Collocation analysis of news discourse and its ideological implications

Huei-ling Lai
National Chengchi University

This study investigates the use of an ethnic term in news discourse from linguistic, discursive, and social-cultural aspects. A more rigorous computational procedure than hitherto used is employed to measure the collocational strength of collocates in news corpora. The results indicate diversified distributions of the collocates regarding their frequency, distance, and semantic connections. The findings enhance the meaning specificity of the term by revealing the characterized reference of this ethnic group, the trends in the choice of news topics, and the ideological representation of this ethnic group in a wider social-cultural context. The findings deepen an understanding of news discourse as the representations of the minority ethnicity in the news media are analyzed through three layers – the linguistic, the discursive, and the social-cultural context. A more precise method of analyzing news texts uncovers ideological effects brought about by media, in turn implying different construal of newsworthiness in news discourse.

Keywords: collocation strength, Hakka ethnic groups, news discourse, ideological representation, discursive

1. Introduction

Language use as a form of social action is dynamically shaped and reshaped to construct or to manipulate social realities. Analyzing the language use in news corpora helps understand the structure, the discourse and the ideology represented in news media (Breeze 2016). Language use as meaning-making suggests that news discourse needs to be considered as multi-layered (Blum-Kulka and Hamo 2011; Jiwani and Richardson 2011). Lexical or grammatical expressions encode textual information. However, while the denotative meanings of linguistic devices can uncover textual values, ideological effects aroused can be underestimated or mispresented if the use of linguistic resources is taken at face value,

as linguistically-encoded meanings can be under-determined. Hence, in addition to linguistic devices, social-cultural contexts that provide communal common ground for interpretation are crucial for reinforcing a better understanding of newsworthiness (Potts et al. 2015).

From the discursive perspective, what makes certain stories worthy of being news has been constructed through news discourse; the question for analysis is how to use language resources to show the ways in which an event is considered newsworthy (Bednarek and Caple 2014). Corpus linguistic techniques have been employed to analyze news texts and to identify linguistic resources that represent news values, including the indicators, their definitions, and corresponding linguistic devices (Potts et al. 2015; Bednarek and Caple 2014). These studies bring to light the significance of linguistic analysis to newsworthiness. In terms of ideological effects projected by news media, Van Dijk (1992) proposed the ideological square whereby a conceptual apparatus that dominates texts about ethnic others is operated. The operation can be observed through the study of various linguistic aspects of a text, and is defined by a positive self-presentation and at the same time a negative other-presentation. News media provides a strategic ground in which the choice of the words used in naming or portraying people is especially significant in analyzing "the positive presentation and negative other-presentation integral to the 'ideological square'" (Jiwani and Richardson 2011, 243).

Among the linguistic devices available for the analysis of news discourse, collocation analysis that identifies collocates and phrases around significant words is considered useful as the presence of recurring patterns can identify specific indicators established around words immediately adjacent to each other (Baker et al. 2013; Potts et al. 2015). However, "since it [collocation analysis] focuses on the *immediate co-text* (italics original), rather than the whole news article, it can only provide a partial view of the news values constructed around a particular issue" (Potts et al. 2015, 161). In other words, the identification of only next-to-each-other phraseologies may miss significant recurring collocates that occur farther away from the key word in a news text. Hence, a more rigorous computational procedure is needed to analyze the strength of the collocations in a news text where collocates can be found beyond the immediate co-text. This study – investigating the use of a specific ethnic term *kèjiā* 'Hakka' used in the news corpora in Taiwan – is an endeavor to demonstrate a more rigorous measurement of collocational strength in news discourse on the one hand, and to link the analytic results to the ideological effects toward the Hakka ethnic groups created by the news media on the other. It is also implied that a better understanding of a text can be accomplished by a multi-layered manner.

The word *kèjiā* 'Hakka', literally meaning 'guests' in Mandarin Chinese, symbolizes the social status of the Hakka ethnic groups, and at the same time carries

immense ethnic and cultural complexities with diasporic characteristics due to historical migration (Chappell and Lamarre 2005). The word is conceptually under-specified, yet discursively salient as it can represent anything associated with the Hakka ethnic groups, such as culture, language, or people in a news text. Measurement of collocational association that quantifies the strength and the reliability of the term is to be implemented so as to reveal how the under-specified referential meaning of *kèjiā* 'Hakka' can be strengthened by its noun and verb collocates in nearby news texts. The purpose is two-fold. First, the definition of the concept of ethnicity is quite imprecise, as its features, such as language, religion, country of birth, and family origins, cannot be easily measured (Bhopal 2004, 442). News media provides a strategic ground in which the choice of the words used in naming or portraying people is especially significant in analyzing the positive ideological presentation and negative other-presentation. Nouns as referring expressions that represent things and entities can reveal the referential management strategies used in news media – anchoring to the referential strategies of the ideological square in regard to how certain ethnic groups are referred to. In addition, verbs as predication that represent states of affairs can reveal the significant information to be communicated – anchoring to the predicational strategies of the ideological square in regard to how certain ethnic groups are represented. The analysis of noun and verb collocates can hence bring to the fore the textual and discursive manifestation of news events in the news discourse. Second, examination of the language use in the news media over a certain period of years helps foreground the news topics which are most frequently covered in association with the Hakka ethnic groups, providing a more rounded picture of the social-cultural practices in the whole society. The analysis will encompass multiple layers: the referential meaning from linguistic collocation, the textual theme manifestation from the discursive context, and the implicit construal from the social-cultural context. The results will bring to light how ideological effects are established in the news discourse as certain news topics are considered worthy of being reported.

In what follows, Section 2 presents an overview of the studies of collocation analysis in the extant literature, Section 3 introduces the data and methods, Section 4 presents the results, Section 5 presents the discussion and implications, and Section 6 concludes the study.

2.   Collocation analysis

The description of apparent multiple associations of the meaning of a single word form have always been a basic problem in lexical semantics (Cruse 1986). The

complexities lie in the fact that the meaning of any lexical item can be modulated in various ways in any distinct context in which it is used. Semantic vagueness occurs when a lexeme is under-specified, and hence has a lower level of semantic specificity (Cruse 1986). Denotative vagueness refers to a situation in which a particular element of the meaning of a word needs to be provided by the context in which the word occurs (Cruse 1986). For instance, words of kinship relations and occupation are often underspecified. The sentence *His favorite aunty is coming for dinner* is low in information because the word *aunty* is underspecified with regard to whether there is either a motherhood or fatherhood relationship. Another kind of indeterminacy has to do with referential identity. For example, the word *elbow* does not have a clear boundary of its range, so does the color term *blue*, which is fuzzy in terms of its possible range of hues.

To resolve such a lack of information exhibited by a semantically vague term requires an increase in its specificity. One way of increasing specificity is to add syntagmatic modifiers. The syntagmatic arrangement of two or more lexical items is a very fundamental strategy for use in the creation of novel expressions, and is widely used to modulate information for specificity. Sequences of lexical items can thus give rise to semantically more fine-grained compounds, collocations or lexically complex idiomatic expressions. For instance, *curry chicken* delimits the meaning of *chicken* to meat instead of animal. Such a syntagmatic co-occurrence of two nouns formulating a modifier-head compound is also common in Mandarin Chinese. Examples such as *tǐwēn* 'body temperature' or *jīdàn* 'chicken egg' can illustrate.

Words are less freely co-occurring with each other, but rather tend to be used in a limited number of grammatical structures or patterns (Sinclair 2004). Even for words whose meanings are under-specified, it is found that they co-occur with other words not in a scattered manner, but in a more fixed way than might be imagined. Such a fixed sequence of two or more words that corresponds to a conventionally understood meaning is called a collocation. Dating back to the earliest definition advocated by Firth (1957), "collocations are to be defined as the habitual and recurrent juxtaposition of semantically related words" (Bartsch and Evert 2014, 48). The significant investigation of collocations on a larger scale has become viable thanks to the construction of corpora of substantial sizes, leading to wider possibilities in the study of collocations and their applications. At the same time, considerable advances in computational methods and statistical approaches have helped effectively find and measure collocations. (McEnery and Wilson 2001; Gries 2013; Bartsch and Evert 2014). To illustrate, in order to exactly measure how the co-occurrence of two lexical items can to be considered as a collocation, Sinclair et al. (2004) propose three criteria for identifying collocations: distance, frequency, and exclusivity. Distance specifies the span, i.e., the colloca-

tion window, around a node word and its collocates. The distance of the collocates from the node can be as short as one word or as long as a cluster of four or five words on each side of the node. Frequency is an important indicator of the typicality of a word association. For instance, the noun *fire* occurs frequently with the preposition *on* and therefore *on fire* is a common chunk in English. Nevertheless, the relationship between *fire* and *on* is not exclusive since *on* can co-occur with many other nouns as in *on foot*. On the other hand, *fire* is much more strongly and exclusively connected with the noun *forest*. When the word *fire* appears in a text, there is a large probability that the preceding word may be *forest*. Endeavors to provide a more precise computation of association measures that quantify the strength and the reliability of collocation continue in the extant literature with significant findings (e.g., Manning and Schütze 1999; Pecina 2010). To illustrate, Manning and Schütze (1999) propose the principal approaches to find and measure the strength of collocations, including the selection of collocations by frequency, the mean and variance of the distance between the focal word and its collocates, hypothesis testing, and mutual information. Pecina (2010) compares more than 80 measures for use in the extraction of collocations, and finds out that no single measure can be served for all-purpose measurement.

The meaning of a collocation is assumed to be compositionally derived from the combination of the meanings of the co-occurring collocates. Compositionality allows a one-to-one correspondence between the syntactic structures and the semantic representations. For instance, the meaning of *body temperature*, which refers to a kind of temperature, is the combination of the meaning of *body* and of the meaning of *temperature*. It is suggested that there is a requirement for an increase in the specificity of a semantically vague term so as to resolve the lack of information exhibited by such term; hence, a plausible hypothesis is that information specificity can be elaborated through investigation of its syntagmatic co-occurring collocates. Hence, how to precisely measure the strength of a collocation so as to improve the information specificity is important.

Given the many endeavors of probing into collocation analysis, full agreement has yet to have been reached, however, regarding the definition of collocation, as stated in Gries (2013, 138): "the notion of 'collocation' is probably best characterized as a radial category whose different senses are related to each other and grouped around one or more somewhat central senses, but whose senses can also be related to each other only rather indirectly". Even so, Gries (2013, 138–139) points out that these studies of collocation analysis have proposed useful criteria and done experiments in various dimensions, including the words to be observed, the number of words to be counted, the frequency of the occurrences of these words, the distance of the collocates, the degree of lexical and syntactic flexibility of the collocates, and the degree of semantic compositionality

of the collocates. Furthermore, in order to manage larger corpora data, many studies have focused on "how to best extract, identify, and measure collocations given their frequencies of co-occurrence" in the last fifty years or so, as noted by Gries (2013,139). In more recent extant literature, Seretan (2011), employing syntax-based collocation extraction with a focus on using linguistic tools for the corpus-based identification of collocations, argues for better efficiency with the use of syntax-driven extraction as an alternative to co-occurrence-driven extraction. Huang et al. (2015), based on Information-based Case Grammar, show the importance of integrating existing grammatical information so as to tackle the automatic extraction of grammatical knowledge from large corpora. Yang et al. (2015), employing Chinese mono-syllabic characters as seed words and utilizing higher Pointwise Mutual Information collocations, have identified compounds that have different semantic polarities depending on the contexts. Brezina et al. (2015) develop a new tool *GraphColl* that builds collocation networks from user-defined corpora and demonstrate how collocation networks can provide important insights into meaning relationships in language and into the relationship between the discourse and the language users. Up until now, it seems that more and more techniques utilizing different parameters have been proposed to cope with collocation while various corpora of different kinds are becoming larger and more diversified in nature. Still, even though reasonably good results have been obtained, Gries (2013,159) contends that methods or measures can be improved by further refinement since "collocation has been, and will remain, one of the most important concepts in corpus linguistics".

From the extant literature, it should be fair enough to say that there seems to be no one best method or approach alone for collocation analysis. For the purpose of the current study, a four-way combination for the analysis of Chinese collocation is employed with the aim of providing a more rigorous procedure and hence a more solid result for discursive and social-cultural interpretations. Based on Firth's (1957) very first idea of collocation analysis, the methods of the study also adopt approaches from Gries (2013) and Bartsch and Evert (2014). Essentially, a standard range of well-known association measures – including co-occurrence frequency, mean distance and standard deviation, Pearson's chi-squared test, and different variants of mutual information – is adopted, and then the results are compared and contrasted. The outcome of such a more rigorous procedure is to investigate two questions. The first question is how collocates with a stronger strength within a news text can be seriously identified by the four types of method so as to help specify the referential meaning of *kèjiā* 'Hakka'. Then, the second question is to examine what textual themes can be discursively drawn from the news corpora and what profound significance can be demonstrated within the social-cultural context. The next section will present the data and the methods.

3. The data and methods

3.1 Materials

The corpora data come from newspaper online databases in Taiwan – *Udn Data* (including the United Daily News, Economic Daily News, and United Evening News), *Knowledge Management Winner* (KMW, from the *China Times*), *Liberty Times Net* (from the *Liberty Times*), and *Apple Daily* (*Apple Daily*, Taiwan) – the four main newspapers by circulation in Taiwan. The United Daily News, launched in 1951, provides news on current events, and is deemed as strongly pan-blue; it covers more news related to the KMT (Chinese Nationalist Party), while providing less news about the DPP (Democratic Progressive Party). The Economic Daily News and the United Evening News are subdivisions of the United Daily News Group. The Economic Daily News, launched in 1967, covers news including commerce, industry analysis, equity and bond markets, commodities, creative management skills and technological innovation. The United Evening News, launched in 1988, reports the latest political, social, and financial and leisure news. The China Times, launched in 1950, is also considered to align more with the KMT than with the DPP. The Liberty Times, launched in 1980, tends to give wider news coverage to the DPP instead of to the KMT. The Apple Daily (Taiwan), launched in 1995 by a Hong Kong-based company, shows relatively less political alignment. The selected database can be considered quite representative and balanced due to the high reputation for soundness in reporting, wide circulation, and different political alignments of the newspapers included in it. The news articles were all extracted from the period beginning 1st January, 2005 until 31st December, 2015, inclusive.

3.2 Procedure

Searching for the target node in the news corpora

Articles which contain the keyword *kèjiā* were extracted to establish the database. In total 168,116 tokens of *kèjiā* were found; the distribution of the data from the four newspapers is given in Table 1. Then, all of the articles were coded with a metadata format including title, subtitle, main topics, subtopics, and author information. Consistency and reliability were maintained through two rounds of cross-coding and revisions. Table 2 gives the nature of all of the news topics and subtopics.

Table 1. The distribution of *kèjiā* in four newspapers

Database	Time	Number of articles (selected / total)	Number of tokens (selected / total)
Udn	2005/01/01–2015/	21847 / 320745	77174 / 544562
KMW	12/31	13793 / 282314	45057 / 325711
LTN		8701 / 167470	32027 / 190476
AD		4385 / 129603	13858 / 1678920
Total		**48726 / 900132**	**168116 / 2739669**

Table 2. Topics and subtopics

Topic	Subtopic	Number of articles
Politics	central, local, political party, politicians	6058
Society	community constructions	4346
Lifestyles	health information, family, tourism, climate & environments, human characters	15881
Education & Technology	education, technology, human characters	4365
Global issues	international, cross-strait	706
Business & Economics	human characters	868
Sports	events, human characters	156
Entertainment	events, human characters	2292
Artistic activities	human characters	14054

Segmenting the text

Word segmentation that separates words in a text is considered an important step for Chinese since Chinese words can be composed of multiple characters in a sequence. Sentences and words need to be divided correctly to obtain word frequency and find collocations. Patricia tree (PAT Tree) was used to segment texts and retrieve collocations. With the text segmented, we then found *kèjiā*'s collocates, and detected their collocational strength.

Computing the strength of a collocation

The statistical evaluations of *kèjiā* and its collocates were calculated in terms of four dimensions: frequency, mean and variance of the distance between *kèjiā* and its collocates, chi-square test, and pointwise mutual information.

First, in order to find the content of the collocates of *kèjiā*, we first removed all of the function words such as conjunctions, articles, auxiliary verbs, and pronouns, and also words with a frequency of less than 100. We then chose the ten most frequently-occurring nouns. Second, although frequency-based search works well for fixed phrases, many collocations consist of two words that may not stand next to one another and the fixed phrase approach will not work if the distance between two words is not constant. However, we will be able to determine whether two words form a collocation or not if there is sufficient regularity in the patterns in which they occur. One way of discovering the relationship between two words is to compute the mean and variance of the distances between the two words. This information can be used to discover collocations by looking for pairs with low variances. If the differences are randomly distributed, then the variance would be high. The use of this variance-based method is appropriate for finding combinations of words that are in a looser relationship than fixed phrases and that are variable with respect to intervening material and relative positions. A low variance means that the two words usually occur at about the same distance from the target node. No variance indicates that the two words always occur at exactly the same distance from the target node. Third, high frequency and low variance can be accidental. Two words may co-occur many times by chance, and thus they will not form a collocation. Hence, a further aim is to compare the observed frequency of occurrence with the expected frequency of occurrence under the independence hypothesis. A chi-square test can be applied to evaluate how likely it is for two words to co-occur by chance. This test is calculated by summing the square of the differences between observed and expected frequencies of occurrence divided by the expected frequency. If the value is larger than the critical value, the independence hypothesis, which suggests that the two words co-occur by chance, is rejected. In other words, we can claim that the two words form a collocation. Last, pointwise mutual information (PMI) is to measure the mutual information between the occurrence of two words, namely, how much one of the words tells us about the other. PMI works well for locating empirical collocations since it expresses the nature of the association (attraction vs. repulsion) and the strength of the association by a logarithmic measure. The value of PMI maximizes when the two words are greatly associated.

Coding of verb collocates

The next task was to code the co-occurring verbs, categorize their object themes, and identify the associated news topics. The fifteen most frequent verbs – including verbs of presenting or characterizing such as *zhǎnxiàn* 'show', verbs of creation or transformation such as *chénglì* 'establish', and verbs of improving such as *tuīguǎng* 'promote' – were then selected for a more in-depth analysis. The purpose

for this further examination is to identify what specific Hakka affairs are reported by the short eye-catching news headlines, such as *tuīguǎng kèjiā* 'to promote Hakka'. While such a headline often attracts the attention of the reader audiences, the referent to be promoted cannot be determined by this verb phrase structure which contains the verb and *kèjiā* alone. Curiosity arises as to what Hakka issues are reported when verbs such as *tuīguǎng* 'promote', *jǔbàn* 'hold', or *zhǎnxiàn* 'show' collocate with *kèjiā*. Two aspects are considered. First, for each of the verbs, the noun following *kèjiā* was coded so as to identify the referential domain of the object *kèjiā* + Noun. Second, the news topic associated with each verb was identified. The coding process was conducted by two groups of four research assistants who were graduate students with training in linguistics for at least two years. The two groups exchanged their data for a second round of examination to guarantee the credibility of the results.

4.   Results

First of all, Table 3 shows the ten most frequently-occurring noun collocates with *kèjiā* in terms of the four dimensions, respectively.

The result shows that some collocates occur at a high frequency and show strong semantic relations; some occur at a long distance and show strong semantic connections; others occur at a high frequency, yet at a long distance and show weak semantic connections; still others show stronger semantic connections, but occur at a low frequency and at a long distance. Some variations are observed, showing interesting correlations. First, *xiāngqīn* 'fellow folks', *zúqún* 'ethnic groups', *wénhuà* 'culture', *chuántǒng* 'tradition', *měishí* 'cuisine', and *tónghuā* 'tong blossoms' steadily score higher than other cases in terms of the four calculations, indicating that they carry a stronger collocational strength with *kèjiā*. Next, the case *yuánqū* 'resort areas' occurs at a farther distance from *kèjiā*, but at a high collocation frequency. Third, the collocate *yìshù* 'arts' occurs at a high frequency, but scores quite low in the other three indicators, showing clearly that it does not have a strong association with *kèjiā* by meaning or distance. These results come as no great surprise. Hakka fellow folks, traditions and cuisines are items that profile Hakka festival activities and hence are often reported in the news. In addition, a number of popular resort areas are in areas that are historically associated with Hakka ethnic groups and provide Hakka-style or Hakka-theme artistic performances, such as singing, dancing or the making of Hakka cuisines associated with the festivals, and hence these resorts and activities are often highlighted in the news.

Table 3. The noun collocates based on the four dimensions

Ranking		Frequency		Mean (SD)		Pearson's chi-square		PMI
1	wénhuà 'culture'	10771	xiāngqīn 'local people'	0.17 (4.31)	wénhuà 'culture'	23547.91	yìngjǐng 'stiff neck'	2.3
2	huódòng 'activity'	3056	tónghuā 'Tung blossom'	0.27 (6.61)	chuántǒng 'tradition'	7073.94	měishí 'cuisine'	1.78
3	chuántǒng 'tradition'	3050	yìmín 'brave people'	0.34 (8.17)	měishí 'cuisine'	6408.82	wénhuà 'culture'	1.78
4	měishí 'cuisine'	2444	měishí 'cuisine'	0.43 (5.22)	yuánqū 'resort areas'	5441.67	yuánqū 'resort areas'	1.76
5	yuánqū 'resort areas'	2121	guānguāng 'tourism'	0.45 (13.04)	xiāngqīn 'local people'	3479.16	chuántǒng 'tradition'	1.7
6	tónghuā 'Tung blossom'	1882	wénhuà 'culture'	0.53 (4.51)	zúqún 'ethnic group'	2553.92	xiāngqīn 'local people'	1.7
7	zúqún 'ethnic group'	1577	chuántǒng 'tradition'	0.60 (5.37)	tónghuā 'Tung blossom'	2491.11	zúqún 'ethnic group'	1.43
8	xiāngqīn 'local people'	1511	kèyǔ 'hakka language'	0.68 (14.13)	huódòng 'activity'	1808.93	biǎoyǎn 'performance'	1.28
9	biǎoyǎn 'performance'	1365	zúqún 'ethnic group'	0.98 (5.99)	biǎoyǎn 'performance'	1775.41	tónghuā 'Tung blossom'	1.27
10	yìshù 'art'	988	yìngjǐng 'stiff neck'	1.30 (5.71)	yìngjǐng 'stiff neck'	1605.11	yìmín 'brave people'	1.2

These syntagmatic co-occurring collocates increase the semantic specificity of the vague term *kèjiā*, giving a more concrete specification of the referential meaning expressed by the concept Hakka, pinpointing down the reference reported in the news. The following examples give the context in which the collocations occur in the nearby co-texts in the news articles. The underlined collocates explicate what Hakka topic is reported.

(1) 畢竟總是有**文化**需要被保護，包括**客家**、原住民等　（LTN-2006-03-22）
     *bìjìng zǒngshì yǒu **wénhuà** xūyào bèi bǎohù bāokuò **kèjiā** yuánzhùmín děng*
     after all-always-exist-culture-need-passive marker-protect-include-Hakka-
     aborigine
     'After all, there are always cultures that need to be protected, including those of
     the Hakka and aborigines, among others.'

(2) **客家**影音網路平台服務網友和海外**鄉親**　　　　　　（LTN-2011-07-16）
     *<u>**kèjiā**</u> yīngyīn wǎnglù píngtái fúwù wǎngyǒu hàn hǎiwài **xiāngqīn***
     Hakka-audio-video-internet-platform-serve-internet-user-and-overseas-fellow
     folks
     'The Hakka Audio and Video Internet Platform serves the internet users and
     fellow folks overseas.'

(3) 提起**客家**，不能不談**桐花**這個美麗的印記　　　　（LTN-2005-06-03）
     *tíqǐ **kèjiā** bù néng bù tán **tónghuā** zhè ge měilì de yìnjì*
     mention-Hakka-not-can-not-discuss-tong flower-this-CL-beautiful-icon
     'When it comes to Hakka, there can be no ending of the discussion without
     mentioning Tong flower – the beautiful icon.'

(4) 為東南亞鄉親演繹<u>**客家**</u>歌謠、戲曲、語言文化、<u>美食</u>與產業。
     　　　　　　　　　　　　　　　　　　　　　（LTN-2015-03-02）
     *wèi dōngnányǎ xiàngqīn yǎnyì **kèjiā** gēyáo xìqǔ yǔyán wénhuà **měishí** yǔ
     chǎnyè*
     for-southeast Asia- folks-perform-Hakka-song-drama-language-culture-
     cuisine-and-industry
     'To display Hakka songs, dramas, languages, culture, cuisine and industry to
     the Southeast Asia fellow folks'

What is worthy of attention is the two items *yìmín* 'brave people' and *yìngjǐng* 'stiff-neck, toughness/indomitability'. The item *yìmín* 'brave people' ranks third by Mean (SD), but falls out of the top ten by frequency. The two words do not show a high frequency with the target word *kèjiā*. However, the item *yìmín* occurs at a close distance with *kèjiā*, in contrast to *yìngjǐng*, which does not occur at a close distance. Another difference lies in their PMI values. While *yìngjǐng* shows the highest PMI value, *yìmín* comes as tenth in terms of its PMI value. These two items can be considered as culturally-loaded words of Hakka. The first word *yìmín* has to do with a major type of Hakka religious temple, which worships those brave Hakka people who fought for their survivals during the early land reclamation and cultivation period, hence serving as a symbolic representation of the loyal and indomitable spirit of the Hakka ethnic groups. Every July, there is a Hakka *yìmín* festival with worship and celebrations. Since the festival only occurs once every year, the word *yìmín* hence does not enjoy a high frequency. However,

since it is a word characteristic of Hakka culture, it enjoys a high PMI value. The second word *yìngjǐng* is used to portray the most prototypical character associated with the Hakka ethnic groups. Hence the two words show high PMI values and strong collocation strength with *kèjiā* since their occurrence explicates that Hakka news is being reported. The example given in (5) illustrates such a co-occurrence.

(5) **義民**廟供奉的義民爺是**客家**族群忠義**硬頸**精神的代表

(LTN-2006-08-27)

*yìmín miào gòngfèng de yìmínyé shì **kèjiā** zúqún zhōngyì **yìngjǐng** jīngshén de dàibiǎo*

Yimin Temple-worship-nominalizer-Lord of the Righteous-be-Hakka-ethnic group-loyal-righteous-indomitable-spirit-nominalizer-symbol

'the Lord of the Righteous worshipped in the Yimin Temples is the symbolic representation of the loyal and indomitable spirit of the Hakka ethnic groups.'

Next, Table 4 shows the 15 chosen verbs which collocate with *kèjiā* in terms of the four dimensions, respectively. Verbs including *jǔbàn* 'hold', *zhǎnxiàn* 'show', *chénglì* 'establish', and *tǐyàn* 'experience' score relatively higher than the other cases in terms of the four dimensions. Particularly, *zhǎnshì* 'display' shows the second highest PMI value than the other collocates. However, this word does not occur at a high frequency, nor does it occur at a close distance to *kèjiā*. Some of the verb collocations seem quite synonymous as illustrated by *zhǎnxiàn* and *chéngxiàn*, which may reasonably be considered as a result of the writing style or personal word-choice of the news reporters. However, after double-checking the data and examining the news being reported, the factor of personal writing styles can be regarded not that influential for the following two reasons. First, the coverage of data ranges over eleven years with a total of 900,132 articles as provided by Table 1. These many articles were produced by more than forty different reporters. The verb collocates which have been identified can be regarded as evenly distributed among them. Second, what is essentially examined here is what theme or topic that is related to Hakka is being displayed or shown, no matter whether it is *zhǎnxiàn* or *chéngxiàn* that is being used to indicate such topic or theme. The high frequency of the occurrence of these verbs of display is what needs to be further delved into.

Further examination of the referential meaning denoted by the object theme associated with *kèjiā* expands the meaning domain and helps identify more aspects of the concept of Hakka projected discursively in the whole news article. Table 5 shows the correlation between the verb collocate and the co-occurring object theme *kèjiā* + Noun.

The result identifies what Hakka matters are associated with each verb in a news report. For instance, *guīhuà* 'to plan' is mostly related to Hakka cultural

Table 4. The verb collocates of kèjiā based on the four dimensions

Ranking	Frequency		Mean (SD)		Pearson's chi-square		PMI	
1	jǔbàn 'hold'	3672	zhǎnshì 'display'	−1.56	zhǎnxiàn 'show'	1682.06	tuīguǎng 'promote'	1.42
2	zhǎnxiàn 'show'	1841	zhīchí 'support'	−1.44	jǔbàn 'hold'	798.46	zhǎnxiàn 'show'	1.37
3	chénglì 'establish'	1641	fāzhǎn 'develop'	−0.99	chuánchéng 'inherit'	793.83	chuánchéng 'inherit'	1.34
4	tǐyàn 'experience'	1619	tǐyàn 'experience'	−0.85	chénglì 'establish'	715.83	chénglì 'establish'	1.28
5	fāzhǎn 'develop'	1544	chéngxiàn 'appear'	−0.51	guīhuà 'plan'	495.12	bǎocún 'preserve'	1.26
6	chuánchéng 'inherit'	1423	chuánchéng 'inherit'	−0.27	tuīguǎng 'promote'	437.88	tǐyàn 'experience'	1.21
7	tuīguǎng 'promote'	1173	zhǎnxiàn 'show'	−0.16	tǐyàn 'experience'	418.4	jǔbàn 'hold'	1.18
8	tuīdòng 'execute'	968	tuīguǎng 'promote'	−0.13	bǎocún 'preserve'	332.87	zhǎnshì 'display'	1.12
9	guīhuà 'plan'	922	chénglì establish'	−0.05	zhēngqǔ 'strive for'	161.6	chéngxiàn 'appear'	0.99
10	bǎocún 'preserve'	903	bǎocún 'preserve'	−0.05	tuīdòng 'execute'	146.33	tuīdòng 'execute'	0.95
11	zhīchí 'support'	723	zhēngqǔ 'strive for'	0	zhǎnshì 'display'	137.85	dǎzào 'create'	0.91
12	zhēngqǔ 'strive for'	684	dǎzào 'create'	0.09	fāzhǎn 'develop'	124.78	guīhuà 'plan'	0.76
13	zhǎnshì 'display'	582	tuīdòng 'execute'	0.54	zhīchí 'support'	119.62	fāzhǎn 'develop'	0.71
14	dǎzào 'create'	570	jǔbàn 'hold'	0.54	chéngxiàn 'appear'	114.11	zhēngqǔ 'strive for'	0.7
15	chéngxiàn 'appear'	554	guīhuà 'plan'	0.55	dǎzào 'create'	94.76	zhīchí 'support'	0.43

parks, or Hakka earthen buildings; *zhīchí* 'support' is all related to presidential or political candidates; *tuīguǎng* 'promote' is related to Hakka culture, followed by Hakka language and music. The following three examples give the context in which the collocations occur in the news articles, whereby the underlined noun phrases specifically profile which aspect of *kèjiā* is being predicated by the verb.

Table 5. The object theme of the verb collocates

Ranking	1	2	3
guīhuà	wénhuà yuánqū 'culture park'	kèjiā yuánlóu 'Hakka earthen buildings'	kèjiā tǔlóu 'Hakka earthen buildings'
tuīguǎng	kèjiā wénhuà 'Hakka culture'	kèjiā huà 'Hakka language'	kèjiā gēyáo 'Hakka music'
tuīdòng	kèjiā shìwù 'Hakka affairs'	kèjiā wénhuà 'Hakka culture'	kèjiā yǔ 'Hakka language'
zhǎnxiàn	kèjiā jīngshén 'Hakka spirit'	kèjiā fēngqíng 'Hakka custom'	kèjiā tèsè 'Hakka character'
zhǎnshì	kèjiā wénwù 'Hakka cultural relic'	kèjiā bǎnzǎi 'Hakka rice noodles'	kèjiā mǐshí 'Hakka food'
bǎocún	kèjiā wénhuà 'Hakka culture'	kèjiā jùluò 'Hakka village'	kèjiā wénwù 'Hakka cultural relic'
zhēngqǔ	kèjiā piào 'vote'	jīngfèi 'founding'	bǔzhù 'subsidy'
chéngxiàn	kèjiā fēngqíng 'Hakka custom'	jīngshén 'spirit'	tèsè 'character'
chénglì	kèjiā hòuyuánhuì 'Hakka supporting fan'	kèjiā wěiyuánhuì 'Hakka Affairs Council'	kèjiā diàntái 'Hakka radio'
dǎzào	kèjiā táohuāyuán 'Hakka Utopia'	Táiwān 'Taiwan'	kèjiā yuánlóu 'Hakka earthen buildings'
zhīchí	Cài yīng-wén 'Tsai Ing-wen'	Mǎ yīng-jiǔ 'Ma Ying-jeou'	Zhū lì-lún 'Chu, Li-luan'
fāzhǎn	shèqū 'community'	kèjiā chǎnyè 'Hakka industry'	kèjiā zúqún 'Hakka group'
chuánchéng	kèjiā jīngshén 'Hakka spirit'	kèjiā xìqǔ 'Hakka music'	kèjiā mǔyǔ 'Hakka language'
jǔbàn	kèjiā gē bǐsài 'Hakka music contest'	yīnyuèhuì 'concert'	jiānniánhuá 'carnaval'
tǐyàn	kèjiā fēngqíng 'Hakka custom'	kèjiā léichá 'Hakka tea'	kèjiā wénhuà 'Hakka culture'

(6) 他呼籲選民**支持**客家子弟   (LTN 2011/07/09)
*tā hūyù xuǎnmín **zhīchí** kèjiā zǐdì*
'He urged voters to support Hakka candidates.'

(7) 他在高雄市長任內**成立**第一個客家電台   (LTN 2007/09/26)
*tā zài gāoxióng shìzhǎng rènnèi **chénglì** dì-yí-ge kèjiā diàntái*
'He established the first Hakka radio station when he was the Kaohsiung mayor.'

(8)   充分**展現**客家人的人文精神   (KNW 2009/05/14)
      *chōngfēn **zhǎnxiàn kèjiārén de rénwén jīngshén***
      '(This activity) showed the spirit of humanism of the Hakka people.'

Finally, Figure 1 gives the distribution of the fifteen verb collocates found in various news topics. The content shows strong correlation between the verb collocate of *kèjiā*, its referential domain, and the news topic projected discursively. For the fifteen predicates, the most frequently reported topic in the news is artistic activities, followed by lifestyle topics and politics.
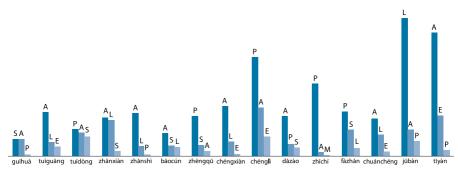

Figure 1. The distribution of verb collocates in different news topics
(*Note.* S = society; A = art; P = politics; E = education and technology; L = lifestyles; M = media.)

5.   Discussion and implications

This research has integrated computational tools and methods of corpus calculation for the elaboration of information on an ethnic term used in news discourse. The findings profile the importance of using more rigorous computational and statistic methods than hitherto to measure collocational strength in news texts. From the linguistic perspective, the identification of collocational strength resolves semantic indeterminacy in a more precise way. The denotative vagueness of *kèjiā* is enhanced by elevating semantic specificity through the semantic contents of collocates. The noun collocates help specify clearer the referential domain, hence showing what Hakka matters are referred to by the news. The verb collocates together with their object themes with *kèjiā* help reveal the most commonly-reported news topics and how they are being characterized. From the discursive perspective, the syntagmatic clues in the proximal co-text within a whole news article are identified, explicitly projecting the thematic information of the context. Rather than measuring collocates only within the immediate co-text,

in which fractional perspectives on the construal of news value may result, our methods that measure the strength of a collocation within a complete news article increase the credibility of the description of the strength of a collocation, in turn unfolding the referential strategies operating in the news discourse in a more clear way than hitherto. The methods used provide a promising way for an intra-textual linguistic analysis of news discourse. Furthermore, our findings reveal the predicational strategies involved in news reporting. A tendency of the most prevalently discussed and concerned topics regarding Hakka affairs in Taiwan society is discursively detected. A strong correlation is observed between *kèjiā*, its verb collocates and news topics. A further analysis of the news topics across the articles included in the study is given in Figure 2.
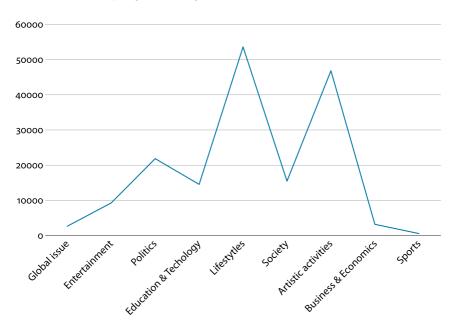

Figure 2. The distribution of *kèjiā* in different topics

Among the topics, the most frequently reported Hakka topics in a descending order are: lifestyle topics (34%) that encompass family matters, leisure activities, or travel; artistic activities (32%) that encompass cultural festivals, art exhibition, creation of artistic or craft objects, or performances; and politics (11%). The fact that *xiāngqīn* 'fellow folks' *wénhuà* 'culture', *chuántŏng* 'tradition' *měishí* 'cuisine', and *tónghuā* 'tong blossoms' score relatively higher than other cases in terms of the four calculations correlates with the topics most reported in news.

Both the linguistic and the discursive findings underscore certain fundamental or commonsensical ideological beliefs or ideas that are otherwise implic-

itly characterized under the framing of the news discourse (Verschueren 1999). From a social-cultural perspective, the operation of such discursive strategies in news discourse carries certain implications. For one, the results show that the news media appeals to the popular assumptions and premises that Hakka arts and crafts, traditional Hakka activities, and Hakka cuisines have become more acceptable and accessible to the public than before, and hence are worthy of being reported. For the other, the results also show that the scenes of bustle and excitement associated with Hakka activities are the most common Hakka images projected in the media, disseminating political and societal propaganda for the purpose of raising the visibility of the Hakka ethnic groups. The implications are closely related to ideological effects created by newsworthiness that are not only ingrained in language but also influenced by the social-cultural context. Language use is "not simply a socially or politically neutral resource" (Chilton and Schäffner 2011, 320), but rather embodied in historical and political discourses. Language choice is considered to carry strategic functions that give rise to meanings interpreted based on knowledge of the background and of the values in the situations described in the articles (Chilton and Schäffner 2011, 311). The analysis reveals that in total 168,116 tokens of *kèjiā* are found in 48,726 news articles in the database, showing that almost 93 news articles are related to Hakka matters per month from the four major newspapers in Taiwan in the eleven years of 2005–2015 inclusive. The analysis seems to suggest a significant increase in terms of media exposure in the recent years, considering that the Hakka language is a non-dominant one and that the Hakka ethnic groups are socially vulnerable in Taiwan.

In fact, the Hakka ethnic groups have long been considered as invisible in a multi-lingual and multi-ethnic Taiwan society because of the complicated historical, political, economic and social factors that led to the nature of their diaspora (Hsu 1994; Chiou 2006; Wang 2007; Lai 2017). Such invisible and diasporic nature has also led to the decline of their language and culture since the Hakka language is not a language prevalently used in education or in public domains. According to two reports of demographic and language surveys conducted by the Hakka Affairs Council (HAC 2013, 2016) in Taiwan, the Hakka population comprises 19.3% of the total population, approximating 4.5 million people. Further, almost all of the speakers of Hakka who are fluent in the language are over sixty years old; and only 15% of Hakka people in the age group of thirteen-eighteen can speak Hakka. The use of the Hakka language is therefore suffering a serious decline as a result of the loss of speakers among the younger generations. Dating back, along with the declaration of the ending of martial law in 1987 in Taiwan, the Hakka ethnic groups launched their first social movement in 1988 – the *Give My Mother Tongue Back Movement*. This event was a milestone as it aroused the Hakka people's awareness of their social and political inferiority and the crisis of the possibility of the loss of

their culture, language, and tradition. Attempts by Hakka people were made to get involved with the process of democratization and to participate in the construction of Taiwanese consciousness (cf. Wang 2007, 880). Then, the year 2000 was of great significance since it was the year that the DPP first won the presidential election as opposed to the long politically dominant KMT. The president-elect Shuibian Chen, in order to realize his promises to the Hakka ethnic groups stated in the Hakka White Paper, implemented several Hakka-based policies, including the promotion of Hakka culture, the preservation and development of the Hakka language, and the enhancement of the media exposure of the Hakka ethnic groups. The Council for Hakka Affairs was established in 2001 at the central government level and various administrative units were established in the cities and counties to promote Hakka-related affairs. Hakka cultural parks, local-level public buildings constructed to include Hakka-style village motifs, annual Hakka traditional festivals, activities, and events have appeared in significant numbers ever since. Accordingly, it is these aspects of Hakka matters that are mostly covered in the news. Indeed, in the findings, the topic-associated collocates show that the news topics most concerned with Hakka were in the form of soft straight reports of news in a neutral and factual manner without an expression of opinion or creation of tensions (Peng 2008). In other words, a display of nostalgic attachment to Hakka cultural traditions is foregrounded as being newsworthy in the news related to the Hakka ethnic groups. The public propaganda appeals to the collective memory of Hakkaness – grassrootness, tradition, and nostalgia – by highlighting the ideological bond of the Hakka ethnic groups.

Such an emphasis, however, perpetuates a stereotypical impression as to the issues that are significant for Hakka people. This appeal shows the media's efforts in raising the visibility of the Hakka ethnic groups in recognizing their deep anxiety in preserving their cultural inheritance and assets (Lai 2017). Yet linguistic analysis of the verb collocates evidences that the Hakka ethnic groups have been placed in the peripheral as opposed to the general social mainstream. Specifically, while the top fifteen verb collocates seem to carry positive connotations, nevertheless, within the news topics, they all refer to Hakka matters that need to be planned, promoted, or enhanced – from local public constructions in the traditional Hakka style and Hakka cultural parks to Hakka traditional cuisines, and Hakka languages. Thus, a strong implicit social-cultural construal of the discursive discourse is made known. That is, news related to Hakka people has to do not with the mainstream issues in society but mostly with their needs for assistance or support from the government or the society. Their lower, secondary status is implied by reference to a lack of modernization, lack of pro-activeness, and hence lack of hard news that deserves serious wider dissemination on the front page. Cottle (2000, 11) points out "how ethnic minorities are now often portrayed in

deliberate 'multiculturalist' ways through a (superficial) focus on cultural festivals, individual success stories and the cultural exotica of ethnic minority cultures." And "…such 'multiculturalist' representations…, may actually serve to reinforce culturally sedimented views of ethnic minorities as 'Other' and simultaneously appear to give the lie to ideas of structural disadvantage and continuing inequality," as claimed by him.

Furthermore, another interesting social-cultural implication has to do with the category of politics, as shown by the two collocates – *zhīchí* 'support', and *yìngjǐng* 'stiff neck'. The fact that the verb collocate *zhīchí* 'support' occurs naturally related to political candidates indicates that Hakka people are reported as a slogan and are politically manipulated as supporting certain political parties. Hakka people have been courted by both the KMT and the DPP, the two major political parties in Taiwan, as holding swing votes that have the potential to influence the results of elections. The co-occurrence of *zhīchí* 'support' and certain politicians hence reveals such an underpinning social-cultural basis in Taiwan society.

This observation correlates with the highest PMI value of *yìngjǐng* 'stiff neck' in our analysis. Lai (2017) investigates the semiotic innovation of this particularly unique Hakka symbolic code *yìngjǐng* 'stiff neck' in news media. The longitudinal analysis of this symbolic code in news media shows that it has come to be used to represent a simplistic image with familiar judgement and values that can be attributed to the discourse situation reported in the news. Portraying an embodied experience – people making their necks stiff to show toughness – this code projects an association with an ideological stereotype of Hakka ethnic groups. As pointed out by Lai (2017), the strong semantic connection of *kèjiā* 'Hakka' and *yìngjǐng* 'stiff neck' reveals how Hakka people are considered tough and indomitable as metaphorically encoded by the use of this symbolic code, which projects a prototypical conception of the Hakka ethnic groups. The salient and frequent usage of *yìngjǐng* 'stiff neck' in media discourse has seemed to increase the visibility of the Hakka ethnic groups. At the same time, the word, emerging as a handy attractor in news discourse, also results in an ideological impact upon the understanding of *kèjiā* 'Hakka'. A news report about the current president Ing-wen Tsai using this handy symbolic code as shown in (9) clearly reveals such an ideological manifestation. In a constituency wherein the Hakka population is the majority, President Tsai uses this symbolic code to solicit the emotional empathy of Hakka people and support for the purpose of winning an election.

(9) 拿出「**硬頸**」精神！蔡英文：越艱困選區 越認真經營   (LTN 2016-01-08)
*náchū **yìngjǐng** jīngshén càiyīngwén yuè jiānkùn xuǎnqū yuè rènzhēn jīngyíng*
pluck up-yingjing-spirit Tsai Ingwen-more-difficult-constituency-more-serious-run
'Let's pluck up the indomitable spirit! Ing-wen Tsai: The more difficult the constituency is, the more seriously we need to run."

As *yìngjǐng* 'stiff neck' carries a strong collocational strength with *kèjiā* 'Hakka', it seems that the Hakka ethnic groups are given a free ride in news exposure whenever *yìngjǐng* is used. Intriguingly, its meaning extends by way of metaphor and metonymy to represent various kinds of objects or concepts. Consider the following examples, taken from Lai (2017, 418, Example (2) and Example (4)). In Example (10), even Angelina Jolie is portrayed as having the quality of *yìngjǐng* in being tough on refugee issues as referred to in a news headline. In Example (11), *yìngjǐng* is metonymically used to refer to the capital city Tokyo, which stands for the Japanese government, and, in turn, stands for the decision makers in the Japanese government.

(10) 裴莉**硬頸**挺難民   (LTN 2015-04-26)
*Qiúlì **yìngjǐng** tǐng nànmín*
Angelina Jolie-stiff neck-back up-refugees
'Jolie strongly backs refugees up.'

(11) 東京還是如此**硬頸**   (Udn 1998-11-29)
*Dōngjīng háishì rúcǐ **yìngjǐng***
Tokyo-still-is-so-stiff neck
'Tokyo is still so tough and persistent!'

Language is the most salient social marker that can shape the identity of ethnic groups. The current study accords with Lai (2017) in showing how over a long period of continuous time, media has tried to sustain a specific ideology by way of the innovative usage of this symbolic code in an effort to recruit news readers to such an ideology.

6.   Concluding remarks

This study has demonstrated more rigorous and precise methods for the computation and calculation of collocation analysis in news corpora than hitherto, aiming to suggest that the findings can provide useful insights into better construal of news discourse. The results were teased out in terms of linguistic, discursive and social-cultural layers for the interpretation of the semantic vagueness of an ethnic

term. Hsiao and Huang (2008) evaluated the implementation and impact of the Hakka Movement in 1988 after 20 years and maintain that its most significant and noticeable effect is the presence of media images. This present study, covering the range of news discourse almost a decade afterwards, provides an in-depth analysis for a more comprehensive interpretation of news discourse toward the Hakka ethnic groups in the recent years.

Nevertheless, while computational tools may be employed for further investigation of other corpora, there are limitations to this study. Since the corpora are extracted from the news in Taiwan, the findings are restricted to the situation in Taiwan. The results and interpretations from this one specific region may not be able to be generalized to other areas in which Hakka ethnic groups can also be found – areas such as certain provinces in Mainland China, Malaysia, or Singapore. Note also that, even though the Hakka ethnic groups can be found in various areas in Mainland China, in some areas in Southeast Asia, and in different countries all over the world, it can be understood that they must be presented with diversified characteristics regarding their linguistic and cultural elements as shaped by time and space, and, most importantly, by the situated political and social-cultural context. Investigation into these Hakka ethnic groups for the comparison and contrast of the various diversified linguistic and cultural paradigms are possible directions for further research.

Moreover, the perspectives raised by this study may provide a complementary reference to the examination of the construction of newsworthiness in the extant literature (e.g., Bednarek and Caple 2014). Note that different categorizations, labels or groupings can be seen in other studies, and that the linguistic devices are not claimed to be exhaustive. It was found that the Hakka people's nostalgic attachment to their cultural traditions is foregrounded as being newsworthy. Such an emphasis can be associated with the news value of consonance – the perpetuation of stereotypical aspects of a culture and the adherence to the expectations of the readers. The projection and labeling of the Hakka ethnic groups as being peripheral and of a secondary lower status bring about a different ingratiating news topic worthy of reporting. The collocational analysis also suggests that the value of negativity can be manifested by positive expressions when certain situated contexts which unpack some of the assumptions underlying language use are added into the news. It implies that the understanding of negativity is a subjective value that should include the examination not only of linguistic devices but also of social-cultural factors. Furthermore, the fact that the Hakka ethnic groups are often politically manipulated owing to their role in changes of allegiance to swing a vote indicates the existence of a paradoxical ideology – the give-and-take conditions of getting votes for an exchange of resources in Taiwan society.

In sum, a deeper understanding of the representation of the views of a minority ethnicity in the news media is achieved when characterized through three layers of meaning – the explicit referential meaning in the linguistic context, the textual manifestation in the discursive context, and the implicit construal in the social-cultural context. The findings of this study hence shed new light on the construal of news discourse, and at the same time suggest the possibility of a more comprehensive examination of the news values indicators in the extant literature.

Funding

Acknowledgements

References

Baker, Paul, Costas Gabrielatos, and Tony McEnery. 2013. "Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word 'Muslim' in the British Press 1998–2009." *Applied Linguistics* 34 (3): 255–278. https://doi.org/10.1093/applin/ams048

Bartsch, Sabine, and Stefan Evert. 2014. "Towards a Firthian Notion of Collocation." *OPAL* 48–69.

Bednarek, Monika, and Helen Caple. 2014. "Why Do News Values Matter? Towards a New Methodological Framework for Analyzing News Discourse in Critical Discourse Analysis and Beyond." *Discourse and Society* 25 (2): 135–158. https://doi.org/10.1177/0957926513516041

Bhopal, Raj. 2004. "Glossary of Terms Relating to Ethnicity and Race: For Reflection and Debate." *Journal of Epidemiology and Community Health* 58 (6): 441–445. https://doi.org/10.1136/jech.2003.013466

Blum-Kulka, Shoshana, and Michal Hamo. 2011. "Discourse Pragmatics." In *Discourse Studies: A Multidisciplinary Introduction*, ed. by Teun A. van Dijk, 143–164. London: Sage. https://doi.org/10.4135/9781446289068.n8

Breeze, Ruth. 2016. "Negotiating Alignment in Newspaper Editorials: The Role of Concur-Counter Patterns." *Pragmatics* 26 (1): 1–19. https://doi.org/10.1075/prag.26.1.01bre

Brezina, Vaclav, Tony McEnery, and Stephen Wattam. 2015. "Collocations in Context: A New Perspective on Collocation Networks." *International Journal of Corpus Linguistics* 20 (2): 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

Chappell, Hilary and Christine Lamarre. 2005. *A Grammar and Lexicon of Hakka*. Ecole des Hautes Etudes en Sciences Sociales.

Chilton, Paul, and Christina Schäffner. 2011. "Discourse and Politics." In *Discourse Studies: A Multidisciplinary Introduction*, ed. by Teun A. van Dijk, 303–330. London: Sage. https://doi.org/10.4135/9781446289068.n15

Chiou, Chang-Tay. 2006. "A Study of the Hakka's Invisibility in Taiwan: An Analysis of Their Language Use and Ethnic Identification." *Hakka Studies* 1 (1): 45–96.

Cottle, Simon. 2000. "Introduction Media Research and Ethnic Minorities: Mapping the Field." In *Ethnic Minorities and the Media*, ed. by Simon Cottle, 1–30. Buckingham: Open University Press.

Cruse, David Alan. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Firth, John Rupert. 1957. "A Synopsis of Linguistic Theory 1930–1955." In *Selected Papers of J.R. Firth 1952–1959*, ed. by Frank R. Palmer, 168–205. London: Longman.

Gries, Stefan Th. 2013. "50-Something Years of Work on Collocations." *International Journal of Corpus Linguistics* 18 (1): 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Hakka Council Affairs. 2013. *Report of National Surveys of the Usage of Hakka Languages Between 2012–2013*.

Hakka Council Affairs. 2016. *Report of National Surveys of Hakka Population and Languages Between 2015–2016*.

Hsu, Cheng-Kuang. 1994. "The Ethnic Relations in Taiwan: An Investigation from Hakka Perspective." *Proceedings of the Conference of Hakka and Culture*, 381–399.

Hsiao, Hsin-Huang Michael, and Shih-Ming Huang. 2008. "Hakka Movements Under the Shift of Politics in Taiwan." In *Hakka and Multi-Ethnic Groups: The 20th Anniversary of Hakka Movement in Taiwan*, ed. by Wei-An Chang, Cheng-Kuang Hsu and Lieh-Shih Lo, 157–182. Taipei: Nantian.

Huang, Chu-Ren, Jia-Fei Hong, Wei-Yun Ma, and Petr Šimon. 2015. "From Corpus to Grammar: Automatic Extraction of Grammatical Relations from Annotated Corpus." *Linguistic Corpus and Corpus Linguistics in the Chinese Context. Journal of Chinese Linguistics Monograph Series* 25: 192–221.

Jiwani, Yasmin, and John Richardson. 2011. "Discourse, Ethnicity and Racism." In *Discourse Studies: A Multidisciplinary Introduction*, ed. by Teun A. van Dijk, 241–262. London: Sage. https://doi.org/10.4135/9781446289068.n12

Lai, Huei-ling. 2017. "Understanding Ethnic Visibility Through Language Use: The Case of Taiwan Hakka." *Asian Ethnicity* 38 (3): 406–423. https://doi.org/10.1080/14631369.2016.1171130

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

McEnery, Tony M., and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Pecina, Pavel. 2010. "Lexical Association Measures and Collocation Extraction." *Language Resources and Evaluation* 44 (1): 37–158.

Peng, Wen-Cheng. 2008. "Variety of Hakka Images in Taiwan Major Newspapers." In *Hakka and Multi-Ethnic Groups: The 20th Anniversary of Hakka Movement in Taiwan*, ed. by Wei-An Chang, Cheng-Kuang Hsu and Lieh-Shih Lo, 274–296. Taipei: Nantian.

Potts, Amanda, Monika Bednarek, and Helen Caple. 2015. "How Can Computer-Based Methods Help Researchers to Investigate News Values in Large Datasets? A Corpus Linguistic Sstudy of the Construction of Newsworthiness in the Reporting on Hurricane Katrina." *Discourse and Communication* 9 (2):149–172. https://doi.org/10.1177/1750481314568548

Seretan, Violeta. 2011. *Syntax-Based Collocation Extraction*. Berlin: Springer. https://doi.org/10.1007/978-94-007-0134-2

Sinclair, John. 2004. "The Search for Units of Meaning." In *Trust the Text. Language, Corpus and Discourse*, ed. by John Sinclair, 190–191. London: Routledge.

Sinclair, John, Susan Jones, and Robert Daley. 2004. *English Collocation Studies: The OSTI Report*. London: Continuum.

Van Dijk, Teun A. 1992. "Discourse and the Denial of Racism." *Discourse and Society* 3 (1): 87–118. https://doi.org/10.1177/0957926592003001005

Verschueren, Jef. 1999. *Understanding Pragmatics*. London: Oxford University Press.

Wang, Lijung. 2007. "Diaspora, Identity and Cultural Citizenship: The Hakkas in Multicultural Taiwan." *Ethnic and Racial Studies* 30 (5): 875–895. https://doi.org/10.1080/01419870701491861

Yang, Heng-Li, and August FY Chao. 2015. "Sentiment Analysis for Chinese Reviews of Movies in Multi-Genre Based on Morpheme-Based Features and Collocations." *Information Systems Frontiers* 17 (6): 1335–1352. https://doi.org/10.1007/s10796-014-9498-1

Publication history