TOKYO UNIVERSITY
OF AGRICULTURE 1891

COMPILING A DOMAIN SPECIFIC CORPORA WITH THE SKETCH ENGINE

Aika Miura
東京農業大学
TOKYO UNIVERSITY OF AGRICULTURE

19/08/2016

The Faculty of Agriculture, Tokyo University of Agriculture

- ❑ **Department of Agriculture**
  - Crop Science, Genetics and Plant Breeding, Plant Pathology, Entomology, Pomology, Vegetables, Floriculture, Horticulture-Biotechnology, Postharvest Physiology and Technology
- ❑ **Department of Animal Science**
  - Animal Reproduction, Animal Genetics and Breeding, Animal Physiology, Animal Feeding, Animal Product Processing, Animal Health, Livestock Farming Management
- ❑ **Department of Human and Animal-Plant Relationships**
  - Plant Conservation, People-Plant Relationships, Wild Animals, Companion Animals, Plant Assisted Therapy, Animal Facilitated Therapy

1. The Purpose of the Study

- To introduce domain-specific corpora using the *Sketch Engine*
- Describing and comparing *'Agriculture Corpus (ver.1)'* (Miura, 2015) and *'Agriculture Corpus 2016 (ver.2)'*, in terms of size, keyness, and lexical behaviours of the genre-specific vocabulary

✓The importance of selecting **"seed-words"** to compile corpora

✓Referring to a balanced-corpus, the British National Corpus, and a mega-corpus, SEKAI Corpus (will be officially released by Shogakkan on 28 August 2016) .

2. The Sketch Engine

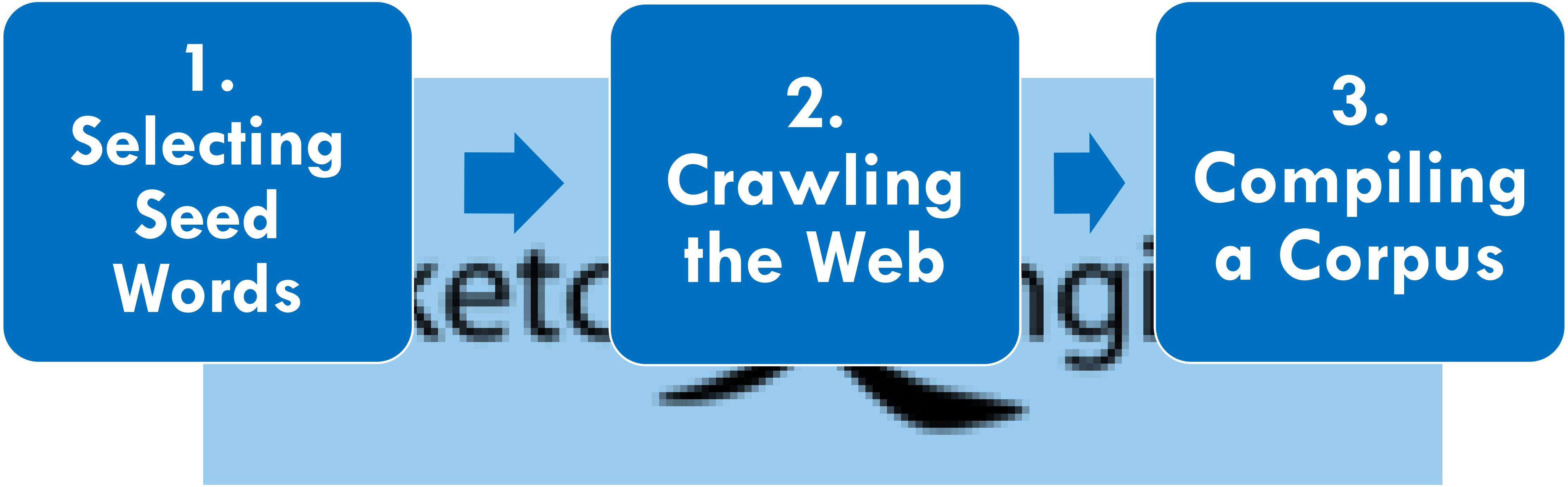
Sketch 文 Engine

- ❑ **A commercial interface that contains a built-in corpus query system – Lexical Computing Ltd.**
  (Kilgarriff, Baisa, Bušta, Jakubíček, Kovář, Michelfeit, Rychlý, & Suchomel, 2014)

- **Ready-made corpora available**
  e.g.) enTenTen, ukWaC, jpWaC, the British National Corpus (BNC)

- **Allowing users to create original corpora with a simple and effortless procedure - *WebBootCat***

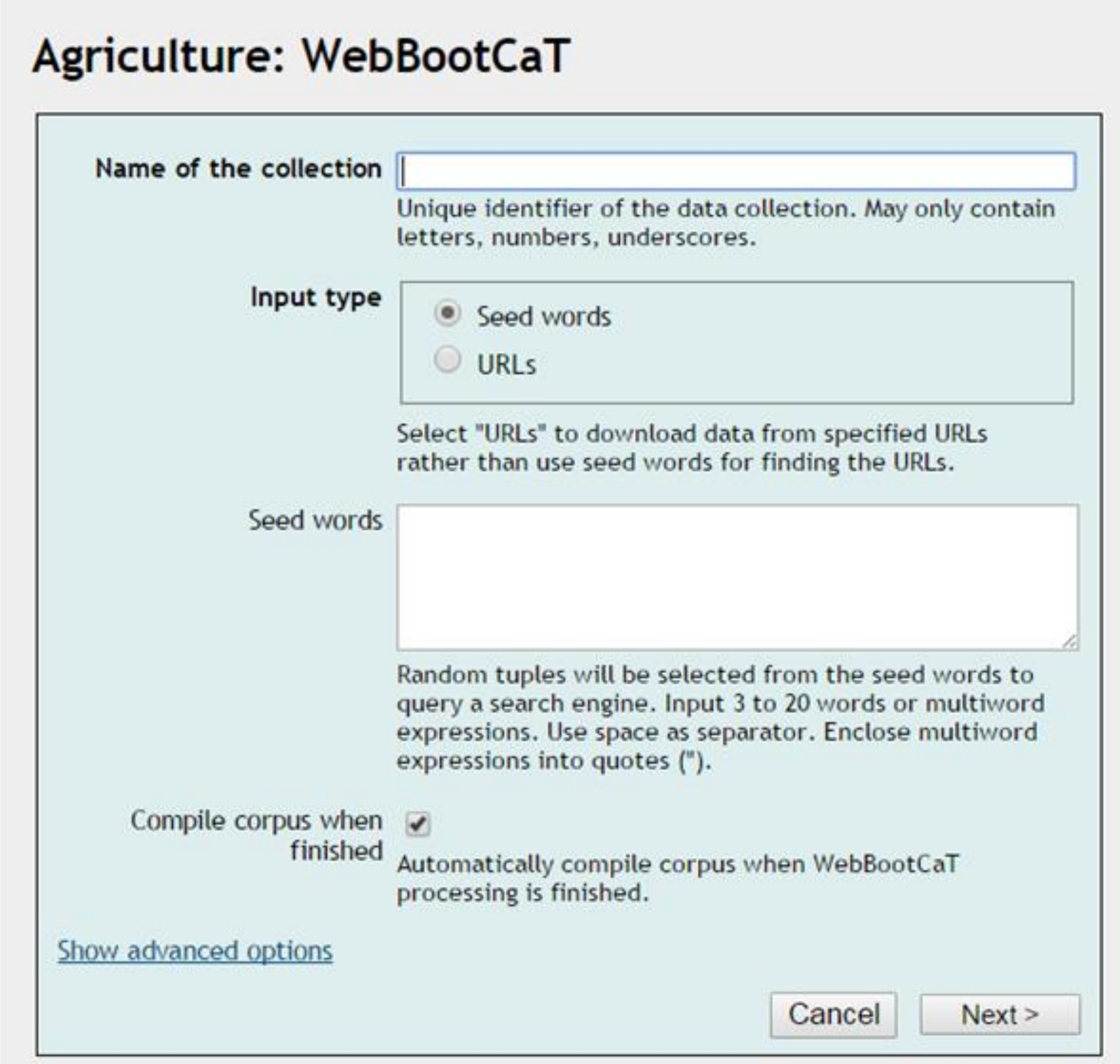
2.1 The Process of Compiling a Domain Specific Corpus With WebBootCat



1. Selecting Seed Words

2. Crawling the Web

3. Compiling a Corpus



Agriculture: WebBootCaT

Name of the collection

Unique identifier of the data collection. May only contain letters, numbers, underscores.

Input type

● Seed words

○ URLs

Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs.

Seed words

Random tuples will be selected from the seed words to query a search engine. Input 3 to 20 words or multiword expressions. Use space as separator. Enclose multiword expressions into quotes (").

Compile corpus when finished ☑

Automatically compile corpus when WebBootCaT processing is finished.

2.2 Seed Words for Crawling the URLs

- *Agriculture (Ver.1) - Trial Version*

  - Miscellaneous

- *Agriculture 2016 (Ver.2) -*

  – Focusing on key words given from students and academics at the Department of Agriculture (excluding the Departments of Animal Science and Human and Animal-Plant Relationships)

2.2.1 Seed Words for *Agriculture (Ver.1)*
136 PHRASES CONTAINING 248 DISTINCT WORDS

Entrance Exam Papers

for Undergraduate Programmes

- **72 phrases containing 80 words**
- species / plant / animal / organism / habitat / breed / variation / physical diversity / biologist / interbreed /

Materials Introduced to "Science English"

for Graduate Applicants

- **54 phrases containing 67 words**
- cloning / ethical / molecular / feasible / cell / foetus / trunk / bud culture / genetically identical etc.

Specific Fields of Academic Staff at Dept. of Agriculture

in the University Prospectus

- **10 phrases 101 words**
- biomass crop production / green manure crops / soil physical condition improvement / compost utilization etc.

2.2.2 Seed Words for *Agriculture 2016 (Ver.2)*

131 PHRASES CONTAINING 216 DISTINCT WORDS

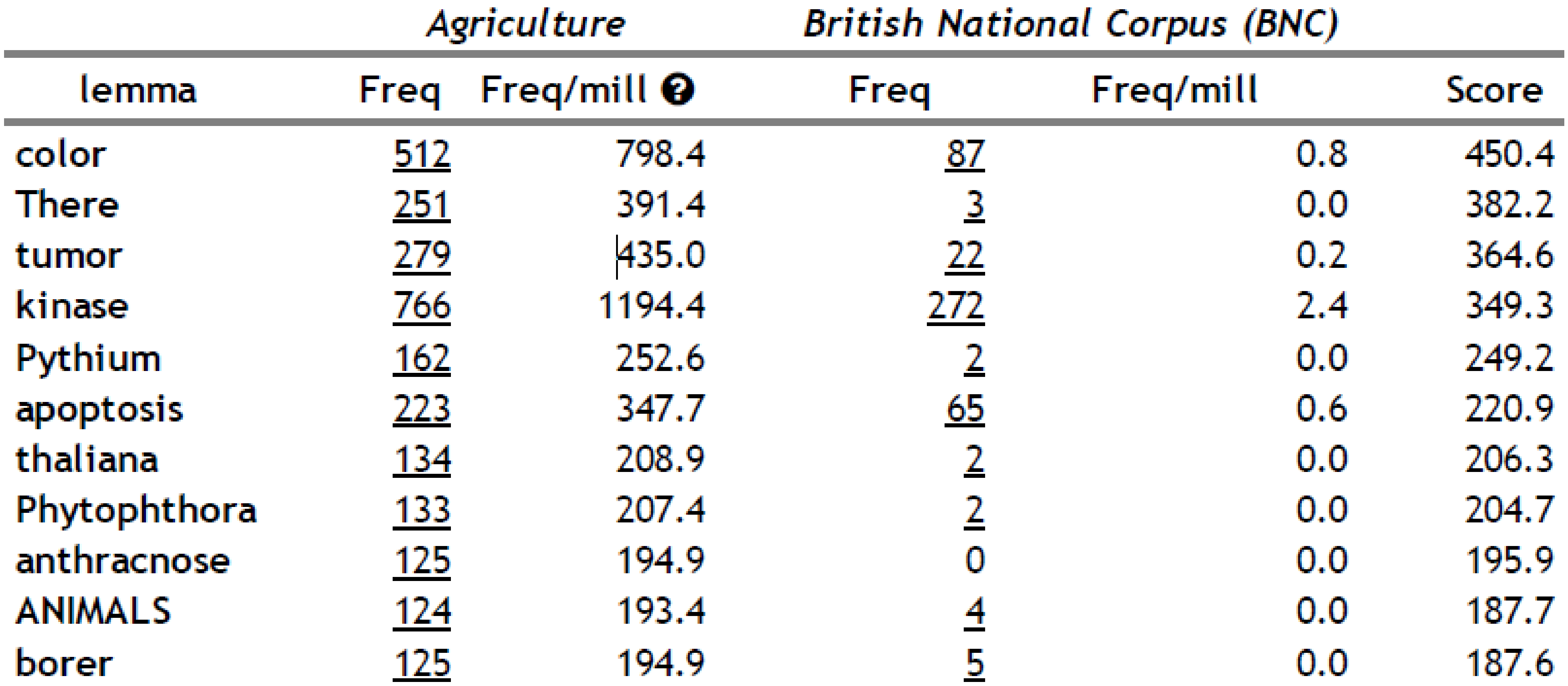
Research Laboratories	Examples of Seed Words
Crop Science	compositing regulation / global climate change, etc.
Genetics and Plant Breeding	chromosome / transposon / genome transformation / tissue culture, etc.
Plant Pathology	phytopathology / plant diseases / biocontrol /plant virus, etc.
Entomology	insect taxonomy / morphology, genus / tribe /cell culture, etc.
Pomology	pomology / permanent crop / fruit production / rootstock, etc.
Vegetables	growth control / development control / environment control, etc.
Floriculture	regulation of growth / flowering / chemical regulation, etc.
Horticulture Biotechnology	horticulture / biotechnology / micropropagation / photomorphogenesis, etc.
Postharvest Physiology Technology	postharvest / fresh food / quality / antioxidant / storage / marketing, etc.

3. RESULTS

3.1 The Statistical Information of Corpora

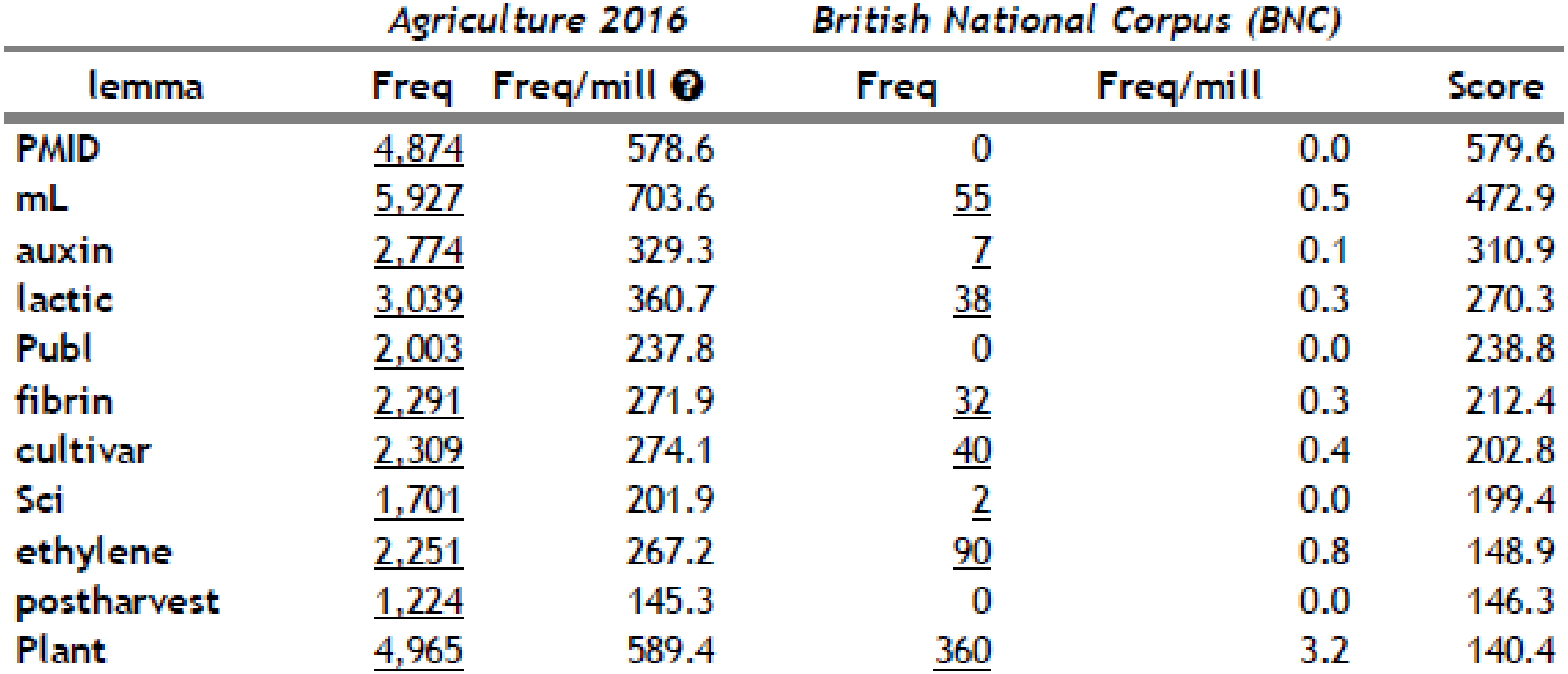
	Agriculture Corpus (ver.1)	*Agriculture Corpus 2016 (ver.2)*	BNC	SEKAI Corpus (incl. BNC)
Documents (Retrieved URLs)	151	596	4,054	574,621
Tokens	641,315	8,424,353	112,181,015	N/A
Words	513,888	6,583,432	96,048,950	2,191,836,612
Sentences	27,785	404,117	6,052,184	N/A
Paragraphs	7,798	82,698	1,514,906	N/A

3.2.1 Key Word Analysis: *Agriculture (ver.1)* vs. BNC



	Agriculture		*British National Corpus (BNC)*		
lemma	Freq	Freq/mill ?	Freq	Freq/mill	Score
color	512	798.4	87	0.8	450.4
There	251	391.4	3	0.0	382.2
tumor	279	435.0	22	0.2	364.6
kinase	766	1194.4	272	2.4	349.3
Pythium	162	252.6	2	0.0	249.2
apoptosis	223	347.7	65	0.6	220.9
thaliana	134	208.9	2	0.0	206.3
Phytophthora	133	207.4	2	0.0	204.7
anthracnose	125	194.9	0	0.0	195.9
ANIMALS	124	193.4	4	0.0	187.7
borer	125	194.9	5	0.0	187.6

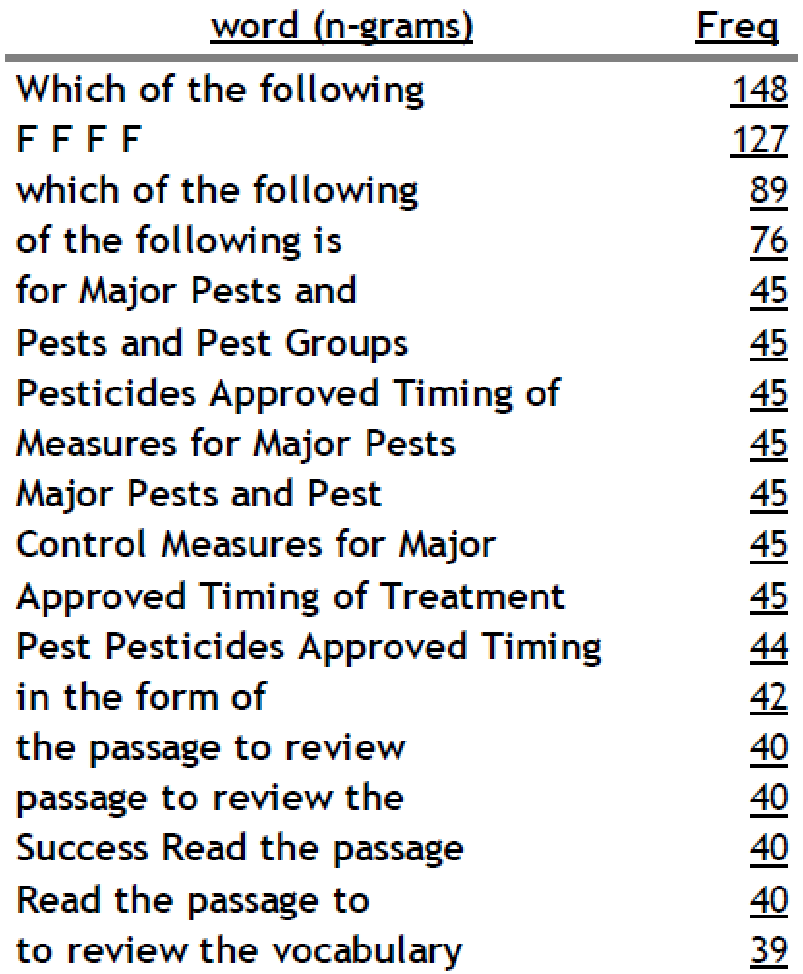
3.2.2 Key Word Analysis: *Agriculture 2016 (ver.2)* vs. BNC



	Agriculture 2016		*British National Corpus (BNC)*		
lemma	Freq	Freq/mill ?	Freq	Freq/mill	Score
PMID	4,874	578.6	0	0.0	579.6
mL	5,927	703.6	55	0.5	472.9
auxin	2,774	329.3	7	0.1	310.9
lactic	3,039	360.7	38	0.3	270.3
Publ	2,003	237.8	0	0.0	238.8
fibrin	2,291	271.9	32	0.3	212.4
cultivar	2,309	274.1	40	0.4	202.8
Sci	1,701	201.9	2	0.0	199.4
ethylene	2,251	267.2	90	0.8	148.9
postharvest	1,224	145.3	0	0.0	146.3
Plant	4,965	589.4	360	3.2	140.4

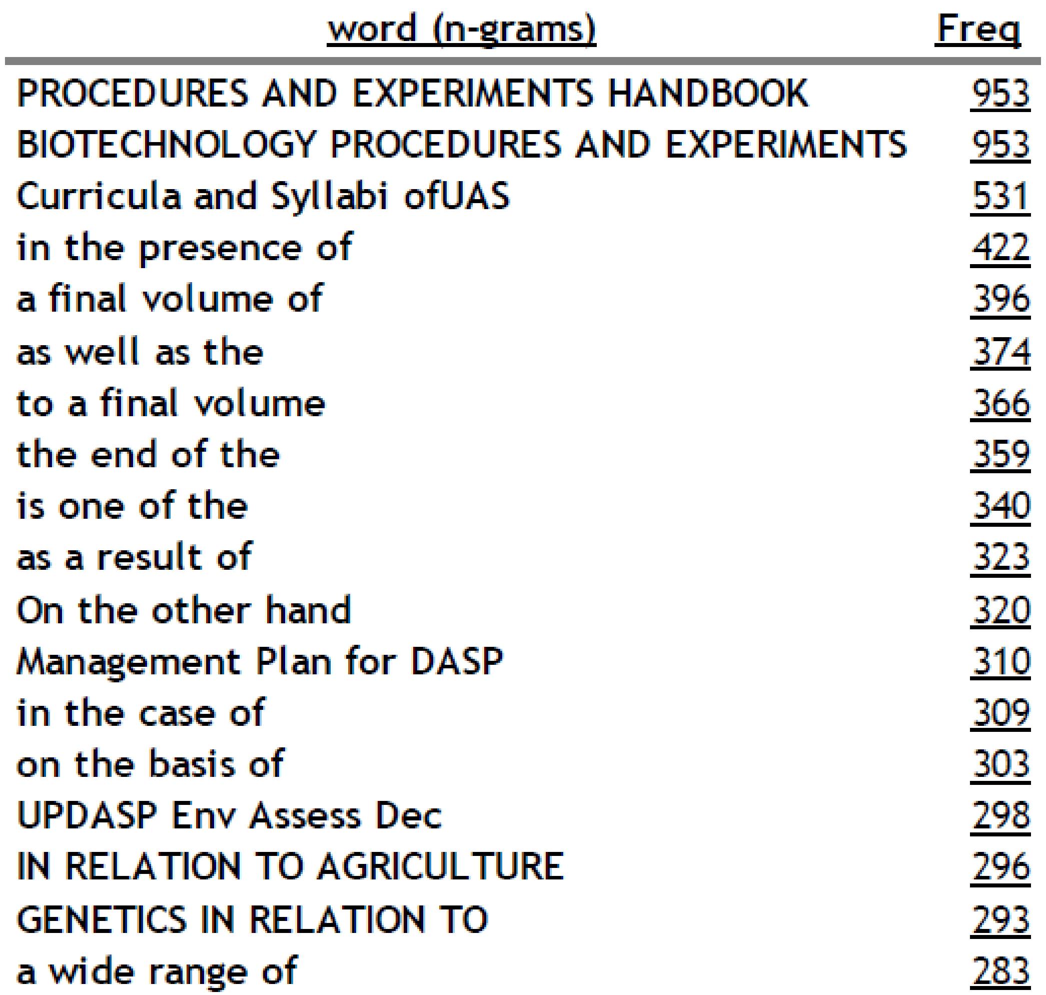
3.3 Collocations: Frequent 4-Grams

Agriculture (ver.1)



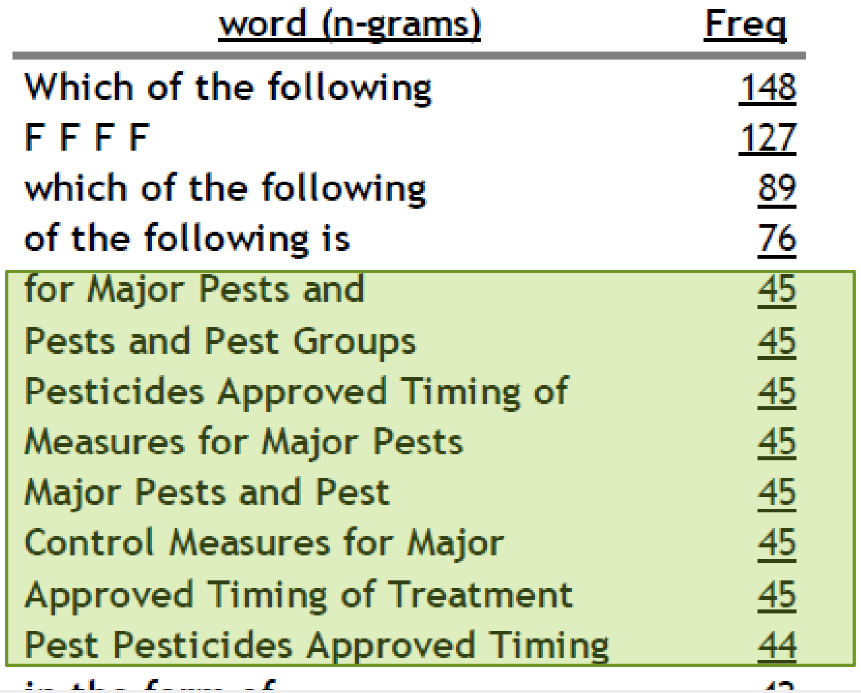
word (n-grams)	Freq
Which of the following	148
F F F F	127
which of the following	89
of the following is	76
for Major Pests and	45
Pests and Pest Groups	45
Pesticides Approved Timing of	45
Measures for Major Pests	45
Major Pests and Pest	45
Control Measures for Major	45
Approved Timing of Treatment	45
Pest Pesticides Approved Timing	44
in the form of	42
the passage to review	40
passage to review the	40
Success Read the passage	40
Read the passage to	40
to review the vocabulary	39

Agriculture 2016 (ver.2)

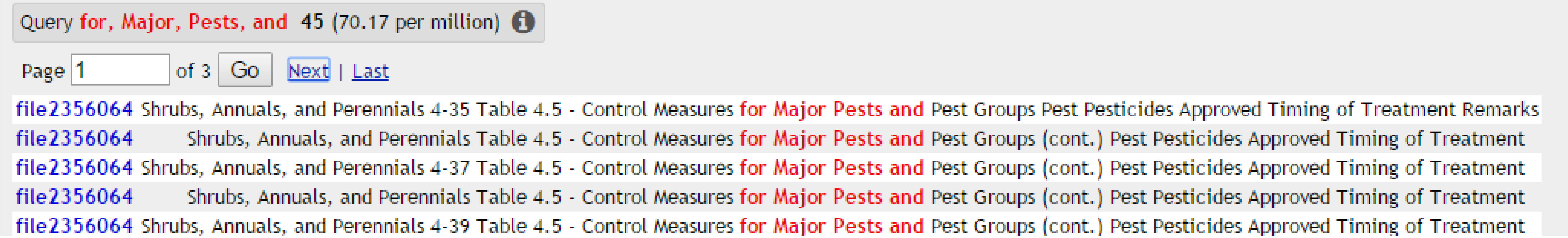


word (n-grams)	Freq
PROCEDURES AND EXPERIMENTS HANDBOOK	953
BIOTECHNOLOGY PROCEDURES AND EXPERIMENTS	953
Curricula and Syllabi ofUAS	531
in the presence of	422
a final volume of	396
as well as the	374
to a final volume	366
the end of the	359
is one of the	340
as a result of	323
On the other hand	320
Management Plan for DASP	310
in the case of	309
on the basis of	303
UPDASP Env Assess Dec	298
IN RELATION TO AGRICULTURE	296
GENETICS IN RELATION TO	293
a wide range of	283

3.3.1 Collocations: Frequent 4-Grams in *Agriculture (ver.1)*



word (n-grams)	Freq
Which of the following	148
F F F F	127
which of the following	89
of the following is	76
for Major Pests and	45
Pests and Pest Groups	45
Pesticides Approved Timing of	45
Measures for Major Pests	45
Major Pests and Pest	45
Control Measures for Major	45
Approved Timing of Treatment	45
Pest Pesticides Approved Timing	44
in the form of	43



Query for, Major, Pests, and 45 (70.17 per million)

Page 1 of 3 Go Next | Last

file2356064 Shrubs, Annuals, and Perennials 4-35 Table 4.5 - Control Measures for Major Pests and Pest Groups Pest Pesticides Approved Timing of Treatment Remarks

file2356064 Shrubs, Annuals, and Perennials Table 4.5 - Control Measures for Major Pests and Pest Groups (cont.) Pest Pesticides Approved Timing of Treatment

file2356064 Shrubs, Annuals, and Perennials 4-37 Table 4.5 - Control Measures for Major Pests and Pest Groups (cont.) Pest Pesticides Approved Timing of Treatment

file2356064 Shrubs, Annuals, and Perennials Table 4.5 - Control Measures for Major Pests and Pest Groups (cont.) Pest Pesticides Approved Timing of Treatment

file2356064 Shrubs, Annuals, and Perennials 4-39 Table 4.5 - Control Measures for Major Pests and Pest Groups (cont.) Pest Pesticides Approved Timing of Treatment

4. Analysis (1): Searching Agricultural Vocabulary in the Corpora

4.1 Selecting Vocabulary from Students' Essays (Written in 2015)

Department	Year (No. of students)	Couse	Topic	No. of Words	No. of Selected Words
Agriculture	1st Year Students (41)	Reading Course	My Family's Garden	150 – 200 words	27 Words
Agriculture	2nd Year Students (31)	Writing Course	My Own Major Subject	70 – 100 words	55 Words

4.1.1 An Example Essay of 1st Year Student (152 words)

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.

4.1.2 An Example Essay of 2nd Year Student (115 words)

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.

4.2.1 The Frequencies of Some Words Selected from the Students' Essays in Four Corpora

	Agriculture (ver.1)		BNC		SEKAI Corpus	
	Raw Freq.	Freq. Per Million	Raw Freq.	Freq. Per Million	Raw Freq.	Freq. Per Million
aphid	194	302.5	158	1.41	316	0.23
pollination	76	118.51	78	0.7	466	0.28
horticulture	43	67.05	129	1.15	1155	0.53
hydroponics	17	26.51	9	0.08	86	0.04
entomology	7	10.92	23	0.21	410	0.19
plow	2	3.12	15	0.13	1366	0.57

4.2.2 The Frequencies of Some Words Selected from the Students' Essays in Four Corpora

	Agriculture (ver.1)		*Agriculture 2016 (ver.2)*		BNC		SEKAI Corpus	
	Raw Freq.	Freq. Per Million	**Raw Freq.**	**Freq. Per Million**	Raw Freq.	Freq. Per Million	Raw Freq.	Freq. Per Million
aphid	194	302.5	**323**	**38.34**	158	1.41	316	0.23
pollination	76	118.51	**561**	**66.59**	78	0.7	466	0.28
horticulture	43	67.05	**2564**	**304.36**	129	1.15	1155	0.53
hydroponics	17	26.51	**262**	**31.1**	9	0.08	86	0.04
entomology	7	10.92	**299**	**35.49**	23	0.21	410	0.19
plow	2	3.12	**52**	**6.17**	15	0.13	1366	0.57

4.3.1 The Frequencies of Some Seed-Words for Agriculture (ver.1) in Four Corpora

	Agriculture (ver.1)		BNC		SEKAI Corpus	
	Raw Freq.	Freq. Per Million	Raw Freq.	Freq. Per Million	Raw Freq.	Freq. Per Million
postharvest	154	240.13	0	0	195	0.2
insecticide	144	224.54	181	1.61	1138	0.87
horticultural	59	92	308	2.75	1644	0.86
plant disease	19	29.63	10	0.09	116	0.06
plant pathology	10	15.59	3	0.03	145	0.06
plant nutrition	4	6.24	1	0.01	48	0.03

4.3.2 The Frequencies of Some Seed-Words for Agriculture (ver.1) in Four Corpora

	Agriculture (ver.1)		*Agriculture 2016 (ver.2)*		BNC		SEKAI Corpus	
	Raw Freq.	Freq. Per Million	**Raw Freq.**	**Freq. Per Million**	Raw Freq.	Freq. Per Million	Raw Freq.	Freq. Per Million
postharvest	154	240.13	**2405**	**285.48**	0	0	195	0.2
insecticide	144	224.54	**412**	**48.91**	181	1.61	1138	0.87
horticultural	59	92	**2144**	**254.5**	308	2.75	1644	0.86
plant disease	19	29.63	**421**	**49.94**	10	0.09	116	0.06
plant pathology	10	15.59	**318**	**37.75**	3	0.03	145	0.06
plant nutrition	4	6.24	**73**	**8.67**	1	0.01	48	0.03

5. Analysis (2):Collocates of ‘culture’ (incl. noun and verb)

5.1 Examples of Concordance Lines of ‘Culture’ from *Agriculture (ver.1)*

Plant (栽培)	The Introduction provides background information about tomato botany and ***culture***, seed production and quality assurance, and container production of transplants.
Cell (培養)	The main tool used to diagnose infectious diseases is microbiological ***culture*.**
Animal (養殖)	There are a variety of techniques for growing mussels. Bouchot ***culture*:** Intertidal growth technique, or bouchot technique: pilings, known in French as bouchots, are planted at sea; ropes, on which the mussels grow, are tied in a spiral on the pilings; some mesh netting prevents the mussels from falling away.

5.2.1 Frequent Collocates Preceding “Culture” in Corpora (X + “CULTURE”)

Agriculture (ver.1)			BNC			SEKAI Corpus (raw freq. based)		
Left	Freq.	logD.	Left	Freq.	logD.	Left	Freq.	%
hydroponic	9	10.72	popular	169	8.13	political	113	3.84
container	3	9.37	youth	85	7.62	popular	79	2.69
cell	5	7.25	tissue	62	7.37	corporate	45	1.53
N/A			western	83	7.37	american	35	1.19
N/A			dominant	60	7.21	cell	33	1.12
N/A			political	159	7.08	organizational	24	0.82

5.2.2 Frequent Collocates Preceding “Culture” in Corpora (X + “CULTURE”)

Agriculture (ver.1)			*Agriculture 2016 (ver.2)*			BNC			SEKAI Corpus (raw freq. based)		
Left	Freq.	logD.	Left	Freq.	logD.	Left	Freq.	logD.	Left	Freq.	%
hydroponic	9	10.72	**tissue**	**1566**	**31.0**	popular	169	8.13	political	113	3.84
container	3	9.37	**cell**	**559**	**18.43**	youth	85	7.62	popular	79	2.69
cell	5	7.25	**suspension**	**175**	**9.16**	tissue	62	7.37	corporate	45	1.53
N/A			**vitro**	**155**	**8.80**	western	83	7.37	american	35	1.19
N/A			**broth**	**100**	**8.37**	dominant	60	7.21	cell	33	1.12
N/A			**pure**	**98**	**8.329**	political	159	7.08	organizational	24	0.82

5.3.1 Frequent Collocates Following “Culture” in Corpora (“CULTURE” + X)

Agriculture (ver.1)			BNC			SEKAI Corpus (raw freq. based)		
Right	Freq.	logD.	Right	Freq.	logD.	Right	Freq.	%
systems	8	9.67	medium	37	6.484	medium	46	1.53
conditions	3	8.02	shock	39	6.476	condition	13	0.43
system	3	7.42	supernatants	12	5.251	supernatant	13	0.43
N/A			Club	18	5.22	war	12	0.4
N/A			*	14	5.03	collection	10	0.33
N/A			dish	11	4.93	dish	10	0.33

5.3.2 Frequent Collocates Following "Culture" in Corpora ("CULTURE" + X)

Agriculture (ver.1)			*Agriculture 2016 (ver.2)*			BNC			SEKAI Corpus (raw freq. based)		
Right	Freq.	logD.	Right	Freq.	logD.	Right	Freq.	logD.	Right	Freq.	%
systems	8	9.67	**medium**	**222**	**9.219**	medium	37	6.484	medium	46	1.53
conditions	3	8.02	**system**	**295**	**9.318**	shock	39	6.476	condition	13	0.43
system	3	7.42	**T**	**155**	**8.795**	supernatants	12	5.251	supernatant	13	0.43
N/A			**media**	**153**	**8.745**	Club	18	5.22	war	12	0.4
N/A			**systems**	**131**	**8.343**	*	14	5.03	collection	10	0.33
N/A			**conditions**	**106**	**7.98**	dish	11	4.93	dish	10	0.33

5.4.1 Extracts Including Frequent Collocates Following 'Culture' from *Agriculture 2016 (ver.2)*

tissue culture (Seed word for GENETICS and PLANT BREEDING)	LA included in plant ***tissue culture*** media significantly lowered tissue browning and improved transformation efficiency of wheat, soybean and cotton (Dan et al., 2009).
cell culture (Seed word for ENTOMOLOGY)	The expression of recombinant proteins in larvae has the advantage of its low cost in comparison with insect ***cell cultures***.

5.4.2 Extracts Including Frequent Collocates Preceding ‘Culture’ from *Agriculture 2016 (ver.2)*

culture medium	Single cell cultures, plant cell without cell wall (Protoplast), tissues of leaves, or roots can be used to generate plants on ***culture media*** given the required nutrients and Growth regulators. (Seed word for POMOLOGY)
culture system	A novel ***culture system*** suitable for practical application in micropropagation has been developed. (Seed word for HORTICULTURE BIOTECHNOLOGY)

6. SUMMARY

- Domain-specific corpora compiled by the Sketch Engine should be useful for EST practitioners who lack knowledge in the target field.
- Various analyses (keyness, collocations, specific vocabulary search) on the originally made domain-specific corpora were more informative than the BNC and SEKAI Corpus, in terms of retrieving target vocabulary in agricultural contexts.
- Careful selection of seed words is prerequisite for compiling more informative and balanced domain-specific corpora.

References

- 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

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

- Shogakkan (2016). *21okugo no daikbo senmonnbunnya web corpus. [2,100,000,000-word mega specific web corpus].* Retrieved from http://scn.jkn21.com/information/20160322_corpus.pdf

Aika Miura

Tokyo University of Agriculture

E-mail: dawn1110am@gmail.com