Miura, A. (2016). Compiling domain-specific corpora with the sketch engine. The Joint International Conference of the 8th International Conference on ESP in Asia & the 3rd International Symposium on Innovative Teaching and Research in ESP in Japan, Tokyo, August 19, 2016. Tokyo: UEC IGTEE Research Station. Vol. 1, 44-51.

Compiling Domain-Specific Corpora using the Sketch Engine

Aika Miura

Tokyo University of Agriculture
dawn1110am@gmail.com

Abstract
This paper aims to describe domain-specific corpora compiled with the Sketch Engine. The author teaches English at Tokyo University of Agriculture and compiled "Agriculture Corpus Ver.1" (Miura, 2015) and "Agriculture Corpus Ver.2," via WebBootCaT, an online tool that is part of the Sketch Engine. During compilation, WebBootCaT crawls the internet and retrieves the available URLs based on a list of "seed words" that the compiler intends to include within the corpus. The seed words of the former corpus were somewhat random, taken from various sources of topics ranging from "cloning" to "environmental issues," while the latter specifically focused on agricultural vocabulary drawn from academic keywords given by the final-year undergraduates and academic staff at the Department of Agriculture. It was discovered that a narrow, i.e., more restrictive, domain of seed words resulted in richer corpus content as measured by the size, keyness, and collocational behaviors of genre-specific vocabulary.

Key words: corpus compilation, domain-specific corpora, the Sketch Engine, agriculture, genre-specific vocabulary

I. Introduction

Language educators, who are involved with English for specific purposes (ESP) at the tertiary level, are generally unfamiliar with the field in question since they are likely to have academic and educational backgrounds in linguistics, applied linguistics, or literary studies. It is necessary that ESP practitioners expand their knowledge of technical vocabulary outside of their area of expertise (Nation, 2001), and corpora are now a database that anyone can easily access (Chujo & Utiyama, 2006).

The purpose of referring to larger corpora is to explore general linguistic phenomena that are maximally "representative of the language variety" (McEnery, Xiao, and Tono, 2006, p.3). According to Koester (2010), large and balanced corpora such as the British National Corpus (BNC) are useful in providing insight into lexica-grammatical patterns of the language as a whole, while not identifying insight into patterns of language use in particular settings. In the fields of ESP or English for academic purposes (EAP), large corpora do not necessarily meet the teacher and learner requirements (Tribble, 2000). Therefore, smaller specialized corpora, which are domain-or-genre specific, should be compiled by carefully "target[ing] and set[ting] up to reflect contextual features" (Koester, 2010, p.67). Koester (2010) argues that specialized corpora should be designed to answer specific research questions, and any limitations regarding the representatives and balance should be resolved by referring to the contextual information of the target vocabulary.

This paper assists ESP practitioners by introducing a commercial web-based corpus tool, the Sketch Engine. This tool allows users to create original corpora using simple and near effortless procedures, as well as providing numerous ready-made corpora of a number of languages. First, the paper describes the process of automatic compilation of corpora based on the lists of "seed words," specifically targeted to the field of agriculture. Then, the process by which selection of seed words influenced the outcome of compilation is described, in terms of the size, keyness and lexical behaviors of the compiled corpora. It was discovered that the original compiled domain-specific corpora can contain more information, when compared with a larger and more balanced corpus such as the BNC.

II. What is the Sketch Engine?

The Sketch Engine is a useful corpus tool marketed by Lexical Computing Limited (Kilgarriff, Baisa, Bušta, Jakubíček, Kovář, Michelfeit, Rychlý, & Suchomel, 2014). It was originally developed for use by lexicographers to assist in the compilation of dictionaries, but it is now widely used by language learners, educators, practitioners, and researchers (Kilgarrif, Rychly, Smrž, & Tugwell, 2008).

This section provides a brief introduction of the tools that comprise the Sketch Engine. It is a commercial interface that provides corpora in many languages, including corpora such as the BNC and web-based mega corpora including enTenTen, ukWaC, jpTenTen, and JpWac.[1]

A. *Concordance, Word List, Word Sketch, Thesaurus, Sketch-Diff*

There are several basic tools and functions that increase the usability of the Sketch Engine. The first is *Concordance* that identifies the occurrences of a specified word or multi-word unit in various forms such as a simple surface form, lemma, and phrase within corpora. The second tool is *Word List* that creates a word list of the whole corpus or specified subcorpora, along with creating a keyword list that generates keywords unique to the target corpus, which are distinguishable from the reference corpus, i.e., keyness of the target corpora. The third tool is *Word Sketch* that provides lexical behaviors, showing collocational and colligational relations with the target search word, based on the output computed by logDice[2] (Lexical Computing Ltd., 2014, p.2). The fourth tool *Thesaurus* generates a list of words, each of which behaves in a manner similar to the target word retrieved from the corpus. The last tool is *Sketch-Diff* that illustrates the differences between the collocational behaviors between the two target terms.

III. Building Corpora Using the Sketch Engine

This section deals with the process of creating a corpus using the Sketch Engine. There are two methods. WebBootCaT, which is used for the first method, is a tool that automatically creates corpora by crawling the internet. The second method is performed by manual compilation of words from uploaded files in any format such as .doc(x), .pdf, and .txt. This section describes the former procedure, which is more efficient than the latter.

A. The Procedure of Creating a Corpus

Figure 1 shows a sample page of WebBootCaT within the Sketch Engine. To compile a corpus, the user merely clicks the top left link labeled “WebBootCaT” and names a corpus. Seed words are manually inserted into a text window labeled “Seed words” and then the tool crawls the internet retrieving available URLs. The seed words are those words that the compiler intends to include in the corpus.[3]

Table I and Table II show the total numbers and examples of seed words for “Agriculture Corpus Ver.1” and “Agriculture Corpus Ver.2,” as well as the outcome, or retrieved URLs that were generated by WebBootCaT.

For the first version, the author partitioned the set of seed words into three subsets: i) Undergrads, ii) Miscellaneous, and iii) Agriculture. The seed words were selected from available resources that provided easy access to the author as she was not familiar with agricultural fields at all. “Undergrads” is a collection of words that the author regarded as vocabulary belonging to the field of science and agriculture, which were extracted directly from the Tokyo University of Agriculture (TUA) entrance exam materials for undergraduate degrees from 2011 to 2012. “Miscellaneous” consists of a list of seed words taken from reading materials introduced by the author in a course named “Science English” that was created for 4th year undergraduates at the Faculty of Agriculture in 2015. The class students belonged to three different Departments, the Departments of Agriculture, Animal Science, and Human and Animal-Plant Relationships. Therefore, the articles covered various topics such as environmental problems and animal cloning. The third seed word list

[1] A user can gain access to these corpora collections at an annual cost of approximately 60 euros; this fee also allows the user to create a personal corpus with a maximum limit of one million words. Word tokens can be expanded for an additional fee.

[2] LogDice is a computational statistics tool based on the Dice coefficient, which is used to calculate collocations in the Sketch Engine (Lexical Computing Ltd., 2014, p. 2).

[3] The Sketch Engine suggests that we should input 3 to 20 words or phrases, but this study dealt with more.

"Agriculture" contains 10 English translations taken from research fields of some of the academic members from the Department of Agriculture in the University guidebook. In this department, the academics are involved in the areas of crop science, genetics and plant breeding, plant pathology, pomology, vegetables, floriculture, horticulture and other related fields.

Conversely, in compiling "Agriculture Corpus Ver.2," the author created lists of seeds words that were limited to the areas studied in the Department of Agriculture. The author queried those students who took the aforementioned course "Science English" in the 2016 academic year to provide their unique keywords in their field of study in English (or Japanese if they did not know English), while clarifying their laboratory participation as well. Then, the keywords and multi-unit keywords provided by all academic staff members in the Department of Agriculture, as currently described in the English version of the Department website, were added, as illustrated in Table II.

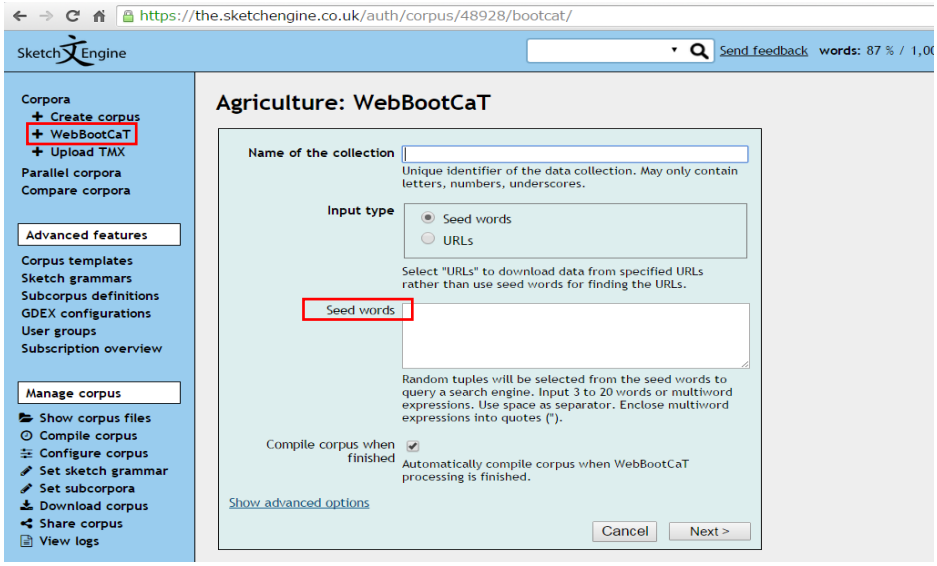

Figure 1. The sample page of WebBootCaT in the Sketch Engine.

Table I

Seed Words and Retrieved URLs for Agriculture Corpus Ver.1

List	Total No. of Phrases	Examples of Seed Words	No. of URLs
Undergrads	72 phrases (80 words)	species / plant / animal / organism / habitat / breed / variation / physical diversity / biologist /interbreed / reproductive isolation / genetic exchange / gene / morphology / pollutant	74
Miscellaneous	54 phrases (67 words)	cloning / ethical / molecular / feasible / cell / foetus / counterproductive /genetic diagnosis / branch / trunk / bud culture / shoot / sprout / cluster / genetically identical / old age	61
Agriculture	10 phrases (101 words)	biomass crop production / studies on non-tillage transplanting rice culture using green manure crops / exploration for natural enemies of some pest thrips	16
Totals	136 phrases (248 words)		151

Table II

Seed Words and Retrieved URLs for Agriculture Corpus Ver.2

List	Total No. of Phrases	Examples of Seed Words	No. of URLs
Crop Science	20 phrases (36 words)	compositing regulation / primary fermentation / spatulationn / global climate change / cultivation	63
Genetics and Plant Breeding	28 phrases (18 words)	chromosome / transposon / genome transformation / tissue culture / plant breeding / genetics / epigenetics	110
Plant Pathology	20 phrases (11 words)	phytopathology / plant diseases / biocontrol /plant viruses / biocontrol / microorganism / plant rescue	39
Entomology	38 phrases (27 words)	insect taxonomy / morphology / genus / tribe /cell culture / insect ecology / anterior / protein / silk	127
Pomology	19 phrases (10 words)	pomology / permanent crop / fruit production / rootstock / fruit tree / germination / seed / rootstock	43
Vegetables	14 phrases (7 words)	growth control / development control / environment control / production systems / vegetable cultivation	26
Floriculture	17 phrases (8 words)	regulation of growth / flowering / chemical regulation / ornamental plant / postharvest handling	44
Horticulture Biotechnology	18 phrases (11 words)	horticulture / biotechnology / micropropagation / photomorphogenesis / fresh herbs / allelopathy	37
Postharvest Physiology Technology	28 phrases (19 words)	postharvest / fresh food / quality / antioxidant / storage marketing / packaging / heat shock	122
Totals	131 phrases (218 words)		440

B. Statistical Results of Compiled Corpora

Using the Sketch Engine, based on the seed words, two corpora were quickly compiled after crawling the internet. The statistical information of these two corpora is summarized in Table III, along with that of the BNC. The size of Agriculture Corpus Ver.2 is approximately 12 times larger than that of Ver.1, although the BNC is larger than that of Ver.2 by a factor of 15.

Table III

The Statistical Information of Agriculture Corpus Ver. 1 and Ver.2 and the BNC.

Corpus	Agriculture Corpus Ver. 1	Agriculture Corpus Ver. 2	BNC
Tokens	641,315	8,424,353	112,181,015
Words	513,888	6,583,432	96,048,950
Sentences	27,785	404,117	6,052,184
Paragraphs	7,798	82,698	1,514,906
Documents (or Retrieved URLs)	151	596	4,054

Note. Partly Adapted from Miura (2015).

IV. Analyzing Agriculture Corpora

A. Keyword Analyses

Table IV shows the results of keyword analyses[4] executed by the tool *Word List* using the Sketch Engine. This data represents frequency information that identifies the six most frequent keywords as they appear within "Agriculture Corpus Ver.1" and "Agriculture Corpus Ver.2," and are distinguished from the reference corpus BNC.[5] It was discovered that the senses of these technical words are related to agricultural fields.

[4] Multi-word units of specialized vocabulary are not included in the keyword analyses.

[5] A couple of general terms such as "color," "There" in Ver.1, and abbreviated words related to published journal articles such as

The keywords listed in Ver.2 show over 1,000 raw frequencies, while the most of the keywords that appear in Ver.1 show less than 300. The word "postharvest," which was selected as a seed word for both agricultural corpora, appears in Ver.2. There were no occurrences of this word within the BNC as this term was probably coined after the compilation of the BNC.

Table IV
The Keyword Lists of Agriculture Corpus Ver.1 and Agriculture Corpus Ver. 2 Relative to the Reference Corpus BNC

Lemma	**Raw Frequency (Freq. Per Million.)** Agriculture Corpus Ver.1	BNC	**Lemma**	**Raw Frequency (Freq. Per Million.)** Agriculture Corpus Ver.2	BNC
tumor	279 (435.0)	22 (0.2)	auxin	2,774 (329.3)	7 (0.1)
kinase	766 (1194.4)	272 (2.4)	lactic	3,039 (360.7)	38 (0.3)
Pythium	162 (252.6)	2 (0.0)	fibrin	2,291 (271.9)	32 (0.3)
apoptosis	223 (347.7)	65 (0.6)	cultivar	2,309 (274.1)	40 (0.4)
thaliana	134 (208.9)	2 (0.0)	ethylene	2,251 (267.2)	90 (0.8)
Phytophthora	133 (207.4)	2 (0.0)	postharvest	1,224 (145.3)	0 (0)

B. Technical Vocabulary Search

As previously mentioned, the size of Agricultural Corpus Ver.2 was much larger than the size of Ver.1. This section investigates the degree to which Agriculture Corpus Ver.2 is more informative relative to Ver.1, in terms of the coverage of vocabulary specific to the agricultural field. Vocabulary analyses were conducted using the following two approaches. First, the seed words prepared for Agriculture Corpus Ver.1 were searched in three corpora: (1) Ver.1, (2) the BNC, and (3) Ver.2. Then the occurrences of non-seed words provided by student-written essays were counted in these three corpora.

First, within Agriculture Corpus Ver.1 and the BNC, a total of 136 seed words prepared for Ver.1 were searched, and their raw frequencies and normalized frequencies per million were counted. After this process, only six terms occurred with frequencies that exceeded those in the BNC; this is illustrated in Table V. Then, a search for the same six terms in Table V was conducted using Ver.2, and it was discovered that these terms occurred with a frequency factor of approximately 15 to 102 greater than Ver.1.

Table V
The Distribution of Selected Seed Words across Three Corpora

	Agriculture Corpus Ver.1		**BNC**		**Agriculture Corpus Ver.2**	
	Raw Frequency	Frequency Per Million	Raw Frequency	Frequency Per Million	Raw Frequency	Frequency Per Million
postharvest	154	240.13	0	0	2405	285.48
plant disease	19	29.63	10	0.09	421	49.94
plant pathology	10	15.59	3	0.03	318	37.75
plant nutrition	4	6.24	1	0.01	73	8.67
soil science	3	4.68	1	0.13	128	15.19
plant virus	2	3.12	1	0.01	204	25.64

In the second analysis, the author deliberately chose words and multi-word units that were likely to be specific within the domain of agriculture, all taken from the collections of short essays written by undergraduates of the Department of Agriculture at TUA. The collections consisted of two types of essays written during the 2015 academic year. The first type contained the thirty-one 2nd year students' essays with

"PMID" and "Publ" in Ver.2 were deliberately excluded by the author.

a size ranging from 70 to 100 words written on the topic of their major subject, describing the nature of their major, the types of courses and lectures attended by the students, and their future goal. Figure 2 shows one of the samples; the bold letters are seed words. The second type was composed of essays written by the forty-one 1st year students from the same department. These students were asked to write about their experiences related to growing plants and vegetables, the most difficult part of the agricultural process of caring for the plants, the end results of their experiences, and future plans for growing plants. The length of essays was approximately 200 words. Figure 3 shows an essay written by a 1st year student, and "thinning out" was chosen as a seed word. Both samples shown below contain spelling mistakes but they are shown verbatim.

I am majoring in Agriculture. In agriculture, we study the science and practice of farming. Related areas are food and environmental. I'm taking agricultural production science. I also have **plant pathology** and **crop production** studies. Also I'm going to get training in **genetics and breeding**, which is required for my future job. I hope to be a scientist someday. A scientist is an expert in the area of making new tipe flowers, improvement of flower's **pigment**. In order to become a scientist, it is necessary to study hard. Especially to study English, to go a graduate school and to get a lot of knowledge of **genetics and breeding** is needed. I should stick it out.

Figure 2. An example of essays written by the 2nd year students.

My family experienced growing plants when I was 13 years old. We planed green soy beans and avocado. I was concered about growing plants because my garden didn't get a lot of sunshine. The most difficult things about caring for the plant were pulling out all of the weeds and **thining out**. These were a lot of trouble to do. In the end, both of green soy beans and avocado didn't die. But green soy beans's color was black so we couldn't eat them. Avocado didn't bear fruit. Now, soy beans died but Avocado grows. I wish that avocado bears fruit someday.

The next time we plant a garden, I want to grow flowers. I am member of the department of agriculture. I want to make use of knowledge I leared in class. I am interested in Bonsai. At first I will pull up the weeds in my garden to plant flower.

Figure 3. An example of essays written by the 1st year students.

As is illustrated in Table VI, 55 terms from the 2nd year group and 27 terms from the 1st year group were selected. In this analysis, a total of 56 term searches (excluding seed words for Agriculture Corpus Ver.1) were conducted in Agriculture Corpus Ver.1, the BNC, and Ver.2.

Table VI
The Number and Examples of Selected Terms Extracted from Students' Essays

	Terms from the 2nd Year Students	**Terms from the 1st Year Students**
Total Numbers	55	27
Total Numbers of Seed Words	21	5
Total Numbers of Non-Seed Words	34	22
Examples of Non-Seed Words	self-sufficient / earth science / shearing dietetics / meteorology / biological pesticide preservation / entomology / stability / germ hydroponics / biotechnology / pruning / genetics / conservation / microbial	aquatic plant / thin out / well-bred citron / ill-bred / thermal management air permeability / plow / louse / turf / scorch lily / bulb / cultural knowledge / aphid

Table VII illustrates the top three terms that had the smallest differences between Agriculture Corpus Ver.1 and the BNC in terms of their raw frequencies. Of the 56 terms, only two terms, "aphid," and "hydroponics" experienced larger raw frequencies than the corresponding raw frequencies within the BNC, while the word "pollination" had a slightly lower raw frequency. However, the occurrences contained in Agriculture Corpus Ver.2 were greater than the occurrences within the BNC by a factor ranging from 2 to 30. It is assumed that the raw frequencies of the remaining 53 non-seed words in Ver.2 would exceed the

corresponding frequencies in Ver.1 and the BNC.

Table VII
The Distribution of Selected Non-Seed Words across Three Corpora

	Agriculture Corpus Ver.1		**BNC**		**Agriculture Corpus Ver.2**	
	Raw Frequency	Raw Frequency	Frequency Per Million	Frequency Per Million	Raw Frequency	Frequency Per Million
aphid	194	302.5	158	1.41	323	38.34
pollination	76	118.51	78	0.7	561	66.59
hydroponics	17	26.51	9	0.08	262	31.1

C. Collocates of "Culture" in the Corpora

"Culture" is a polysemy, but can be treated as a technical term in the field of agriculture as well as science. According to Macmillan Dictionary Online (Macmillan Publishers Limited, 2016), in biology, "culture" is defined as "a group of bacteria or other cells that have been grown in a scientific experiment" and "the process by which a group of bacteria or other cells is grown in a scientific experiment," while it is also defined as "in agriculture, the process of growing crops or breeding animals."

From this analysis, any forms of "culture" (including verbs and nouns) were retrieved using the *Concordance* tool of the Sketch Engine, with the intention of identifying the nature of collocates that tend to co-occur with "culture" across three corpora. First, it was discovered that there were 119 (185.56 per million) occurrences of "culture" contained in Agriculture Corpus Ver.1, and 10,281 (91.60 per million) contained in the BNC, and 9,333 (1,107.86 per million) contained in Agriculture Corpus Ver.2.

Based on values derived from logDice, Tables VIII and IX identify collocates that precede "culture" with the highest frequency as well as collocates that immediately follow "culture" in the three corpora.

In the two agricultural corpora, all of the words that occur adjacent to "culture" were used in agricultural or biological contexts; for example, "hydroponic culture," "container culture," "cell culture," "tissue culture," "suspension culture" and others. However, with the exception of "tissue culture," the BNC contained frequent combinations such as "popular culture," "youth culture," and "western culture," which were used in a more general sense, or "activities involving music, literature, and other arts" (Macmillan Publishers Limited, 2016).

It seems that "culture system(s)," "culture conditions," and "culture medium(a)" were the typical co-occurrences across the three corpora. Compared to the preceding collocates, the BNC included scientific combinations such as "culture supernatants" and "culture dish," neither of which occurred in the two agricultural corpora, although "culture shock" and "culture Club" occurred as general combinations.

It was discovered that the collocational behaviors with both positions in the two agricultural corpora indicated that the term "culture" was genre-specific, and Agriculture Corpus Ver.2 contained a greater frequency of "culture" than the frequency of occurrence in Ver.1 and the BNC.

Table VIII
The Frequent Collocates Immediately Preceding "Culture" in the Corpora

Agriculture Corpus Ver.1			**BNC**			**Agriculture Corpus Ver.2**		
Left	Freq.	logDice	Left	Freq.	logDice	Left	Freq.	logDice
hydroponic	9	10.72	popular	169	8.13	tissue	1566	31
container	3	9.37	youth	85	7.62	cell	559	18.43
cell	5	7.25	tissue	62	7.37	suspension	175	9.16
N/A			western	83	7.37	vitro	155	8.8
N/A			dominant	60	7.21	broth	100	8.37
N/A			political	159	7.08	pure	98	8.329

Table IX
The Frequent Collocates Immediately Following "Culture" in the Corpora

Agriculture Corpus Ver.1			**BNC**			**Agriculture Corpus Ver.2**		
Right	Freq.	logDice	Right	Freq.	logDice	Right	Freq.	logDice
systems	8	9.67	medium	37	6.484	medium	222	9.219
conditions	3	8.02	shock	39	6.476	system	295	9.318
system	3	7.42	supernatants	12	5.251	T	155	8.795
N/A			Club	18	5.22	media	153	8.745
N/A			*	14	5.03	systems	131	8.343
N/A			dish	11	4.93	conditions	106	7.98

V. Summary

The Sketch Engine is a powerful and innovative tool for ESP practitioners in any domain, especially those who are not familiar with the target field. WebBootCaT, a tool in the Sketch Engine, offers effortless procedures of compilation to support domain-specific corpora. However, it was suggested that the selection of seed words was very effective not only in terms of the entire size of the corpora, but also in terms of the coverage of the genre-specific vocabulary, according to the analyses on vocabulary search and collocational behaviors. Compilers should collect seed words that come from narrower domains, as Agriculture Corpus Ver.2 was more informative and larger than Agriculture Corpus Ver.1. In the future, Agriculture Corpus Ver.2 can be combined with several domain-specific corpora that are based on seed words collected from the Department of Animal Science and others, so that it is possible to build a more balanced and larger corpus on agriculture containing narrower domain-or-genre specific subcorpora. This kind of corpus can be useful not only for teaching English to agricultural undergraduates, but can also serve as a database for use by postgraduate students as well as academic researchers in the target field, in presenting their research papers in scientific journals and conferences.

References

Chujo, K., & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistic measures. *System 34*, 255-269.

Kilgarrif, A., Rychly, R., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. *Proceedings of Euralex, Lorient, France, July 2004,* 105-116.

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

Koester, A. (2010). Building small specialised corpora. In A. O'Keefe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66-79). New York: Routledge.

Lexical Computing Ltd. (2014). *Statistics used in the Sketch Engine.* Retrieved from https://www.sketchengine.co.uk/documentation/raw-attachment/wiki/SkE/DocsIndex/ske-stat.pdf

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* Oxon: Routledge.

Macmillan Publishers Limited. (2006). *Definition and synonyms of culture from the online English dictionary*. Retrieved from http://www.macmillandictionary.com/dictionary/british/culture_1

Miura, A. (2015). Building a domain-specific corpus of agriculture and applying it in the classroom. *Annual Report of JACET SIG on ESP 17*, 25-29.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Tribble, C. (2002). Corpora and corpus Analysis: New windows on academic writing. In J. Flowerdew (Ed.), *Academic Discourse* (pp. 131-149). London: Longman.