

Noun	Gloss	CS (*mí*)	Semantic category
ňg-è (魚仔)	fish	24.2	fish
ňg (魚)	fish	11.0	fish
sǎ-gó (蛇哥)	snake	10.4	snake
sǎ (蛇)	snake	8.1	snake

7.2.7 *Ton*

The classifier *ton* collocates with nouns denoting linear objects, both concrete and abstract. Linguistic contents are temporal and thus are conceptually linear objects. Below is a list of nouns found to follow *ton* in our data, sorted in descending order by their CS values of *ton*:

Table 15. Nouns that collocate with *ton*

Noun	Gloss	CS (*ton*)	Semantic category
sĭi-gién (時間)	time	37.8	linear abstract object
ngìd-è (日仔)	day	4.0	linear abstract object
gu-sii (故事)	story	1.8	content (linguistic, musical)
lu (路)	road	1.4	linear object

Despite differences in their semantic categories, the nouns are all linear objects that can be segmented, either physically or conceptually.

7.3 Comparison of general classifiers and specific classifiers

In a strict sense, we can say the term 'general classifier' is an oxymoron and paradoxical. By 'general' we mean universal and non-discriminative, so ideally a general classifier could collocate with any noun. However, one defining feature of classifiers is having a sortal function. If a classifier ceases to select nouns based on properties such as size or shape, can it still be called a classifier?

We believe that no ideal general classifiers exist that indiscriminately select any noun, as is also observed in Zhang (2013). Classifiers, as function words required by grammar, may still retain their semantic preferences but eventually gain access to other semantically related nouns through extension (for specific classifiers) or semantically unrelated nouns as a default rule (for general classifiers).

In this subsection, we present a comparison of specific classifiers and general classifiers. We argue that general classifiers are more frequently used (in terms of frequency) and have disjointed semantics among member nouns and have the ability to categorize abstract nouns.

The property of being 'most frequently used' requires quantitative measurements. We tackle this issue from both a *word type* perspective and a *word token* perspective.

Based on our data of *ge* and *zàg* previously and the seven specific classifiers in Section 7.2, we came up with the table below. Although it is intuitive to map the three kinds of relations (attractive, neutral, and repulsive) to the three levels of acceptability (acceptable, marginally acceptable, and unacceptable), following the discussion in Section 6.4, we modified our criteria after manually inspecting data with CS values between −1.3 and 0. They are for the large part unacceptable even though they are categorized as neutral. Therefore, the three levels of acceptability were determined using the three sets of ranges: acceptable if CS ≥ 1.3, marginally acceptable if 0 < CS < 1.3, unacceptable if CS ≤ 0.

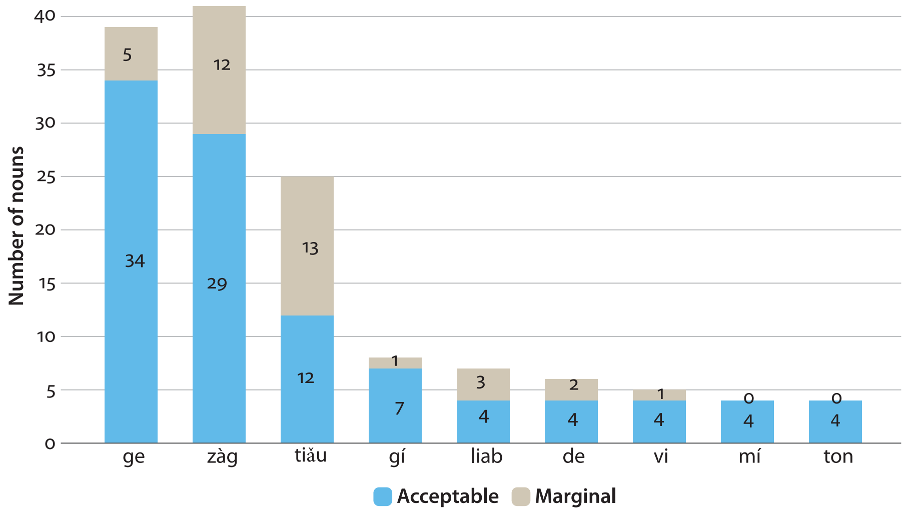

Figure 4. Distribution of acceptability of the classifiers with 116 following nouns

Since as seen in our data a noun can collocate with multiple classifiers, it is natural to note that the accumulated number in the table exceeds 116. It is the *noun type*, not the *noun token*, that participates in the statistics.

We can also measure the relative shares of the top classifiers in the [Det/Num-Cl-N] construction. In Section 5, we see that the total frequency of this construction is 3957, while *ge* and *zàg* have frequencies of 1102 and 1090 respectively. In addition to the two general classifiers, the top ten classifiers of the rest are (with their frequencies in parentheses) *tiǎu* (341), *sǎ* (245), *gí* (154), *liab* (92), *gi* (89), *kien* (81), *de* (61), *luí* (52), *mí* (44), and *ton* (40). The bar graph in Figure 5 shows their relative shares. It is obvious from this graph that the general classifiers *ge* and *zàg* get the lion's share (about 55.40%), while the top ten classifiers of the rest appear in only 30.30% of constructions. Together, the twelve classifiers account for 85.70% of the construction [Det/Num-Cl-N]. Here, it is the *noun token*, not the *noun type*, that participates in the statistics.

It is clear from the previous discussion that the two general classifiers are most frequently used in terms of both noun types and noun tokens.

The property of being semantically disjointed among the member nouns needs inspecting. From the data presented previously, we see that *ge* and *zàg* are also specific in that they favor nouns denoting certain semantic properties, e.g.,

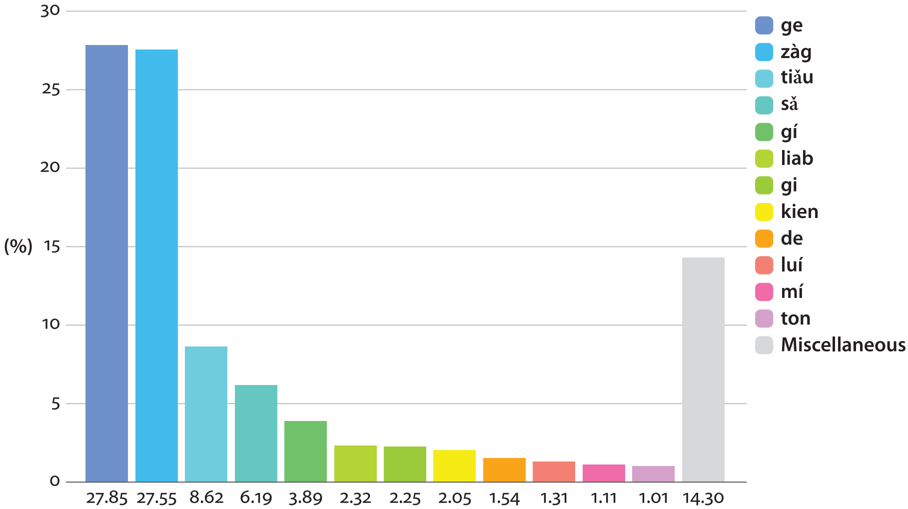

Figure 5. The relative shares of classifiers in the [Det/Num-Cl-N] construction

human-denoting for *ge* and animal-denoting for *zàg*. They are general only in that they combine more freely and in larger quantities with nouns.

If both *ge* and *zàg* have their own preferences in collocating with nouns, can we still say that they are general classifiers? The answer is yes. From the discussion in Section 7.1.3, we see that there is no way to find a common property among all the nouns with high CS values of either *ge* or *zàg*. Being semantically disjointed is a property of general classifiers. Also, both *ge* and *zàg* select abstract nouns. This is yet another property typical of general classifiers.

On the other hand, having central properties is indicative of specific classifiers, as has been justified in our discussion of typical specific classifiers in Section 7.2, where central properties can be usually found among the nouns collocating with them, for example linearity for *tiǎu* and *ton*, roundness for *liab*, straightness for *gí*, and respectfulness for *vi*.

Furthermore, as we can see from Tables 5–7, nouns lacking specific classifiers allow both *ge* and *zàg*, although the CS values may be small. There are nouns preferring *ge* to *zàg*, e.g., *sǐi-jièd* 'time' (4.4, 0.0), *sii-gie* 'world' (3.3, −0.7), *gí-fi* 'chance' (2.2, 0.0), and *sò-cai* 'place' (1.9, 0.3), and there are nouns preferring *zàg* to *ge*, e.g., *guèd-gá* 'country' (0.0, 2.1), *sǐi-toi* 'era' (0.0, 2.1), *miǎng-è* 'name' (−0.7, 2.1), *vi-sò* 'location' (0.0, 1.7), *mun-tǐ* 'problem' (0.4, 1.0), and *lí-bai* 'week' (0.6, 0.9). Here we see that the less-preferred classifier for each noun is still neutral to the noun (mostly having CS values larger than zero), which means that their use is (marginally) acceptable. This implies that both *ge* and *zàg* are general classifiers, or 'default' classifiers, that can be present, if no other specific classifiers gain the upper hand.

Therefore, based on structural and semantic criteria, we claim that both *ge* and *zàg* are general classifiers in Hakka.

8. Conclusion

In this paper, we present a corpus-based account of the properties of the two general classifiers *ge* and *zàg* in Hakka using a collostructional analysis. We have shown that although they correlate to human-denoting and animal-denoting nouns respectively, exceptions remain which must be learned individually. We have also demonstrated the advantages of the collostructional analysis, in which relations between two lexical items within a construction can be quantified and measured. Collocational strength values are direct indicators of degree of acceptability in the classifier-noun combinations.

It is also evident from the distribution of the data that if a noun has a specific classifier, it will have relatively low collocational strength values for *ge* and *zàg*. Since Hakka is a classifier language in which classifiers are obligatory, a classifier must always be present between a determiner/numeral and a noun. For abstract entities or concrete objects lacking physical properties like size or shape, both *ge* and *zàg* can be used, though with varying degrees of acceptability.

This paper contributes to the study of classifiers in Hakka and the understanding of classifiers in general. Previous studies on classifiers focusing on their semantic properties usually provide long lists of nouns that collocate with certain classifiers without further explaining their relative acceptability with respect to those classifiers. The collocational strength values presented in this paper are good indicators of their degree of acceptability.

Furthermore, this study also sheds light on language teaching. When teaching the usage of classifiers, it is crucial to make sure learners know the limitations of general classifiers should a specific classifier exist for a certain noun. Lexical blocking is always at work.

In this study, we provide a big picture of the distribution of nouns collocating with the two general classifiers *ge* and *zàg* in Hakka in terms of collocational strength values. The three-way distinction in the collostructional analysis (attractive, neutral, and repulsive) is directly mapped to acceptability, though we have modified the cut-off values based on native speaker intuition: acceptable if the CS value is above or equal to 1.3, unacceptable if the CS value is below or equal to 0, marginally acceptable if otherwise. This measurement has the advantage of recognizing a continuum of acceptability in language. It is the relative acceptability that matters.

We also observe that although human-denoting and animal-denoting nouns are the majority among the nouns collocating with *ge* and *zàg*, unrelated nouns, especially abstract nouns, are still abundant. There is no way to pick up single semantic features characterizing nouns with high CS values in *ge* and *zàg*.

This leads to our finding that both *ge* and *zàg* are qualified as being general classifiers since they are more frequently used than other specific classifiers in terms of noun types and noun tokens. They are the default classifiers used by native speakers of Hakka if no specific classifiers are available.

Funding

The work in this research article was sponsored by the National Science and Technology Council, Taiwan (NSTC 111-2410-H-007-034-), whose financial support is gratefully acknowledged.

Acknowledgements

We would also like to express our sincere gratitude to the two anonymous reviewers and the editors for their invaluable feedback and suggestions, which helped to improve the quality of this paper.

References

Aikhenvald, Alexandra Y. 2003. *Classifiers: A Typology of Noun Categorization Devices*. Oxford, UK: Oxford University Press.

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

Chiu, Hsiang-Yun. 2007. Minnanyu he kejiahua de liangci: Yu guoyu bijiao [Measure words of Southern Min and Hakka: A comparison with Mandarin]. *Hsuan Chuang Humanities Journal* 7:175–206.

doi Fillmore, Charles J., Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64.3:501–538.

doi Frankowsky, Maximilian, and Dan Ke. 2016. Humanness and classifiers in Mandarin Chinese: A corpus-based study of anthropocentric classification. *Language and Cognitive Science* 2.1:55–67.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: The University of Chicago Press.

Goldberg, Adele E. 2003. *Construction at Work: The Nature of Generalization in Language*. Oxford, UK: Oxford University Press.

doi Gries, Stefan Th., and Anatol Stefanowitsch. 2004a. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9.1:97–129.

Gries, Stefan Th., and Anatol Stefanowitsch. 2004b. Covarying collexemes in the into-causative. *Language, Culture, and Mind*, ed. by Michel Achard and Suzanne Kemmer, 225–236. Stanford, CA: CSLI Publications.

doi Gries, Stefan Th., Beate Hampe, and Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16.4:635–676.

Hakka Affairs Council. 2022. *Taiwan Hakka Corpus*. Retrieved November 26, 2022, from https://corpus.hakka.gov.tw/

Her, One-Soon, and Chen-Tien Hsieh. 2010. On the semantic distinction between classifiers and measure words in Chinese. *Language and Linguistics* 11.3:527–551.

Huang, Han-Chun. 2021. Yi dapei jiegou fenxi kan keyu zhong si ge biaoshi "ren" de fenleici [On four human-denoting classifiers in Hakka: A collostructional analysis]. *Proceedings for the 30th Anniversary of Taiwan Languages and Literature Society*, ed. by Shu-chuan Chen and Min-hua Chiang, 177–193. Taipei: Taiwanese Languages and Literature Society.

doi Jackendoff, Ray. 1997. Twistin' the night away. *Language* 73:534–559.

doi Kay, Paul, and Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The *what's X doing Y?* construction. *Language* 75.1:1–33.

Li, Charles N., and Sandra A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley, CA: University of California Press.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Myers, James. 2000. Rules vs. analogy in Mandarin classifier selection. *Language and Linguistics* 1.2:187–209.

doi Stefanowitsch, Anatol, and Stefan Th. Gries. 2003. Collocations: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8.2:209–243.

doi Stefanowitsch, Anatol, and Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1.1:1–43.

Tai, James Hao-Yi. 1994. Chinese classifier systems and human categorization. In *Honor of Professor William S.-Y. Wang: Interdisciplinary Studies on Language and Language Change*, ed. by Matthew Y. Chen and Ovid Jyh-Lang Tzeng, 479–494. Taipei: Pyramid Press.

Tai, James Hao-Yi. 2006. Fenleici "wei" zai taiwan minnanyu yu keyu zhong de fanchou jiegou zhi bijiao [A comparison of categorial structures of classifier *bue*^{53}/*mi*^{24} in Taiwan Southern Min and Hakka]. *ON AND OFF WORK: Festschrift in Honor of Professor Chin-Chuan Cheng on his 70th Birthday*, ed. by Raung-Fu Chung, Hsien-Chin Liou, Jia-Ling Hsu and Dah-an Ho, 57–73. Taipei: Academia Sinica.

Tai, James Hao-Yi, and Lianqing Wang. 1990. A semantic study of the classifier *tiao*. *Journal of the Chinese Language Teachers Association* 25:35–56.

Tai, James Hao-Yi, and Li-wen Wu. 2006. Taiwan sixian keyu liangci "wei" de fanchou jiegou [Categorical structure of the classifier *mi*^{24} 'tail' in Sixian Hakka]. *Language and Linguistics* 7.2:501–521.

Zhang, Niina Ning. 2013. *Classifier Structures in Mandarin Chinese*. Berlin & Boston: De Gruyter Mouton.

Zubin, David A., and Mitsuaki Shimojo. 1993. How 'general' are general classifiers? With special reference to *ko* and *tsu* in Japanese. *Proceedings of the Nineteenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Semantic Typology and Semantic Universals*, ed. by J. Guenter, B. Kaiser and C. Zoll, 490–502. Berkeley, CA: Berkeley Linguistics Society.

Appendix. Statistics of the 116 nouns in the construction

#	N	F(N)	F(*ge*)	F(*zàg*)	CS(*ge*)	CS(*zàg*)	#	N	F(N)	F(*ge*)	F(*zàg*)	CS(*ge*)	CS(*zàg*)
1	*ngǐn*	219	0	16	−32.0	−13.7	59	*sǎ*	8	2	1	0.0	−0.3
2	*ngied*	121	56	65	4.9	9.1	60	*guì-è*	8	1	7	−0.3	3.1
3	*fa*	79	0	0	−11.2	−11.0	61	*vùg*	8	1	3	−0.3	0.2
4	*lai-è*	66	55	11	20.4	−1.3	62	*lò-moi*	8	8	0	4.4	−0.9
5	*sù*	63	1	48	−7.4	15.1	63	*sǎ-gó*	8	0	0	−0.9	−0.9
6	*se-ngǐn-è*	58	51	7	21.1	−2.1	64	*miěn-iǒng*	8	0	0	−0.9	−0.9
7	*sii* (character)	54	7	45	−1.9	16.9	65	*tǔng-fá-su*	8	0	0	−0.9	−0.9
8	*moi-è*	52	40	11	12.7	−0.5	66	*tǔng-iěu*	7	0	0	−0.7	−0.7
9	*ngiǔn*	45	32	0	8.8	−6.2	67	*xien*	7	0	0	−0.7	−0.7
10	*sii-qǐn*	39	1	0	−4.1	−5.3	68	*xím*	7	3	1	0.4	−0.2
11	*lu*	32	0	0	−4.4	−4.2	69	*se-lǒng*	7	7	0	3.9	−0.7
12	*sii* (matter)	29	0	1	−4.0	−2.8	70	*ňg*	7	0	1	−0.7	−0.2
13	*giòg*	28	0	18	−3.8	4.3	71	*gòg*	7	0	1	−0.7	−0.2
14	*fá*	28	0	0	−3.8	−3.8	72	*sǎm-sǔ*	7	0	5	−0.7	1.7
15	*se-moi-è*	27	17	10	3.9	0.5	73	*bìd*	7	0	0	−0.7	−0.7
16	*fu-ngǐn-gá*	27	23	4	9.1	−0.7	74	*bu-è*	7	0	0	−0.7	−0.7
17	*gièu*	22	1	17	−1.8	5.8	75	*lòn*	7	1	1	−0.2	−0.2
18	*sǐi-gién*	21	0	2	−2.7	−1.1	76	*ngie-gúng*	7	1	6	−0.2	2.6
19	*tiěn*	20	1	0	−1.7	−2.7	77	*gá*	7	5	2	1.7	0.0
20	*gu-sii*	20	7	7	0.3	0.3	78	*sún*	7	7	0	3.9	−0.7
21	*sám*	20	0	0	−2.7	−2.7	79	*fǔ-lǐ*	7	0	3	−0.7	0.4
22	*sén-è*	19	0	17	−2.5	7.6	80	*nǎm-ngǐn*	7	7	0	3.9	−0.7
23	*mùg-zú*	18	1	4	−1.5	−0.1	81	*mun-tǐ*	7	3	4	0.4	1.0
24	*ngiǔ*	18	2	11	−0.7	2.5	82	*se-lai-è*	7	2	5	0.0	1.7
25	*moi*	18	18	0	10.0	−2.2	83	*vi-sò*	7	2	5	0.0	1.7
26	*miang*	17	0	0	−2.3	−2.3	84	*vú-gièu*	7	0	2	−0.7	0.0

Appendix. *(continued)*

#	N	F(N)	F(*ge*)	F(*zàg*)	CS(*ge*)	CS(*zàg*)	#	N	F(N)	F(*ge*)	F(*zàg*)	CS(*ge*)	CS(*zàg*)
27	*sïi-jièd*	17	13	4	4.4	0.0	85	*gúng-iěn*	7	2	5	0.0	1.7
28	*sag-těu*	16	1	2	−1.3	−0.6	86	*bú-ngiǒng*	6	4	2	1.3	0.2
29	*gièu-è*	16	0	13	−2.0	4.9	87	*fò-fǒng*	6	0	6	−0.7	3.4
30	*lò-ngǐn-gá*	15	12	0	4.4	−1.8	88	*miǎng-è*	6	0	5	−0.7	2.1
31	*lò-tái*	15	14	1	6.8	−1.1	89	*tǔng-hog*	6	5	0	2.1	−0.7
32	*ňg-è*	15	0	0	−1.8	−1.8	90	*xiu-cǒi*	6	6	0	3.3	−0.7
33	*pěn-iú*	14	13	0	6.2	−1.8	91	*fǔng-báu*	6	0	6	−0.7	3.4
34	*heu-sáng-è*	14	10	1	3.1	−0.9	92	*san-è*	6	0	0	−0.7	−0.7
35	*iěn-ngoi*	14	14	0	7.8	−1.8	93	*biàng-è*	6	0	3	−0.7	0.4
36	*hǒ-ba*	13	2	0	−0.3	−1.6	94	*guèd-gá*	6	1	5	0.0	2.1
37	*gié*	12	0	12	−1.6	6.7	95	*hi*	6	1	0	0.0	−0.7
38	*se-á-moi-è*	12	12	0	6.7	−1.3	96	*qiěn*	6	2	0	0.2	−0.7
39	*fu-kièu-miǎng-pú*	12	0	0	−1.6	−1.3	97	*gùg*	6	1	0	0.0	−0.7
40	*gié-è*	11	0	11	−1.4	6.2	98	*su-è*	6	0	0	−0.7	−0.7
41	*sò-cai*	11	7	4	1.9	0.3	99	*xin-è*	6	0	0	−0.7	−0.7
42	*sù-zìi*	11	2	7	−0.1	1.9	100	*gié-mǎ*	6	0	6	−0.7	3.4
43	*nǎm-è-ngǐn*	11	11	0	6.1	−1.4	101	*diáu-è*	6	0	5	−0.7	2.1
44	*cá-è*	11	0	0	−1.4	−1.4	102	*bu*	6	0	0	−0.7	−0.7
45	*se-ngǐn*	11	10	1	4.6	−0.5	103	*tai-sii*	6	0	6	−0.7	3.4
46	*gí-fi*	10	7	3	2.2	0.0	104	*síi*	6	0	0	−0.7	−0.7
47	*ngìd-è*	10	1	6	−0.5	1.5	105	*hiúng-ti*	6	6	0	3.3	−0.7
48	*heu-sáng-ngǐn*	10	6	3	1.5	0.0	106	*pǐ*	6	0	0	−0.7	−0.7
49	*diáu*	10	0	10	−1.1	5.6	107	*lò-fo-è*	6	6	0	3.3	−0.7
50	*hěu-è*	10	0	8	−1.1	3.1	108	*xín-sáng*	6	5	0	2.1	−0.7
51	*hǎi*	9	0	1	−1.2	−0.3	109	*lo-cù*	6	0	4	−0.7	1.3
52	*á-gó*	9	9	0	5.0	−1.1	110	*sò-sìi*	6	0	0	−0.7	−0.7
53	*gó-è*	9	0	0	−1.2	−1.1	111	*lug-è*	6	0	0	−0.7	−0.7
54	*xím-kiú*	9	6	3	1.8	0.1	112	*sïi-toi*	6	1	5	0.0	2.1
55	*lí-bai*	9	4	5	0.6	0.9	113	*biàg*	6	0	0	−0.7	−0.7
56	*ngiǔ-è*	9	0	4	−1.2	0.6	114	*cò*	6	0	0	−0.7	−0.7
57	*ti*	9	1	0	−0.3	−1.1	115	*ién*	6	0	0	−0.7	−0.7
58	*giéu*	8	0	0	−0.9	−0.9	116	*sii-gie*	6	6	0	3.3	−0.7

Note. # stands for ranking of Ns (nouns), sorted in descending order by f(N), the frequency of N in the [Det/Num-Cl-N] construction; f(*ge*) and f(*zàg*) stand for the co-occurrence frequencies of N with *ge* and *zàg* in the construction; CS(*ge*) and CS(*zàg*) stand for the collocational strength values of N with *ge* and *zàg* in the construction.

Address for correspondence

Han-Chun Huang
Department of English Instruction
National Tsing Hua University
Hsinchu, TAIWAN
hanchun@gapp.nthu.edu.tw
https://orcid.org/0000-0002-5072-7940

Publication history

Date received: 30 January 2023
Date revised: 6 April 2023
Date accepted: 11 April 2023