International Journal of English Linguistics; Vol. 5, No. 4; 2015
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Using Sketch Engine to Investigate Synonymous Verbs

Chunyu Hu[1] & Bei Yang[2]

[1] School of English for International Business, Guangdong University of Foreign Studies, China

[2] School of English and Education, Guangdong University of Foreign Studies, China

Correspondence: Bei Yang, School of English and Education, Guangdong University of Foreign Studies, China. E-mail: yangbeixc@163.com

Received: May 28, 2015 Accepted: June 13, 2015 Online Published: July 30, 2015

doi:10.5539/ijel.v5n4p29 URL: http://dx.doi.org/10.5539/ijel.v5n4p29

The research was supported by Foundation for Outstanding Young Teachers, Guangdong Province, China (GWTP-SY-2014-08) to BY and Innovative School Project in Higher Education of Guangdong, China (GWTP-BS-2014-13) to BY.

Abstract

Synonymy is an important yet intricate linguistic feature in the field of lexical semantics. Using the 100 million-word British National Corpus (BNC) as data and the software Sketch Engine (SkE) as analyzing tool, this study examines the usage differences between *raise* and *increase*, two synonymous verbs notorious for their complex semantic and syntactic usage patterns. In addition to examining the collocates of the verbs, the study also investigated the syntactic patterns that the verbs typically occupy in the sentence structure and their functional implications. The data analysis yields an informative delineation of the internal semantic structure of the synonym set. The results also show the need for the corpus approach to go beyond collocational analysis in the study of synonymous verbs. The limitations of using SkE to extract and disambiguate synonyms are also addressed. This paper ends by discussing the pedagogical implications that this research may have when the results are introduced into the classroom.

Keywords: synonymy, lexical semantics, collocation, colligation, BNC, Sketch Engine

1. Introduction

Synonymy, or semantic equivalence, is an important yet intricate linguistic feature in the field of lexical semantics. Synonyms are not completely interchangeable; rather, they differ in shades of meaning and vary in their connotations, implications, and register (DiMarco et al., 1993). Any natural language consists of a considerable number of synonymous words. Due to historical reasons, English is particular rich in synonyms, which enables English speakers "to convey meanings more precisely and effectively for the right audience and context" (Liu & Espino, 2012, p. 198), but also constitute a thorny area for EFL (English as a Foreign Language) learners because of their subtle nuances and variations in meaning and usage.

It thus comes no surprise that an important aspect of English linguistics is to find the proper measures of automatically identifying and extracting synonyms (Peirsman, Geeraerts, & Speelman, 2015) and of distinguishing one word from its synonyms or near-synonyms (Hanks, 1996; Biber et al., 1998; Gries, 2001; Xiao & McEnery, 2006; Divjak, 2006; Gries & Otani, 2010; Liu, 2010). Although the two orientations of researching synonyms are equally important, I will in this paper focus more attention on the second one. The main purposes of this study are methodological, in that I would like to discover what the relative strengths and weaknesses of using Sketch Engine to research synonyms are, and what their relative scope of applicability is.

The rest of this paper is structured as follows. In the next section, I will give an overview of related work by introducing corpus studies of lexical semantics in the first place, and then discussing corpus-based automatic extraction and discrimination of synonymous words. Section 3 will present corpus data and tools used in this study. The results of this study are presented and discussed in Section 4, where I show the success of Sketch Engine in researching synonyms. The final section summarizes major findings and pointers for future research.

2. Related Work

2.1 Corpus Studies of Lexical Semantics

In the field of lexical semantics, there are a number of closely related key issues such as "How do we know what words mean? What evidence do we have? Is this evidence observable and objective? How can large text collections (corpora) be used to study what words mean?" (Stubbs, 2001, p. 4). For centuries, researchers, language teachers, and dictionary makers have used both their own intuitions and also attested uses of words, often in the form of thousands of quotations from printed books. However, it is only since the mid-1980s that corpus methods have been able to provide evidence about word meaning by searching across large text collections.

The approach of using corpus evidence to study meaning of words or phrases is often labeled as corpus semantics or empirical semantics, and the most active and influential scholars are called neo-Firthian corpus linguists. The leading figure is John Sinclair who might as well be one of the first people to bring Firth's ideas together with a corpus linguistic methodology (Stubbs 1996). Other important neo-Firthians include Michael Hoey, Susan Hunston, Bill Louw, Michael Stubbs, Wolfgang Teubert and Elena Tognini-Bonelli (McEnery & Hardie, 2012, p. 122).

At the core of the neo-Firthian school of corpus linguistics is searching for the units of meaning. The assumption that single words or lemmas are the main unit of meaning has underlain the construction of English-language dictionaries for hundreds of years. However, the work of Sinclair and associates provides a considerable amount of evidence that units of meaning are phraseological units instead of single words. Inspired by Firth's (1957, p. 179) maxim that "you shall know a word by the company it keeps", Sinclair has paid much attention to the context in which a word is used. He firmly believes in the principle of 'trust the text' (Sinclair, 2004) and claims that 'the language looks rather different when you look at a lot of it at once' (Sinclair, 1991, p. 100).

Reading concordance and calculating collocates from corpus are two important ways to study a lexical item in its context used by Sinclair, hence his well-cited book is entitled as *Corpus*, *Concordance, Collocation* (Sinclair, 1991). The concordance is the basic tool for anyone working with a corpus. Even far before the emergence of corpus linguistics, concordances to major works such as the Bible and Shakespeare have been available. The computer has merely made concordances easy to compile. For Sinclair (1991, p. 32), "A concordance is a collection of the occurrences of a word-form, each in its own textual environment. In its simplest form, it is an index. Each word-form is indexed, and a reference is given to the place of each occurrence in a text." In corpus linguistics, a simple and effective convention called KWIC (Key Word In Context) has been widely used.

Closely related to concordance is the notion of collocation. Firth (1957, p. 181) defines collocations of a given word as "statements of the habitual and customary places of that word". Nevertheless, Firth's research on collocation is largely intuition-based, which is in sharp contrast with most corpus linguists' belief that the only way to reliably identify the collocates of a given word is to study patterns of co-occurrence in a corpus. For example, Hunston (2002, p. 68) argues, "Collocation may be observed informally in any instance of language, but it is more reliable to measure it statistically, and for this a corpus is essential."

The idea that Firth proposed is operationalized by Sinclair and associates' early work from 1970 (reprinted in 2004) which may be considered a methodological elaboration on the concordance. A collocation is a cooccurrence pattern that exists between two items that frequently occur in proximity to one another—but not necessarily adjacently or, indeed, in any fixed order. Closely related to collocation is the notion of node and collocates. A node is an item whose total pattern of co-occurrence with other words is under examination; a collocate is any one of the items which appears with the node within a specified span (Sinclair et al., 2004, p. 10). Collocates are also determined within particular spans: "Two other terms . . . are span and span position. In order that these may be defined, imagine that there exists a text with types A and B contained in it. Now, treating A as the node, suppose B occurs as the next token after A somewhere in the text. Then we call B a collocate at span position +1. If it occurs as the next but one token after A, it is a collocate at span position +2, and so on."(Sinclair et al., 2004, p. 34)

In order to test whether two words are significant collocates, four pieces of data are required: the length of the text in which the words appear, the number of times they both appear in the text, and the number of times they occur together (Sinclair et al., 2004, p. 28). The optimal span is 4:4, as demonstrated in Sinclair's (1991: 170) definition of collocation, "Collocation is the co-occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening". On the basis of Sinclair's work, Hoey (2005, p. 5) defines collocation as "a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is

explicable in terms of random distribution".

The units Sinclair argues for, units which reach beyond the word and thus incorporate the collocations of words, are referred to either as extended units of meaning or as lexical items (Sinclair, 1996, 2004). Stubbs (2001, 2009) develops Sinclair's ideas into a systematic account of how the extended lexical units around a word may be studied by the successive analysis of collocations, colligations, semantic preferences and semantic or discourse prosodies. Colligation, semantic preference and discourse prosodies are all abstractions of collocation – that is, they are built upon a collocation analysis.

In sum, Sinclair and his associates have shown that lexical items tend to occur in particular linguistic contexts, e.g. they tend to co-occur or collocate with certain other words, phrases, and/or grammatical structures, and these distributional tendencies help define their meanings. Sinclair's pioneering work has shaped contemporary research on lexical semantics, leading to experimental and corpus approaches to the synonymous words.

2.2 Corpus Approaches to Synonyms

Boosted by the advent of the computer era and the central ideas of corpus semantics, the past decades have witnessed significant advances in the studies on synonymy. Based on the Brown Corpus, Miller & Charles (1991) found that the more two words are judged to be substitutable in the same linguistic context (i.e. the same location in a sentence), the more synonymous they are in meaning. Employing a "lexical substitutability" test in a corpus study of the near-synonyms *ask for*, *request* and *demand*, Church et al. (1994) produced the same finding: the substitutability of lexical items in the same linguistic context constitutes a good indicator of their semantic similarity. Gries (2001) quantifies the similarity between English adjectives ending in *-ic* or *-ical* (like *economic* and *economical*) on the basis of the overlap between their collocations. Gilquin (2003) investigates the difference between the English causative verbs *get* and *have*, Glynn (2007) compares intra- and extralinguistic factors in the contexts of *hassle*, *bother* and *annoy*, and Gries & Otani (2010) study the synonyms *big*, *great* and *large* and their antonyms *little*, *small* and *tiny*. Other sets of synonyms that have attracted attention include *strong* and *powerful* (Church et al., 1991), *absolutely*, *completely* and *entirely* (Partington, 1998), *big*, *large* and *great* (Biber et al., 1998), *quake* and *quiver* (Atkins & Levin, 1995), *principal*, *primary*, *chief*, *main* and *major* (Liu, 2010), and *actually*, *genuinely*, *really*, and *truly* (Liu & Espino, 2012)

One corpus-based approach to synonyms is sometimes labeled as corpus-based behavioral profile (BP) study. Generally, a BP study uses corpus data to examine the distributional patterns of lexical items, such as the linguistic contexts a word is typically used in and the words it usually collocates with, so as to identify its unique semantic and usage patterns. For instance, Hanks (1996) examined the syntactic and collocational patterns of the verbs *urge* and *incite*, including the types of subjects (such as animate or inanimate) and the types of complementation structures each verb typically takes (such as a simple object complement vs. a complement involving an object noun plus an infinitive complement as shown in "Rice *urged* the president to resolve the issue"). He also investigated, among other things, the semantics of the complement structures (i.e., whether the instances of the typical complement structure of a verb are positive or negative in meaning). The results of the examination helped uncover the behavioral profiles of the verbs, which in turn Behavioral Profile study of near-synonymous adverbs revealed the primary and secondary meanings of each verb and differentiated it from its synonyms. For instance, in the case of the verb *urge*, its behavioral profile distinguishes it from its near-synonyms like *ask*, *request*, and *order*, because the latter verbs do not share the same complement collocation patterns, among other profile features, with the verb *urge*.

In recent years, Gries and associates (Divjak & Gries, 2006; Gries, 2001; Gries & Otani, 2010) have developed a more sophisticated BP approach in examining both adjectives and verbs. In this approach, they first imported all the relevant corpus data into a spreadsheet, then manually annotated all the linguistic and contextual features they considered relevant, and finally analyzed the annotated data using a statistical program designed specifically for BP research called "R script BP 1.0". The types of linguistic and contextual features they annotated for synonymous verbs included, among others, tense/aspect, the types of complements, and clause types. By examining the various distributional features of the synonyms, such corpus-based BP studies have been able to effectively identify the internal semantic structures of the synonym sets being examined, including the fine-grained semantic differences among the synonyms in each set, an important type of information in the study of synonymy that traditional research methods had difficulty uncovering.

Nevertheless, the BP approach developed by Gries and associates might be complex for pedagogical purpose and thus the scope of its application may be limited. This study, based on a leading corpus tool Sketch Engine, aims to introduce a simple method that can be widely used by researchers, language teachers and even EFL students.

3. Method

3.1 Corpus Data: BNC

The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written (Aston & Burnard, 1998). The written part of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10%) consists of orthographic transcriptions of unscripted informal conversations and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

BNC is, by nature, monolingual, synchronic, general and sample-based, in that it deals with modern British English, it covers British English of the late twentieth century, it includes many different styles and varieties instead of being limited to any particular subject field, genre or register, and that it contains many samples which allows for a wider coverage of texts within the 100 million limit. The corpus is encoded according to the Guidelines of the Text Encoding Initiative (TEI) to represent both the output from CLAWS (automatic part-of-speech tagger) and a variety of other structural properties of texts (e.g. headings, paragraphs, lists etc.). Full classification, contextual and bibliographic information is also included with each text in the form of a TEI-conformant header.

3.2 Corpus Tool and Analysis Procedure

The Sketch Engine (SkE) is a leading corpus tool, widely used in lexicography, language teaching, translation and the like (Kilgarriff et al. 2004). It actually refers to two different things: the software, and the web service. The web service includes, as well as the core software, a large number of corpora pre-loaded and 'ready for use', and tools for creating, installing and managing users' own corpora. Corpora in SkE are often annotated with additional linguistic information, the most common being part of speech information (for example, whether something is a noun or a verb), which allows large-scale grammatical analyses to be carried out.

SkE has a number of core functions: Thesaurus, Wordlist, Concordance, Collocation, word sketches, and Sketch Diff. I will introduce most of them that are relevant for this study.

3.2.1 Thesaurus

In Sketch Engine the automatic identification of synonymy is achieved by the tool Thesaurus. SkE prepares a 'distributional thesaurus' for a corpus, a thesaurus created on the basis of common collocation. If two words have many collocates in common, they will appear in each other's thesaurus entry. For example, if we find instances of both *raise revenue* and *increase revenue*, that is one small piece of evidence that the two verbs *raise* and *increase* are similar. We can say that they 'share' the collocate *revenue* (noun), in the OBJECT relation. In a very large computation, for all pairs of words, we compute how many collocates they share, and the ones that share most (after normalization) are the ones that appear in a word's thesaurus entry. The thesaurus entry for the verb *raise* is shown in Figure 1. The similar words of *raise* are clustered into three categories: *need*, *increase*, and *spend*.

raise *(verb)*
British National Corpus freq = 19,118 (170.4 per million)

Lemma	Score	Freq	Cluster
need	0.31	64,447	receive [0.307, 24,431] mean [0.297, 65,954] provide [0.297, 50,239] give [0.295, 123,424] offer [0.291, 28,778] produce [0.29, 30,045] show [0.289, 58,782] bring [0.288, 42,286] include [0.287, 35,139] create [0.285, 21,207] change [0.284, 26,629] involve [0.278, 22,678] require [0.278, 28,257] put [0.277, 67,367] keep [0.273, 48,581] allow [0.271, 33,577] consider [0.266, 28,517] affect [0.266, 13,195] accept [0.265, 19,841] represent [0.265, 15,402] want [0.264, 87,784] take [0.264, 173,412] expect [0.258, 28,077] see [0.256, 185,151] follow [0.256, 45,071] find [0.251, 95,850] discuss [0.245, 14,764] leave [0.245, 61,413] establish [0.244, 17,397]
increase	0.303	19,059	lose [0.297, 26,605] reduce [0.291, 17,679] hold [0.277, 46,534] suggest [0.274, 28,246] reach [0.269, 22,409] support [0.266, 18,597] set [0.265, 38,838] draw [0.264, 21,631] carry [0.263, 30,571] achieve [0.263, 16,710] help [0.258, 40,484] seek [0.258, 16,605] claim [0.258, 18,672] introduce [0.257, 14,271] meet [0.254, 32,929] generate [0.253, 5,433] decide [0.25, 23,825] indicate [0.249, 12,196] cover [0.249, 18,771] reflect [0.248, 11,127] present [0.247, 14,079] begin [0.245, 40,126] form [0.245, 18,198] reveal [0.244, 10,176] express [0.243, 11,959] place [0.243, 14,640]
spend	0.278	22,062	pay [0.26, 37,379]

Figure 1. Clusters of similar words of *raise* in BNC

3.2.2 Concordance

The basic method in SkE to generate concordance lines is from the simple search form, as in Figure 2.

Figure 2. Simple search form

Users, however, often want more control than the simple search offers. By clicking on‘Query types’ they see the options as in Figure 3, and can specify a lemma (with optional word class, e.g. verb, noun, adjective) or a specific phrase or word form (with an option to match for case).

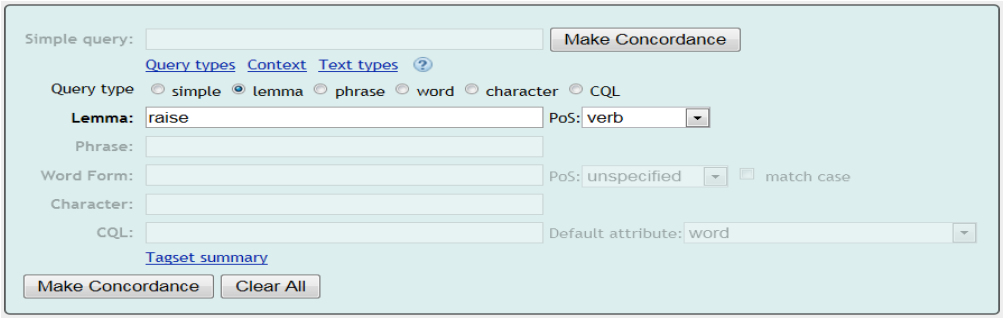

Figure 3. Query type for searching the lemma *raise* as verb

If uses click the button ‘Make Concordance’, the software will generate a number of lines.

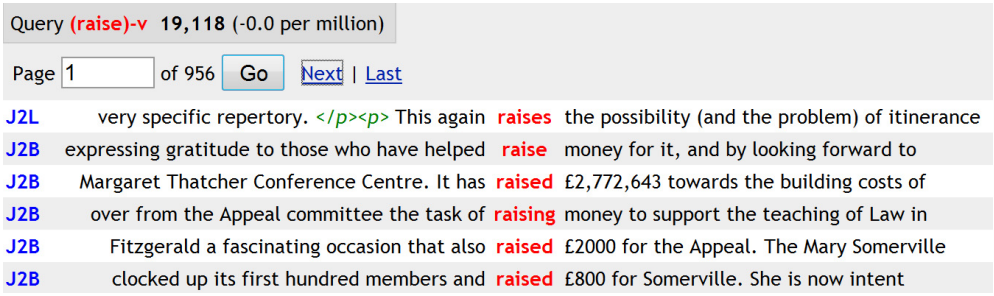

Figure 4. Search hits for the verb *raise* in BNC

3.2.3 Collocations

Closely related to Concordance is Collocates. In the corcordancing interface like Figure 4, if we click the ‘Collocations’ menu, a new box will jump up, as demonstrated in Figure 5 below:

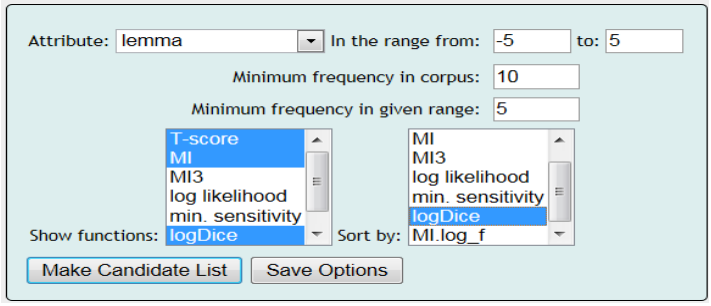

Figure 5. Collocation candidates

Since some collocates of *raise* may have different forms (for example, *rate* and *rates*), in ‘Attribute’ we choose

lemma. The span (the number of words left and right of the search word) is (-5, 5), the minimum frequency of each collocate being set 10 and minimum frequency in given range (in our case -5, 5) 5. Of seven measures to calculate the strength of collocation (T-score, MI, MI3, log likelihood, min. sensitivity, and LogDice), I choose the default one *logDice* which is considered more reliable than the frequently used MI (mutual information) measure.

3.2.4 Word Sketch

The function that gives the Sketch Engine its name is the word sketch: a one-page summary of a word's grammatical and collocational behavior. Figure 6 demonstrates the word sketch for *increase* (verb). Its collocates are grouped according to grammatical relations in which they occur. In the first column, for example, a number of words such as *wage*, *cent*, *population*, *pay*, *spending* and *tax* are grouped into the category *subject*, i.e., they are used as the subject of *increase*.

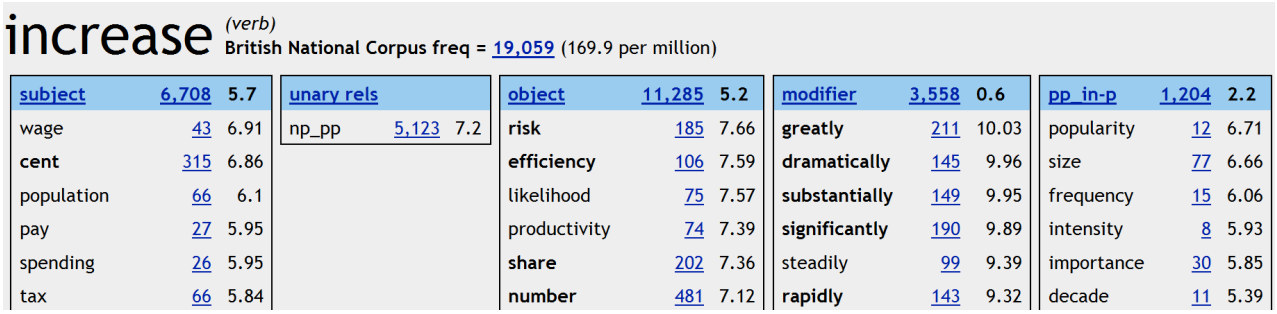

Figure 6. Word sketch for the verb *increase* in BNC

3.2.5 Sketch-Diff

This function is probably the most straightforward when researching synonymous words. Figure 7 presents the interface of Sketch-Diff. When users click the button 'Show Diff', the software will generate a summary-list of two synonymous words in terms of collocates arranged by grammatical categories.

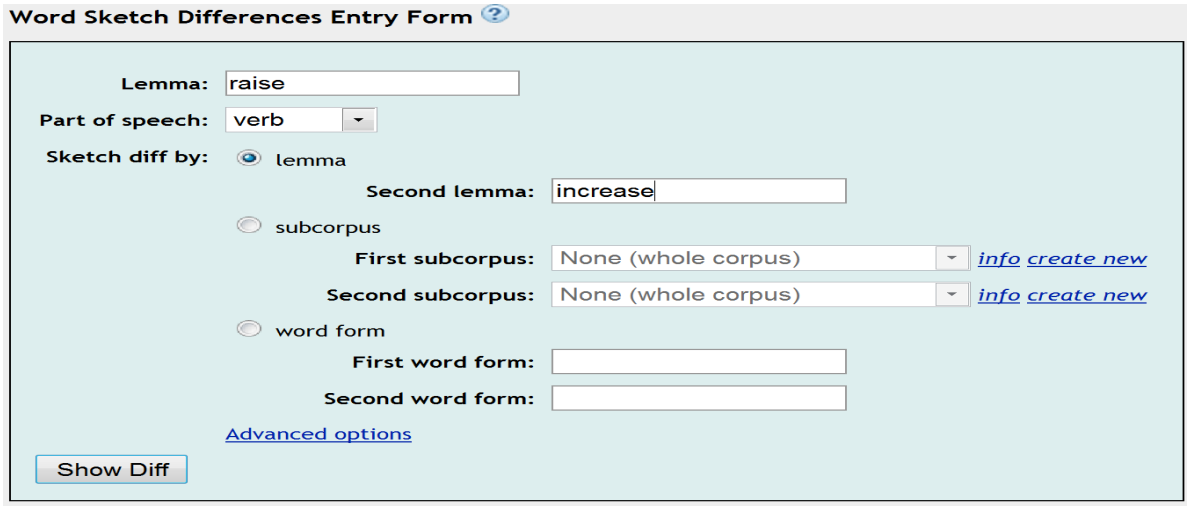

Figure 7. The interface of Sketch-Diff

4. Results and Analysis

4.1 The Frequencies of Raise and Increase

Concordance enables researchers to compare frequencies of synonymous words. As shown in Table 1, while *increase* as a noun is much more frequent than *raise*, the two words as verb are quite close in terms of frequency.

Table 1. Frequency of *raise* and *increase* in BNC (per million)

	Total	Verb	Noun
raise	171.4	170.4	0.9
increase	275.4	169.9	105.5

4.2 The Collocates of Raise and Increase

Table 2 and Table 3 list the top 50 collocates of the verb *raise* and *increase* automatically generated by the software:

Table 2. The top 50 collocates of *raise* (verb) in BNC

Rank	Collocates	Freq.	logDice	Rank	Collocates	Freq.	logDice
1	eyebrow	516	9.697	26	capital	174	7.426
2	question	1,382	9.53	27	price	251	7.424
3	issue	1,113	9.408	28	level	302	7.407
4	money	1,156	9.396	29	appeal	166	7.332
5	fund	634	9.282	30	leg	147	7.31
6	awareness	207	8.226	31	above	213	7.3
7	tax	340	8.163	32	profile	101	7.232
8	cash	214	7.977	33	doubt	153	7.208
9	objection	166	7.966	34	problem	328	7.166
10	head	487	7.949	35	eye	242	7.141
11	hand	553	7.926	36	whether	233	7.128
12	arm	284	7.887	37	aim	141	7.079
13	charity	170	7.886	38	temperature	100	7.04
14	alarm	157	7.868	39	sum	101	6.99
15	standard	275	7.768	40	his	1,558	6.895
16	revenue	155	7.7	41	her	795	6.871
17	rate	309	7.647	42	pound	134	6.864
18	voice	280	7.644	43	knee	81	6.821
19	million	265	7.64	44	expectation	81	6.809
20	point	483	7.618	45	help	232	6.755
21	glass	171	7.521	46	by	1,748	6.751
22	matter	281	7.52	47	concern	111	6.741
23	possibility	156	7.486	48	hon.	96	6.728
24	hope	254	7.439	49	important	187	6.727
25	finance	148	7.436	50	about	679	6.72

As shown in Table 2, the dominant collocates of *raise* are nouns which can be grouped into four categories:

- Physical organs: eyebrow, head, hand, arm, leg, eye, knee;
- Physical items: voice, glass, hon;
- Business and economic terms: money, fund, tax, cash, revenue, rate, finance, capital, price;
- Abstract nouns: question, issue, awareness, objection, charity, standard, point, possibility

Other collocates such as *above*, *whether*, *by, about*, *his* and *her* have much to do with the grammatical relation which will be analyzed in the next section.

Table 3. The top 50 collocates of *increase* (verb) in BNC

Rank	Collocates	Freq.	logDice	Rank	Collocates	Freq.	logDice
1	per	929	8.709	26	productivity	119	7.537
2	number	994	8.674	27	cost	315	7.532
3	cent	680	8.598	28	demand	225	7.493
4	rate	590	8.582	29	amount	210	7.474
5	greatly	221	8.34	30	steadily	111	7.459
6	significantly	218	8.264	31	gradually	121	7.452
7	dramatically	169	8.072	32	spending	121	7.445
8	substantially	167	8.042	33	level	308	7.437
9	pressure	264	8.006	34	speed	150	7.429
10	size	260	7.989	35	value	240	7.424
11	risk	258	7.964	36	reduce	191	7.412
12	production	256	7.9	37	considerably	114	7.412
13	share	331	7.869	38	expenditure	126	7.381
14	price	338	7.855	39	total	186	7.34
15	tax	274	7.854	40	by	2617	7.334
16	rapidly	166	7.852	41	volume	127	7.324
17	decrease	146	7.8	42	increase	243	7.317
18	population	214	7.708	43	capacity	122	7.305
19	output	162	7.701	44	its	839	7.258
20	income	197	7.634	45	million	202	7.251
21	efficiency	138	7.631	46	investment	146	7.25
22	concentration	148	7.601	47	export	114	7.236
23	sale	225	7.593	48	awareness	103	7.223
24	proportion	156	7.554	49	power	252	7.158
25	profit	176	7.543	50	budget	123	7.126

It is clear that the dominant collocates of *increase* are also nouns which fall mainly into two categories:

- Business and economic terms: number, rate, size, risk, production, share, price, tax, population, output, income, sale, profit, productivity, cost, demand, spending, expenditure, volume, investment, export, budget;
- Abstract nouns: pressure, efficiency, concentration, value, awareness, power.

In addition to nouns, adverb collocates are also quite salient. Of 50 collocates there are 8 adverbs: *greatly*, *significantly*, *dramatically*, *substantially*, *rapidly*, *steadily*, *gradually* and *considerably*. Words describing numbers or percentage (such as *per*, *cent*, *million*) also frequently collocates with *increase*. It seems that the dominant collocates of *increase* are also nouns which have much to do with amount, number, or value.

4.3 The Syntactic Patterns of Raise and Increase

The syntactic patterns of the two verbs are based on the Word Sketch function of SkE as demonstrated in Figure 5. In order to present a fine-grained comparison, I summarized the 17 patterns of *raise* and 21 patterns of *increase* in Table 4 and Table 5. In the first example of Table 4, the redden word *eyebrows* functions as the object of *raise*.

Table 4. The syntactic behavior of *raise* (verb) in BNC

Categories	Freq	Score	Example
object	15786	6.9	Malcolm **raised** his ***eyebrows*** at me
subject	4273	3.4	***Gentleman*** again **raised** the question of law and order.
modifier	2128	0.3	but ***inevitably*** **raising** the perennial debate
pp_by-p	708	4.1	the monies **raised** by carbon ***taxation*** will
pp_in-p	620	1.1	a voice **raised** in ***anger***
and/or	241	0.1	By **raising** and ***lowering*** the handle it is possible to
pp_to-p	241	0.7	seven of these were **raised** to the ***status*** of embassies
pp_from-p	182	1.3	Many herbs can be **raised** from ***seed***.
pp_on-p	155	0.7	The dome is **raised** on a ***drum*** which has 16 windows
part_trans	125	1.0	felt his strong arms **raising** her ***up***
pp_for-p	109	0.4	£50 million has been **raised** for ***charity***
pp_at-p	101	0.7	The subject had been **raised** at a ***meeting*** between

np_adj_comp	69	2.0	**raised** himself a little ***higher*** on his elbow
pp_through-p	55	2.2	a third of this will be **raised** through ***sponsorship***
part_intrans	50	0.2	The disk is round, sack-like, often **raised *up*.**
adj_comp	35	0.4	stood with her slender arms **raised *high***
pp_above-p	35	10.9	His arms were **raised** above his ***head*…**

Table 5. The syntactic behavior of *increase* (verb) in BNC

Categories	Freq	Score	Example
object	11285	5.2	Any combination of these factors **increases** the **risk** dramatically
subject	6708	5.7	a statutory national minimum **wage** would **increase** unemployment
modifier	3558	0.6	This will **dramatically increase** the risk of subsidence
pp_in-p	1204	2.2	Employee trusts have **increased** in **popularity** over the past decade
pp_by-p	1036	6.3	Weight **increased** by an **average** of 5.2 kg.
pp_to-p	390	1.3	the potential energy **increases** to a **maximum**
pp_with-p	357	1.9	Now what we've got there is income **increasing** with **age** .
pp_from-p	321	2.4	the potential energy **increases** to a **maximum**
and/or	229	0.1	we reserve the right to **increase** or **decrease** brochure prices
pp_at-p	130	0.9	If CFC emissions continue to **increase** at the current **rate**
pp_over-p	126	5.4	petrol prices will **increase** over the next few **years**
pp_as-p	109	1.9	tax revenues … **increased** as a **proportion** of national income
pp_during-p	96	7.8	urban overcrowding probably **increased** during this **period**
pp_for-p	85	0.3	the price has not been **increased** for more than 10 **years**
adj_comp	83	0.9	Complaints have **increased five-fold** in two years.
part_intrans	82	0.4	Carbon dioxide emissions will **increase by** between 9 and 23 per cent
np_adj_comp	46	1.4	in order to **increase** the memory **available**
part_by-a_obj	21	273.2	death rate among teenagers **increased** by over 40 per **cent**
pp_towards-p	19	2.2	The use of gunboats **increased** towards the **end** of the war
part_trans	14	0.1	Did "Toothless" sales **increase over** the six week period?

It has to be noted that although the syntactic patterns of the two verbs are similar in many ways, there also exist apparent differences, which can be easily shown when using Sketch-Diff function of SkE.

4.4 Direct Comparison of Lexical and Grammatical collocates

The Sketch-Diff function of SkE allows users to visually compare and contrast synonymous words according to their salient collocational context. Figure 8 is part of the result when clicking 'Show Diff' in Figure 7. In the figure, the greener a word is, the more closely it relates to *raise*. The redder a word is, the more closely it relates to *increase*. For example, it is more usual to say *to raise or lower* than *to increase or lower* and similarly, it is more fluent to say *to increase or decrease* than *to raise or decrease*.

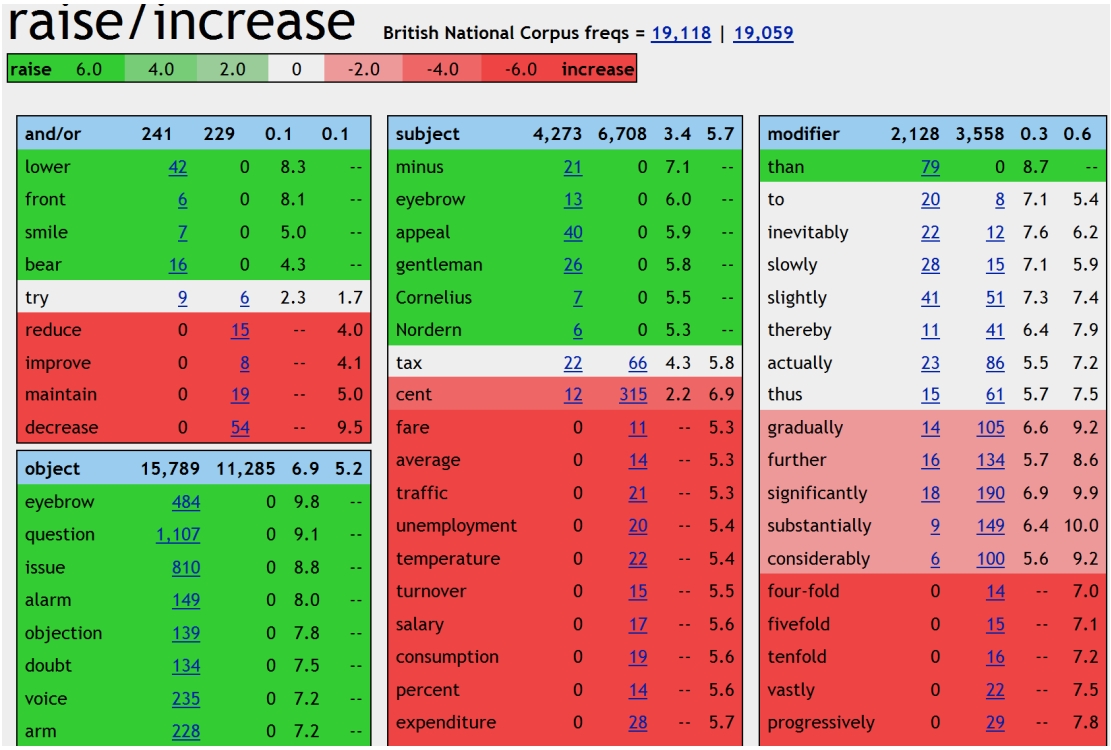

Figure 8. Comparison of *raise* and *increase* in terms of collocational patterns

Apparently, despite that the two verbs *raise* and *increase* share a number of syntactical patterns, the collocates in each pattern differ considerably. In the 'and/or' pattern, for example, *lower* frequently collocates with *raise* but never used with *increase*. On the other hand, *decrease* occurs 54 times with increase, but there is no occurrence of *decrease* with *raise*.

In the 'modifier' pattern, there are many words (such as *fourfold*, *fivefold*, *vastly*) that only collocate with increase. Even if some words (such as *gradually*, *further*, *significantly*, *substantially*, *considerably*) do collocate with *raise*, their occurrences with *increase* are much higher. It is thus no surprise that collocation tokens of for *raise* are only 2128 but 3558 for *increase.*

In the 'object' pattern, the collocation tokens for *raise* are 15789 and only 11285 for *increase,* indicating that there are more words used as objects of *raise*. Words like *eyebrow, arm, head, issue, question, doubt, matter*, only collocate with *raise* instead of *increase*. Words that collocate with both verbs have substantially different frequencies as illustrated in Table 6.

Table 6. Frequency comparison of words used as objects of both *raise* and *increase*

	raise	*increase*		*raise*	*increase*
money	1066	13	productivity	23	74
cash	175	6	number	92	481
fund	489	18	efficiency	10	106
standard	228	10	share	9	202
profile	85	6	risk	8	185
possibility	135	28	likelihood	0	75

It is an interesting observation that possibility and likelihood are semantically similar, but the former mainly collocates with *raise* and the latter with *increase* (Seretan, 2011, pp. 15-17).

5. Discussion

So far we have demonstrated how to use some core functions of SkE to research two synonymous verbs *raise* and *increase*. Each function, as noted before, has its advantages and disadvantages. Concordance not only enables us to look at re-occurring patterns of the words under investigation, it can also provide the frequency

information of the synonymous words as demonstrated in Table 1. Concordance can also make the invisible patterns visible as wisely pointed out by Tognini-Bonelli (2001, p. 18): "In an individual text, we can observe neither repeated syntagmatic relations nor any paradigmatic relations at all, but it is precisely these two things which concordances make visible". Because it gives access to many important language patterns in texts, the concordance is considered "at the centre of corpus linguistics" (Sinclair 1991, p. 170).

Important as concordance is, given the situation that the concordance of both *raise* and *increase* consist of nearly two hundred thousands of concordance lines, it would be more valuable to find a list of collocates which tend to occur near or next to the target item under investigation. Collocation thus plays a central role in the research of synonyms, as strongly articulated by Gries (2001, p. 82): the meaning of words can be defined "in terms of their significant collocates". Word sketch enriches the traditional study of collocation by providing syntactic patterns between the node (*raise* or *increase* in our case) and the collocates, as demonstrated in Figure 6 and Table 4 & 5.

On the top of all these, Sketch-Diff seems to be the easiest and the most straightforward method to distinguish one word/phrase from the other. Nevertheless, Sketch-Diff alone is not sufficient to demonstrate the semantic and syntactic features of words/phrases under investigation." To begin with, the summary list like Figure 8 is incomplete. Many important collocates may be missing. For example, the word *rate* is an important collocate for both *raise* and *increase* as shown in Table 2 & 3. Nevertheless, it is neither found in the 'object' pattern nor in the 'subject' pattern generated by Sketch-Diff. Below are two examples in which *rate* is used as the object of both verbs.

(1) Germany promises not to ***raise*** **interest rates** but refuses to **lower** them. (BNC-HLP)

(2) For example, if the government were to ***increase*** the **rate** of VAT gross turnover may increase (and with it the rent) while the tenant's net profit remains static. (BNC-J6R)

Further investigation indicates that while increase can make rate its object (other attested examples include *increased respiratory rate*, *increased the rate of soil evaporation*, etc.), it is *raise* rather than *increase* that typically collocates with *interest rate(s)*.

In addition to the above limitation, using Sketch-Diff alone will make users of SkE lose the opportunity to take an overall look at the collocates and syntactic patterns of synonyms as a whole. For example, we might have lost the opportunity to observe and categorize the noun collocates of *raise* and *increase* and then find the subtle differences between the two verbs in terms of noun collocates.

In a nutshell, while using Sketch-Diff function alone can give researchers a quick glance at the apparent differences between synonyms in the light of both collocations and syntactic patterns, it would be more rewarding to examine synonyms by way of other core functions of SkE.

It has to be pointed out that SkE has not without its limitations. One apparent limitation is its automatic extraction of similar words. In Figure 1, some of the similar words provided by the tool Thesaurus seems to have little similarity with *raise*. The measure of automatically identifying and extracting synonyms in a recent study (Peirsman *et al*. 2015) might be able to help SkE to improve its accuracy.

Another problem facing SkE is its accuracy of grammatical annotation (part of speech). As demonstrated in Figure 9, some uses of *raise* are typical of verb instead of noun.

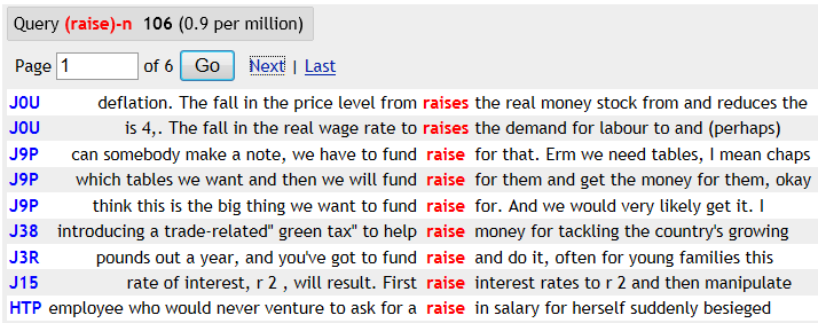

Figure 9. Search hits for *raise* (noun) in BNC

At the present stage, SkE still cannot semantically annotate a corpus as does another web-based corpus tool Wmatrix. This is not a problem, of course, but a direction that the SkE team may wish to endeavor in the near future.

6. Conclusion

In view of its importance and intricacy, researching synonymy is a crucial task in the field of lexical semantics. This paper has introduced the leading corpus tool SkE and its advantages in investigating synonymous verbs. The results show that different functions of SkE can make different contributions to the discrimination of *raise* and *increase*.

This study has also a number of pedagogical implications. In our teaching, we have noticed that students tended to confine their use of *raise* into a limited scope, such as *raise your hands* or/and *raise your voice*. Instead, they tended to overuse increase (such as *increase money*, *increase interest rate*, etc.) where raise might be more appropriate.

Studies in first language acquisition show that children memorize not only words in isolation, but also, to a large extent, groups (or chunks) of words. These chunks are viewed as the building blocks of language. They are available to speakers as ready-made or prefabricated units, contributing to conferring fluency and naturalness to their utterances. Thus, if EFL teachers aim to help their students to achieve a great amount of fluency and accuracy, they may hope to use examples extracted from corpus as in Table 4 & 5.

In view of the fact that there exist a huge amount of synonyms in English, it would be unlikely for teachers to teach each pair of them to students. It might be more promising to teach students how to use SkE to conduct their own research, hence the so-called Chinese saying, 'It's better to teach one fishing than to give him fish'.

References

Arppe, A. (2008). *Univariate, bivariate and multivariate methods in corpus-based lexicography: A study of synonymy*. (Unpublished doctoral dissertation). University of Helsinki, Helsinki, Finland.

Arppe, A., & Jävikivi, J. (2007). Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory, 3*(2), 131-159. http://dx.doi.org/10.1515/CLLT.2007.009

Aston, G., & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Atkins, B., & Levin, B. (1995). Building on a corpus: A linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography, 8*(2), 85-114. http://dx.doi.org/10.1093/ijl/8.2.855

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511804489

Church, K. W., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon* (pp. 115-164). Hillsdale, NJ: Lawrence Erlbaum.

Divjak, D., & Gries, S. Th. (2006). Ways of trying in Russian: Clustering behavioral profiles. *Journal of Corpus Linguistics and Linguistic Theory, 2*(1), 23-60. http://dx.doi.org/10.1515/cllt.2006.002

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In Philological Society (Eds.), *Studies in Linguistic Analysis* (pp. 1-32). Oxford, UK: Blackwell.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, and Computers, 28*(2), 197-202. http://dx.doi.org/10.3758/BF03204765

Fung, P., & McKeown, K. (1997). Finding terminology translations from nonparallel corpora. In J. Zhou & K. Church (Eds.), *Proceedings of the Fifth Workshop on Very Large Corpora* (pp. 192-202). Hong Kong/Beijing, China: The Hong Kong University of Science and Technology & Tsinghua University.

Geeraerts, D. (2010a). Lexical variation in space. In P. Auer & J. E. Schmidt (Eds.), *Language and Space. An International Handbook of Linguistic Variation* (pp. 820-836). Berlin, Germany: De Gruyter Mouton.

Geeraerts, D. (2010b). *Theories of Lexical Semantics*. Oxford, UK: Oxford University Press.

Geeraerts, D., Grondelaers, S., & Speelman, D. (1999). *Convergentie en Divergentie in de Nederlandse Woordenschat*. Amsterdam, Netherlands: Meertens Instituut.

Gilquin, G. (2003). Causative 'get' and 'have': So close, so different. *Journal of English Linguistics, 31*(2), 125-148. http://dx.doi.org/10.1177/0075424203031002002

Glynn, D. (2007). *Mapping meaning. Towards a usage-based methodology in Cognitive Semantics*. (Unpublished doctoral dissertation). University of Leuven, Leuven, Belgium.

Gries, S. Th. (2001). A corpus-linguistic analysis of -ic and -ical adjectives. *ICAME Journal*, *25*, 65-108.

Gries, S. Th., & Otani, N. (2010). Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal*, *34*, 121-150.

Gries, S. Th., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspectives on 'alternations'. *International Journal of Corpus Linguistics, 9*(1), 97-129. http://dx.doi.org/10.1075/ijcl.9.1.06gri

Hanks, P. (1996). Contextual dependency and lexical sets. *International Journal of Corpus Linguistics, 1*(1), 75-98. http://dx.doi.org/10.1075/ijcl.1.1.06han

Kilgarriff, A., & Yallop, C. (2000). What's in a thesaurus? In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Stainhauer (Eds.), *Proceedings of the 2nd Language Resources and Evaluation Conference (LREC 2000)* (pp. 1371-1379). Athens, Greece: European Language Resources Association.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7-36. http://dx.doi.org/10.1007/s40607-014-0009-9

Liu, D. (2010). Is it a chief, main, major, primary, or principal concern? A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics, 15*(1), 56-87. http://dx.doi.org/10.1075/ijcl.15.1.03liu

Liu, D., & Espino, M. (2012). Actually, Genuinely, Really, and Truly: A corpus-based Behavioral Profile study of near-synonymous adverbs. *International Journal of Corpus Linguistics, 17*(2), 198-228. http://dx.doi.org/10.1075/ijcl.17.2.03liu

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam, Netherlands: John Benjamins. http://dx.doi.org/10.1075/scl.2

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Sinclair, J. (1992). The automatic analysis of corpora. In J. Svartvik (Ed.), *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm*, *4–8 August 1991* (pp. 379-397). Berlin and New York: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110867275.379

Sinclair, J. (2004a). *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Sinclair, J. (2004b). Intuition and annotation: the discussion continues. In K. Aijmer & Altenberg (Eds.), *Advances in Corpus Linguistics* (pp. 39-60). Amsterdam: Rodopi.

Sinclair, J., Jones, S., Daley, R., & Krishnamurthy, R. (2004). *English Collocational Studies: The OSTI Report*. London: Continuum.

Stubbs, M. (1993). British traditions in text analysis: from Firth to Sinclair. In M. Baker, F. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 1-36). Amsterdam: John Benjamins. http://dx.doi.org/10.1075/z.64.02stu

Stubbs, M. (1996). *Text and Corpus Analysis*: *Computer Assisted Studies of Language and Culture*. Oxford: Blackwell.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Xiao, R., & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A crosslinguistic perspective. *Applied Linguistics*, *27*(1), 103-129. http://dx.doi.org/10.1093/applin/ami045

Yves Peirsman, Y., Geeraerts, D., & Speelman, D. (2015). The corpus-based identification of cross-lectal synonyms in pluricentric languages. *International Journal of Corpus Linguistics*, *20*(1), 54-80. http://dx.doi.org/10.1075/ijcl.20.1.03pei