A Three-layered Collocation Extraction Tool and its Application in China English Studies

[1]Jingxiang Cao, [2]Dan Li and [3]Degen Huang

[1]School of Foreign Languages, [2]School of Foreign Languages, [3]School of Computer Science and Technology,
Dalian University of Technology, Dalian 116024, Liaoning, China
caojx@dlut.edu.cn, linda_2013@mail.dlut.edu.cn,
huangdg@dlut.edu.cn

Abstract. We design a three-layered collocation extraction tool by integrating syntactic and semantic knowledge and apply it in China English studies. The tool first extracts peripheral collocations in the frequency layer from dependency triples, then extracts semi-peripheral collocations in the syntactic layer by association measures, and last extracts core collocations in the semantic layer with a similar word thesaurus. The syntactic constraints filter out much noise from surface co-occurrences, and the semantic constraints are effective in identifying the very "core" collocations. The tool is applied to automatically extract collocations from a large corpus of China English we compile to explore how China English as a variety of English is nativilized. Then we analyze similarity and difference of the typical China English collocations of a group of verbs. The tool and results can be applied in the compilation of language resources for Chinese-English translation and corpus-based China studies.

Keywords: collocation extraction; dependency relation; China English

1 Introduction

Collocation is pervasive in all languages. *Collins COBUILD English Collocations* includes about 140,000 collocations of 10,000 headwords of English core vocabulary. Collocation is of great importance in Natural Language Processing (NLP) as well as in Linguistics and Applied Linguistics.

Various methods of automatic collocation identification and extraction have been proposed. The common procedure mainly consists of two phases: extracting collocation candidates and assigning association score for ranking [1]. Collocation candidates can be extracted based on surface co-occurrence, textual co-occurrence and syntactic co-occurrence [2], among which the syntactic co-occurrence contains the most linguistic information and is suitable for collocation analysis in the perspective of linguistic properties. The association score can be calculated through different association measures (AMs). Frequency method simply takes the collocation as a whole whereas mean and

variance method [3], hypothesis test (including z-test, t-test, chi-square test, log-likelihood ratio) and information theory (MI^k) [2][4] also consider the components, thus getting better performance; other methods using non-compositionality [5] and paradigmatic modifiability [6] further consider the substitutes of the collocation components, which works well for non-compositional phrases or domain-specific n-gram terms. Smadja's X-tract [3] starts from surface co-occurrence, extracts bigrams, n-grams with window-based method and extends them into syntactic co-occurrence with syntactic parser. Reference [7] constructs a tool for NOUN+VERB collocation extraction as well as morpho-syntactic preference detection (active or passive voice).

Those methods and tools are mainly designed and applied in NLP tasks like semantic disambiguation, text generation or machine translation, rarely oriented towards linguists other than computational scientists. But modern linguists have always been in need of appropriate tools. WordSmith [8] may be the popular corpus assistant software most used by linguists with three modules: Concord, Keywords and WordList, among which the Concord can compute collocates of a given word through window-based method, far from enough for collocation studies.

Inspired by the various extraction methods and linguistic properties of collocation, we design a hierarchical collocation extraction tool based on the three-layered linguistic properties of collocation [9]. It considers different linguistic properties of collocation, which agrees more with the human intuitive conceptualization of collocation.

We also apply our collocation extraction tool in the China English studies. China English is a performance variety of English, which observes the norm of standard Englishes (e.g. British English, American English) but is inevitably featured by Chinese phonology, lexis, syntax and pragmatics [10]. Previous studies on China English have ranged from macro aspects, such as the attitudes towards China English [10, 11], the history of English in China [12, 13], the use of English in China [14] and the pedagogic models of English in China, to micro aspects which focus on specific linguistic levels including phonology, morphology, lexis, syntax, discourses, stylistics etc. [15, 16, 17, 18]. Among those linguistic features, lexical innovation, which is argued to be more likely to get social acceptance compared with grammatical deviations [19], is usually the most active during the nativization of English. Collocations are "social institutions" or "conventional labels", which means the entailed concept is culturally recognized within a specific society. Therefore it is innately appropriate to study the nativization of English which focuses on the process to create a localized linguistic and cultural identity of a variety [20].

Due to the limit of applicable tools, lexical studies on China English are limited, either in the small manually-collected data, or in the rough analysis methods such as frequency, proportion comparison and examples relying on researchers' acute observation or introspection. In-depth empirical studies based on large corpus or latest methods from NLP are therefore needed. Moreover, the lack of effective methods to extract long-distance patterns forces most linguists to study consecutive collocations like noun phrase [15] or adjective phrase [17]. Verb phrase as a significant research object in language is downplayed.

In this paper, we build a large corpus of China English by crawling the last-five-year webpages of four mainstream newspapers in mainland China, and automatically extract

all the collocations in the corpus. Then we collect 52 high-keyness verbs with the help of WordSmith Tools 5.0 and analyze similarity and difference of the typical China English collocations of a group of verbs.

2 The three-layered collocation extraction tool

2.1 Three-layered collocation definition

Collocation is often regarded as the bridge between free word combination and idiom [21, 22, 23, 24]. It has broad definition as "a pair of words that appear together more often than expected" [25, 26], and narrow one as "recurrent co-occurrence of at least two lexical items in a direct syntactic relation" [1] [6], or further restricted one as "recurrent co-occurrence with both syntactic and semantic constraints" [5]. The definitions are gradually narrowed from frequency layer, syntactic layer down to semantic layer.

Based on the three layers, Collocates of a Base [23] are classified into core collocates, semi-peripheral collocates and peripheral collocates. Given a base, a word is a core collocate iff it satisfies all the constraints A, B and C, a semi-peripheral collocate iff it satisfies constraints A and B, and a peripheral collocate iff it only satisfies constraint A.

Three defining constraints are

A) Frequency constraint: the frequency over a specific threshold
B) Syntactic constraint: direct syntactic relation
C) Semantic constraint: not substitutable without affecting the meaning of the word sequence

2.2 Collocation extraction architecture

The first step is to extract peripheral collocation. The texts are segmented into sentences with a punctuation package adapted from Kiss and Struct [27] in NLTK [28], and parsed with Stanford Parser [29] to extract syntactically related co-occurrences with no limit on their distances. Then the dependency triples are extracted from parsed texts and lemmatized with WordNet lemmatizer [30] in NLTK [28] in order to reduce data sparsity. We discard triples with "root" relations or stop word components and selected those with no less than 3 occurrences as peripheral collocations, also candidates of semi-peripheral collocation.

The second step employs an integrated association measure (AM) to extract semi-peripheral collocations. The three AMs are designed for different purposes: LLR (log-likelihood ratio) [4] answers "how unlikely is the null hypothesis that the words are independent?" [2], MI^k (revised MI of Lin [6]) answers "how much does observed co-occurrence frequency exceed expected frequency?" [2], and PMS [5] measures the substitutability of the components in a dependency triple.

For any word pair (u, v) adapted from dependency triple (u, rel, v), we have the contingency table as follows:

Table 1. Contingency table of word pair (u, v)

	v	$\bar{v}$
u	a	b
$\bar{u}$	c	d

$\bar{v}$ means the absence of v. a, b, c, d are the counts of word pairs (u, v), $(u, \bar{v})$, $(\bar{u}, v)$, $(\bar{u}, \bar{v})$. Obviously, $a + b + c + d$ is the sample size N. LLR is represented as follows:

$$\begin{aligned} \text{LLR} = 2(&a\,log\,a + b\,log\,b + c\,log\,c + d\,log\,d - (a+b)\,log(a+b) \\ &- (a+c)\,log(a+c) - (b+d)\,log(b+d) - (c+d)\,log(c+d) \\ &+ (a+b+c \\ &+ d)\,log(a+b+c+d)) \end{aligned} \tag{1}$$

The three-variable $\text{MI}^k(u, rel, v)$ here is under the assumption that u and v are conditionally independent given dependency relation rel. As is known that MI biases to low frequency word, we add k-th power to the numerator in order to eliminate the effect.

$$\begin{aligned} \text{MI}^k(u, rel, v) &= log\left(\frac{p(u, rel, v)^k}{p(u|rel)p(rel)p(v|rel)}\right) \\ &= log\left(\frac{(|u, rel, v| - b)^k |rel|}{|u, rel||rel, v|N^{(k-1)}}\right) \end{aligned} \tag{2}$$

u and v are the component words in a dependency triple, rel is the dependency type, p(#) is the frequency of #, |#| is the count of #, b(=0.95 in our experiments) is an adjustment parameter, and N is the sample size.

$$PMS(u, rel, v) = \frac{|u, rel, v|^6}{|u||rel||v||u, rel||rel, v||u, v|} \tag{3}$$

In order to take advantage of the three AMs, we normalize their values in interval [0,1] and integrate them using geometric mean. The integrated measure (LMP^k) is defined as follows:

$$\text{LMP}^k(u, rel, v) = \sqrt[3]{\text{LLR}'(u, v) * \text{MI}^{k\prime}(u, rel, v) * \text{PMS}'(u, rel, v)} \tag{4}$$

$'$ means the normalized AM.

The triples with LMP^k higher than a specified threshold are regarded as semi-peripheral collocations, and the rest of the candidates are peripheral collocations.

The third step filters out the semi-peripheral collocations to reserve the core collocations by assigning semantic constraints, i.e. to compute the probability of substituting the component words without affecting the meaning of the original collocation.

We adopt Lin [31] to measure the probability. First, we compile a thesaurus by taking all the collocations of a word as its features, computing the similarity between any two words, and selecting the top 10 most similar words for each entry. Based on the thesaurus we reserve the collocation whose MI^k is significantly different from its substitutive collocations at the 5% level.

Given a word w_1, we calculate $\mathrm{Simi}(w_1, w_2)$ to rank its similar words.

$$\mathrm{Simi}(w_1, w_2) = \frac{2\mathrm{Info}(\mathrm{F}(w_1) \cap \mathrm{F}(w_2))}{\mathrm{Info}(\mathrm{F}(w_1) + \mathrm{Info}(\mathrm{F}(w_2)} \quad (5)$$

$$\mathrm{Info}(\mathrm{F}(w)) = -\sum_{f \in F} \frac{p(f)}{p(\mathrm{POS}(w))} \quad (6)$$

$\mathrm{F}(w)$ is the feature set of w, $\mathrm{Info}(F)$ is the amount of information of feature set F, $\mathrm{POS}(w)$ is the POS of w, $p()$ is the frequency. For example, for the base *promote*, we extract (promote, dobj, exchange) and (promote, advmod, actively), and thus (dobj, exchange) and (advmod, actively) belong to the feature set of *promote*, F(*promote*).

Then we employ z-test to extract core collocations. A dependency triple X is not a core collocation if:

a) There is a triple Y obtained by substituting the component with its similar word;

b) $\mathrm{MI}^k(\mathrm{Y}) \in [\, log\left(\left(|u,rel,v| - b - Z_\alpha\sqrt{|u,rel,v|}\right)^k * \frac{|rel|}{|u,rel|*|rel,v|*N^{(k-1)}}\right),$

$$log\left(\left(|u,rel,v| - b + Z_\alpha\sqrt{|u,rel,v|}\right)^k * \frac{|rel|}{|u,rel|*|rel,v|*N^{(k-1)}}\right)\,]$$

(α=5%).

2.3 Comparison with other tools

We compare our tool with the window-based method and WordNet[1] [30] to test the performance of different steps in our tool.

As our collocation candidates are directly from dependency triples with syntactic constraints, we want to see how it differs from the traditional window-based method. Window-based method is a standard method in collocation extraction before mature syntactic parsers came out. It is broadly adopted but lack of interpretability due to mixing "true" and "false" instances as well as distance-different instances identified in the source text [1].

The first experiment is to verify the validity of syntactic co-occurrences in the first step compared with surface co-occurrences. The surface co-occurrences are generated with 5-word window size and the syntactic co-occurrences are generated from the dependency triples. We systemically sampled 100 measure points (by one percent interval) in the respective ranking list of surface co-occurrences and syntactic co-occurrences, extracted semi-peripheral collocations in the second step by LLR, and computed the precisions and recalls which are shown in Table 2.

We find that the syntactic co-occurrences perform much better than the surface co-occurrences. The highest F1 of the surface co-occurrences is 18.77%, and that of the syntactic co-occurrences is 30.35%. However, the surface co-occurrences get higher recall, which indicates that, although the surface co-occurrences bring more potential

[1] http://wordnet.princeton.edu/

candidates, they introduce massive noise. The lower recall of the syntactic co-occurrences is due to that the same surface co-occurrence can derive different syntactic co-occurrences which consist of the dependency relation and the original word pair in the surface co-occurrence, making the data sparser.

Table 2. Comparison of surface and syntactic co-occurrences

	Window-based (%)			Syntax-based (%)		
Percentage	P	R	F	P	R	F
10	13.9843	28.5363	**18.7702**	32.7715	21.5252	25.9837
20	10.7229	38.1304	16.7387	28.7933	29.6433	29.2121
30	08.7080	45.2645	14.6061	25.7732	36.9004	**30.3490**
40	07.5996	53.5055	13.3089	22.3979	41.8204	29.1720
50	06.6740	59.9016	12.0099	20.0832	47.4785	28.2267
60	05.8837	63.8376	10.7743	18.3206	53.1365	27.2469
70	05.2647	67.4047	09.7665	15.9734	56.2116	24.8775
80	04.7825	70.6027	08.9583	14.0320	59.2866	22.6930
90	04.4079	72.2017	08.3086	12.5244	63.2226	20.9071
100	04.1159	**73.5547**	07.7956	11.2586	**66.7897**	19.2690

We also compare our thesaurus with WordNet, to see whether such world knowledge base can help to improve the performance of the tool. We adopt the precision for the evaluation. Our gold standard from *Oxford Collocation Dictionary* adopts a broad concept of collocation and contains many semi-peripheral collocations according to our definition (e.g. *great effort*), but our tool may filter out some semi-peripheral collocations in the gold standard (e.g. *great effort*). The recall decreases and thus is not appropriate for evaluation.

WordNet is a well-organized knowledge base which contains 117, 000 synsets "interlinked by means of conceptual-semantic and lexical relations", while our thesaurus only consists of 31,118 entries, with each attached with 10 similar words. Surprisingly, the result in Fig.1 shows that our thesaurus performs better than WordNet before the top 38%, and becomes worse after 38%. Actually WordNet didn't filter many semi-peripheral collocations out. Instead, it is relatively conservative because many substitutions of the collocation candidate which are composed of the synonym and the original base don't appear in our corpus at all, which means the condition a) in the third step is not satisfied let alone condition b), thus misleading the tool to regard the candidate as core collocation. It indicates that the word distribution difference between the created corpus and WordNet should be considered if we want to utilize the semantic information.

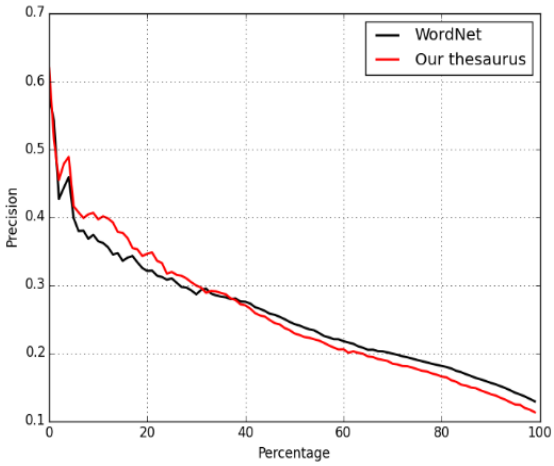

Fig. 1. Comparison of WordNet and our thesaurus

We list some collocations of the following 6 bases (3 (POS type)*2 (keyness type)) in the gold standard set: *effort*, *promote*, *mutual*, *deal*, *pursue*, and *gorgeous*. We set the threshold of four phases (or methods) as 8%, 2%, 42% and 64%, where the F value of the respective collocation ranking list gets the highest value.

Table 3. Extracted collocation examples in different phases

base	Window-based	Peripheral	Semi-peripheral	Core
effort	make effort	make effort	make effort	make effort
	spare effort	spare effort	spare effort	spare effort
	put effort	put effort	put effort	
	strenuous effort	extra effort		
	tireless effort			
promote	promote harmony	promote harmony	promote harmony	promote harmony
	promote cooperation	promote cooperation	promote cooperation	
	promote understanding	promote benefit		
mutual	mutual benefit	mutual benefit	mutual benefit	mutual benefit
	mutual cooperation	mutual cooperation	mutual cooperation	
	mutual suspicion		mutual dependence	
deal	sign deal	sigh deal	sign deal	sign deal
	lucrative deal	good deal	announce deal	
	under-the-table deal			
pursue	pursue dream	pursue dream	pursue dream	pursue dream
	pursue innovation	pursue goal	pursue goal	
		pursue education		
gor-geous	null	null	null	null

For example, as shown in Table 2, the window-based method can extract most collocations (e.g. *make effort, promote harmony, mutual benefit*) that our tool extract except some collocations (e.g. *mutual suspicion, under-the-table deal*).The collocations in our tool are narrowing down from the peripheral to the core. For example, the base *effort* has collocates *make*, *spare*, *put*, *extra* in Peripheral, has collocates *make*, *spare*, *put* in Semi-peripheral, and only has collocates *make* and *spare* in Core. The collocates of *gorgeous* are not extracted because of the absence of its collocates in our test corpus, and null is filled in that raw.

3 Application

3.1 Similarity

We employ Dice Coefficient to evaluate the similarity of two words. Taking each collocate of a word as one of its features, the more common features between two words, the more similar they are.

$$\mathrm{Dice}(v1, v2) = \frac{2|coll(v1) \cap coll(v1)|}{|coll(v1)| + |coll(v2)|} \tag{7}$$

v is the head word, $coll$ is the set of collocates of v.

3.2 Corpus

We build a Corpus of China English (CCE). The corpus size is 126MB, 24 million words and 0.9 million sentences. The texts are crawled by Scrapy[2], a popular crawling framework in Python community, from the official webpages of China Daily[3], Xinhua News[4], the State Council of the People's Republic of China[5], and the Ministry of Foreign Affairs of the People's Republic of China[6]. China Daily and Xinhua News are mainstream comprehensive media that have international influence and publication. The rest two are mainly about politics, economics and diplomacy.

3.3 Test set

Based on the keyword list made from the wordlists of CCE and British National Corpus (BNC) with WordSmith Tool 5.0 (the wordlist of BNC is cited from Scott [8]), we collected 52 verbs from the top 1,000 highest-keyness words. For each verb we extracted 100 collocations (if there exit so many) with our extraction tool, with a total of 5125 collocations. A high-keyness word is defined as one that occurs at least 3 times in

[2] http://scrapy.org
[3] http://www.chinadaily.cn
[4] http://www.news.cn/english/
[5] http://english.gov.cn/
[6] http://www.fmprc.gov.cn/mfa_eng/

CCE and its relative frequency in CCE is statistically significantly larger than in BNC (p-value is 0.05), meaning it is strongly preferred by the editors of the four newspapers.

3.4 Collocations of similar verb in China English

Now that most verbs in our list are positive or neutral, we also wonder, for example in the positive group, whether and to what extent the verbs are similar to each other. We calculated Dice Coefficient of the verbs. As shown in Fig. 2, the red points represent verbs, the orange edges represent similarity between two verbs. The thicker the line is, the more similar the two verbs are to each other.

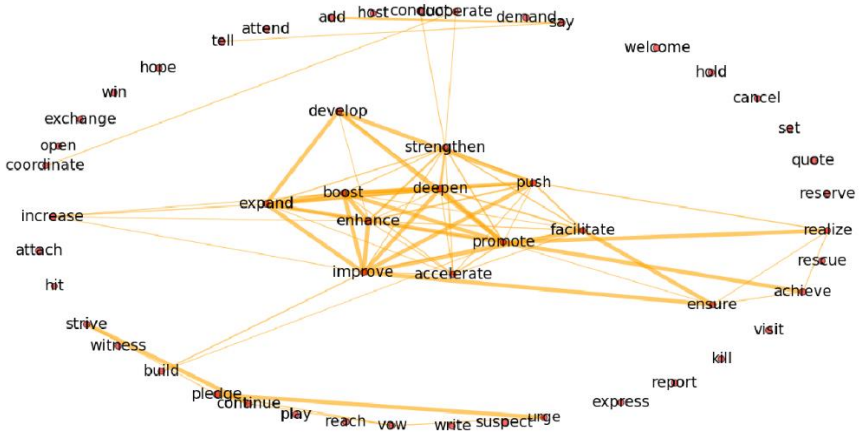

Fig. 2. Verb net based on collocation similarity

We can see clearly that verbs such as *promote, strengthen, enhance, deepen, improve, expand, boost, push, accelerate, facilitate,* and *develop* are strongly connected with several other verbs, usually expressing a positive meaning. We made pairwise comparison of the 11 verbs, and their different collocates are given in Table 4. All the collocations are obviously loan translation rendered from Chinese conventional expressions.

Table 4. Examples of extracted collocaitons of the 11 connected verbs

base verb	Noun collocates	ADV collocates
promote	development, peace, prosperity, growth, stability, integration	actively, vigorously, jointly
strengthen	coordination, communication, supervision, dialogue, trust, management	within ~ framework, ~on ~ issue
enhance	trust, coordination, communication, capability, competitiveness	
deepen	trust, relationship	constantly, continuously, third, ~ in area, within ~ framework

improve	livelihood, quality, efficiency, system, mechanism, environment	constantly
expand	scope, scale, business, demand	at pace, rapidly, continuously
boost	confidence, demand, economy, consumption, vitality, sales, employment	significantly
push	price	forward, up, ahead, for unceasing ~ , to brink, for progress, to limit, along track
accelerate	transformation, pace, negotiation, modernization, restructure	to ~ percent
facilitate	clearance, transformation, flow, interflow, travel, implementation	
develop	economy, industry, country, weapon	rapidly, smoothly, soundly

These collocations in China English reflect conventional expressions of Chinese, especially "various forms of officialese and fixed formulations peculiar to the Chinese political tradition" [33]. In Chinese context our ears are uninterruptedly poured with such expressions, "极大促进", "积极扩大", "大力促进", or "坚定不移地推进". Yet when referring to the *Oxford Collocation Dictionary*, we find varied collocates, like (aggressively, likely) *promote*, (aggressively, playfully, carefully, slowly, blindly*) push*, (radically, exponentially) *expand*, (artificially*) boost*.

These VERB+ADV phrase in China English describe a strong feeling of individual intention and these collocation expressions originate in Chinese expressions appearing extensively in television or newspaper. Due to the quite abstract and opaque meanings of so similar collocations, Chinese people inevitably become confused when they encounter the lexicon selection problem even in Chinese, let alone in English. The collocation comparison may provide a pedagogical reference for China English.

4 Conclusion

The hierarchical collocation extraction tool we propose correspond the output of each phase to the structured definitions. The performance is comparable with the state-of-art extraction methods [2] [26]. By emphasizing broadness in the first two steps and accuracy in the last step, it may offer EFL learners and linguists more choices.

In its application experiment, we built a large corpus of Chinese English and extracted long-distance collocations as well as consecutive ones automatically. We explored how China English is nativilized in terms of verb collocation. Verbs are connected in a network to show their similarity in a collocation perspective instead of traditional semantic perspective. The collocation comparison of similar verbs provides a useful pedagogical reference for China English.

Most of the salient verb collocations are loan translation rendered from Chinese conventional officialeses. They are inevitably influenced by Chinese culture, Chinese linguistic features, and political traditions. We see that China English is exporting Chinese culture and a soft power to expand Chinese influence in the world.

Till now the model is monolingual, not multilingual. As collocation tends to be the one that can't be translated literarily between two languages [33], we plan to add interlingual features so as to utilize multilingual resources such as aligned phrases and so on.

5 References

1. Seretan, V.: Syntax-based collocation extraction. In: Text, Speech and Language Technology Series. Springer Netherlands (2011)
2. Evert, S.: Corpora and collocations. In: Corpus Linguistics. An International Handbook, A. Lüdeling and M. Kytö, (ed.) pp. 1112-1248. Mouton de Gruyter, Berlin (2008)
3. Smadja, F.: Retrieving collocations from text: Xtract. Computational Linguistics. 19(1), 143–177 (1993)
4. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics. 19(1), 61–74 (1993)
5. Wermter J., Hahn U.: Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 843-850 (2005)
6. Lin D.: Extracting collocations from text corpora. In: Proceedings of the First Workshop on Computational Terminology, pp. 57–63. Montreal, Canada (1998)
7. Heid U., Weller M.: Tools for collocation extraction: preferences for active vs. passive. In: Sixth International Conference on Language Resources & Evaluation LREC, 24, pp. 1266-1272 (2008)
8. Scott M.: WordSmith Tools Version 5.0. Lexical Analysis Software, Liverpool (2008)
9. Li, D., Cao, J., Huang D.: A hierachical collocation extraction tool. In: The 5th IEEE International Conference on Big Data and Cloud Computing (BDCloud 2015). August 26-29, Dalian, China (2015) (in press)
10. He, D., Li, D. C. S.: Language attitudes and linguistic features in the "China English" debate. World Englishes. 28(1), 70–89 (2009)
11. Kirkpatrick, A., Zhichang, X. U.: Chinese pragmatic norms and 'China English'. World Englishes. 21(2), 269–279 (2002)
12. Wei Y., Jia, F.: Using English in China. English Today. 19(4), 42–47 (2003)
13. Du, R., Jiang, Y.: China English in the past 20 years. 33(1), 37-41 (2001)
14. Bolton, K., Graddol, D.: English in China today. English Today. 28(03), 3–9 (2012)
15. Yang, J.: Lexical innovations in China English. World Englishes. 24(4), 425–436 (2005)
16. Zhang, H.: Bilingual creativity in Chinese English : Ha Jin's in the pond. World Englishes. 21(2), 305-315 (2002)
17. Yu, X., Wen, Q.: The nativilized characteristics of evaluative adjective collocational patterns in China's English-language newspapers. Foreign Language and their Teaching. 5, 23-28 (2010)
18. Ai, H., You, X.: The grammatical features of English in a Chinese internet discussion forum. World Englishes. 34(2), 211–230 (2015)
19. Hamid, M. B., JR, R. B. B.: Second language errors and features of world Englishes. World Englishes. 32(4), 476-494 (2013)
20. Kachru, B. B.: World Englishes: approaches, issues and resources. Language Teaching. 25(01), 1-14 (1992)
21. Bahns, J.: Lexical collocations: a contrastive view. ELT Journal. 47(1), 56-63 (1993)

22. Benson, M., Benson, I., Robert, E.: The BBI combinatory dictionary of English: a guide to word combinations. pp. x-xxiii. Benjamins John, New York (1986)
23. Sinclair, J.: Corpus, Concordance, Collocation. Shanghai Foreign Language Education Press, Shanghai (2000)
24. Mckeown, K. R., Ravd, D. R.: Collocations. Handbook of Natural Language Processing, Dale, R., Moils, H., Somers, H. (eds.) pp. 1-19. CRC Press (2000)
25. Firth, J. R.: A synopsis of linguistic theory, 1903-1955. In: Studies in Linguistic Analysis (Special volume of the Philological Society), pp. 1-15 (1962)
26. Bartsch, S., Evert, S.: Towards a Firthian notion of collocation. Online publication Arbeiten zui Linguistik. 2, 48-60 (2014)
27. Kiss, T., Strunk, J.: Unsupervised multilingual sentence boundary detection. Computational Linguistics. 32, 485-525 (2006)
28. Bird, S., Loper, E.: NLTK: the Natural Language Toolkit. In: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Association for Computational Linguistics. Philadelphia (2002)
29. Klein, D., Manning, C. D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430 (2003)
30. Miller, G. A.: Wordnet: a lexical database for English. Communications of the ACM. 38(11), 39-41 (1995)
31. Lin, D.: Automatic identification of non-compositional phrases. In: Proceedings of ACL 1999, pp. 317–324. University of Maryland, Maryland, USA (1999)
32. Alvaro, J. J.: Analyzing China's English-language media. World Englishes. 34(2), 260–277 (2015)
33. Pereira, L., Strafella, E., Duh, K., Matsumoto, Y.: Identifying collocations using cross-lingual association measures. In: ACL 2014 14th Conference of the European Chapter of the Association for Computational Linguistics Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pp. 26-27 (2014)