

Corpus-Assisted Editing: Corpora and Tools for Thesis and Dissertation Writers

Maggie Charles
Language Centre
University of Oxford
maggie.charles@lang.ox.ac.uk



Outline

1

- Context and rationale

2

- Building do-it-yourself (DIY) corpora

3

- Tools and examples

4

- Evaluation of the approach

5

- To conclude

Part 1

Context and Rationale

What is Corpus-assisted editing?

- **A corpus** is a collection of electronic texts built according to set criteria and constructed for a specific purpose
- **Corpus-assisted editing** is the use of corpora to edit and revise texts.
- The corpora used here are **DIY corpora** compiled by writers for their own use.
- Used for teaching **translation** (e.g. Kübler 2011), **linguistics** (e.g. Seidlhofer 2000) exploring **disciplinary discourse** (e.g. Charles 2015a, 2015b, 2017; Lee & Swales 2006)

Editing your Thesis with Corpora:

Course Details

Aim:	to improve graduates' editing skills to provide a resource for future use
Frequency:	2-3 times per year (10 in total)
Timing:	One 2-hour session/week for 6 weeks
Venue:	computer laboratory
Class size:	maximum 12
Composition:	multi-disciplinary
Software:	AntConc (Anthony 2014) AntFileConverter (Anthony 2015)

Participants

Doctoral students who have completed at least **1 substantial chapter** of their thesis

66 students (2012 – 2015)

Fields

Natural Science 41%

Social Science 30%

Humanities 29%

Two Types of DIY Corpora

1. DIY Corpus of Research Articles in student's own field/topic area

- based on downloaded files in own bibliography
- may include subcorpora of different topics/genres

2. DIY Corpus of Student's Own Writing

- chapters of thesis as individual files
- may include subcorpora of other writing (e.g. proposals, Master's dissertation)

Course Programme

Topic	Tool
1. Using concordances to answer grammar, vocabulary and usage queries	AntConc Concordance
2. Building your corpus of research articles; answering your own editing queries	AntFileConverter
3. Finding collocations and semi-fixed phrases; building a corpus of your own writing	Clusters Collocates
4. Examining the words you use; checking for consistency; comparing your own writing with expert texts	Word List N-Grams
5. Tracing content, themes, terminology, citation throughout your own writing	Concordance Plot
6. Comparing individual chapters to the whole text; comparing your own writing with expert texts	Keyword List

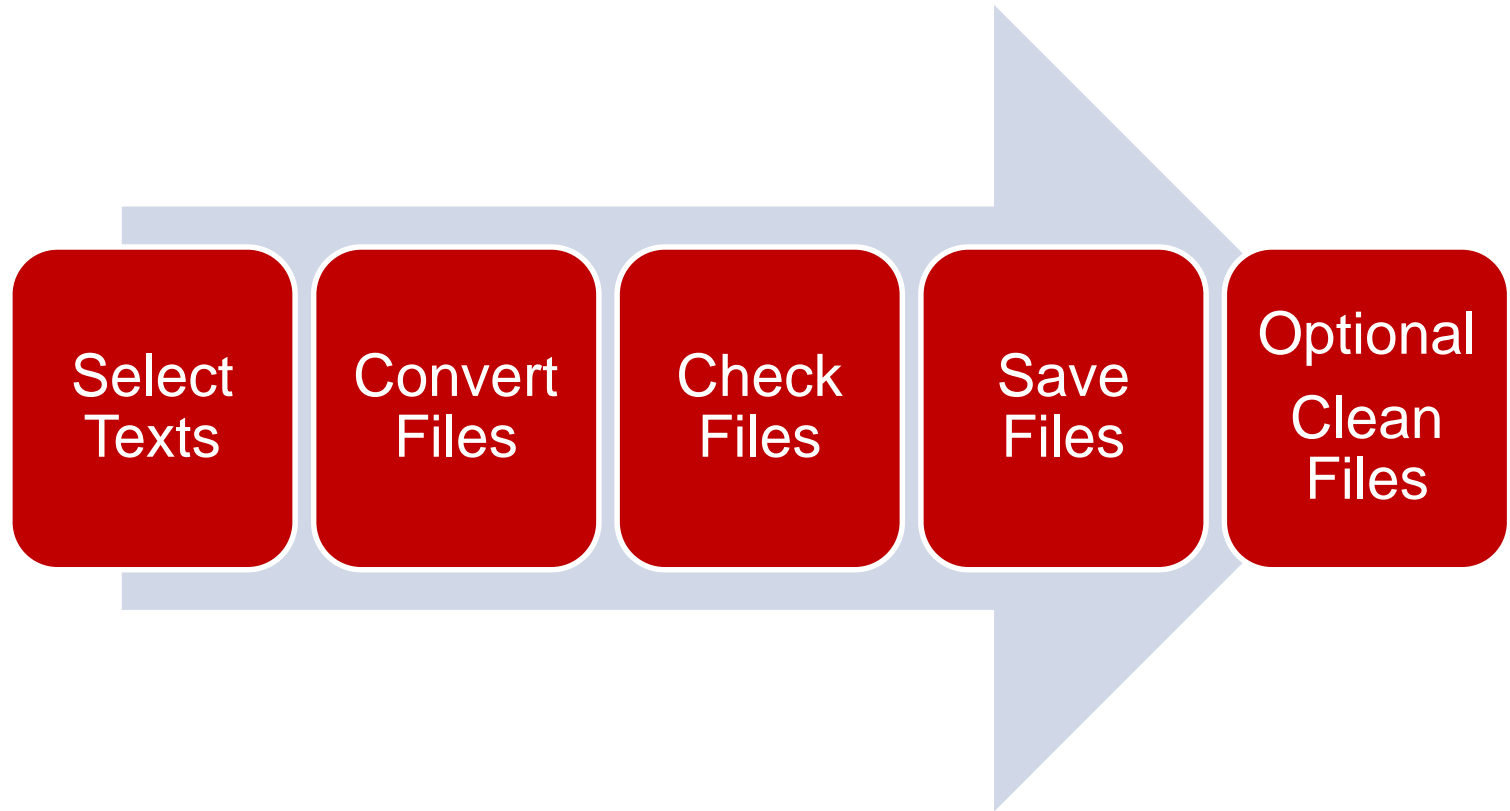
that they exemplify an acceptable construction of stance in their field. Two corpora of theses are examined in two contrasting disciplines, a social science: politics and a natural science: materials. Section 1.7 provides details of the methods used to examine them. The chapter ends with an outline of the structure of the book. 1.4 Justification for a Corpus-Based Approach Corpora have been defined as 'collections of texts (or parts of text) that are stored in a digital format'. This implies that they are instances of naturally occurring texts. This study uses two corpora such as those in this study tend to be relatively small, even such mini-corpora provide large amounts of data in comparison to what would be available through other means. Thus their occurrence and form are not solely due to individual personal choice, but, particularly in specialised corpora, may be characteristic of a discourse community (Stubbs, 2001b: 215). This means that the significance of the frequency of a given pattern can best be seen in comparison with other corpora. The significance of the frequency of a given pattern can best be seen in comparison with other corpora from two different disciplines are compared so that the patterns observed in each corpus can be compared. 1.6.4 The Thesis Writer's Stance towards the Disciplinary Community We have examined two corpora of theses in politics and materials to examine certain grammatical features associated with the expression of stance. The aim is to compare the corpora in terms of the frequency and type of stance constructed by these features and to see if there are any differences between the two corpora in terms of the frequency, operation and type of stance constructed by each group. 1.8 Corpora and Methods 1.8.1 Setting up the Oxford Academic Text Corpora

Part 2

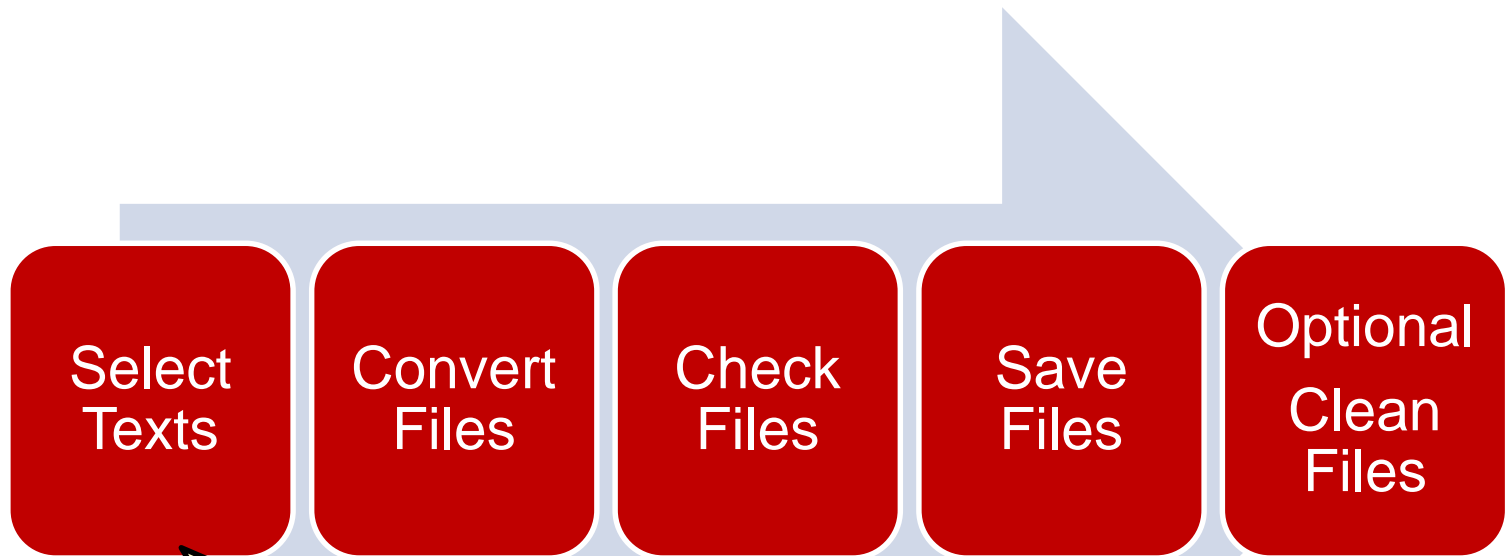
Building Do-It-Yourself (DIY) Corpora

total of 3030 instances, 1998.3 per 100,000 words. 43 adverbs occur in both corpora, 30% of the overall total, indicating a considerable overlap between politics and materials. Each group is named accordingly, primarily, significantly, so, together, unfortunately. However in both corpora, the majority of adverbs are grouped, 88.3% in politics and 92.5% in materials. A similar process occurs with really. About 80% of instances in both corpora occur with negatives, questions, only, if or other markers of doubt. In such cases the difference between the two corpora is again in the expected direction. Because of the recursive and value-laden nature of stance, it is used in both corpora to structure the argument, although more frequently in politics than materials. Both corpora make use of substantially the same textual adverbs; the only exception is too. As shown by the figures for tokens per 100,000 words: 1010.5 in politics compared to 1010.5 in materials. These adverbs account for around 40% of all the occurrences of grouped adverbs. Further, the grammatical marking of stance provides 'an attitudinal or evaluative stance'. corpora contain examples of other introductory it patterns with adjectives, for example the LIKELY group (see section 4.6.5). However, as already noted, corpora, particularly in the choice of adjective. Indeed that study suggested that materials corpora, using it as the search word followed by 2-6 words of context before either the to-infinitive is considerably more common than the that-clause pattern. The corpora, about a third of the total in the two corpora taken together. Thus there is considerable overlap between the two disciplines. corpora use a present tense link verb and in most of these the necessity to infer the meaning of the infinitive that occur after a present tense link verb denote different research actions, which result in observable differences between the two corpora in both the frequency and function of the it v-link ADJ to-inf pattern. 4.5.3 The frequency of the NECESSARY and USEFUL groups is similar in the two corpora, while that of the POSSIBLE and INTERESTING groups is higher in materials. All of the writer's stance and reveal rather different patterns of use in the two corpora. I begin by considering the DIFFICULT subgroup. 4.5.4.1 The DIFFICULT Subgroup

4(5) Steps for Building a Corpus

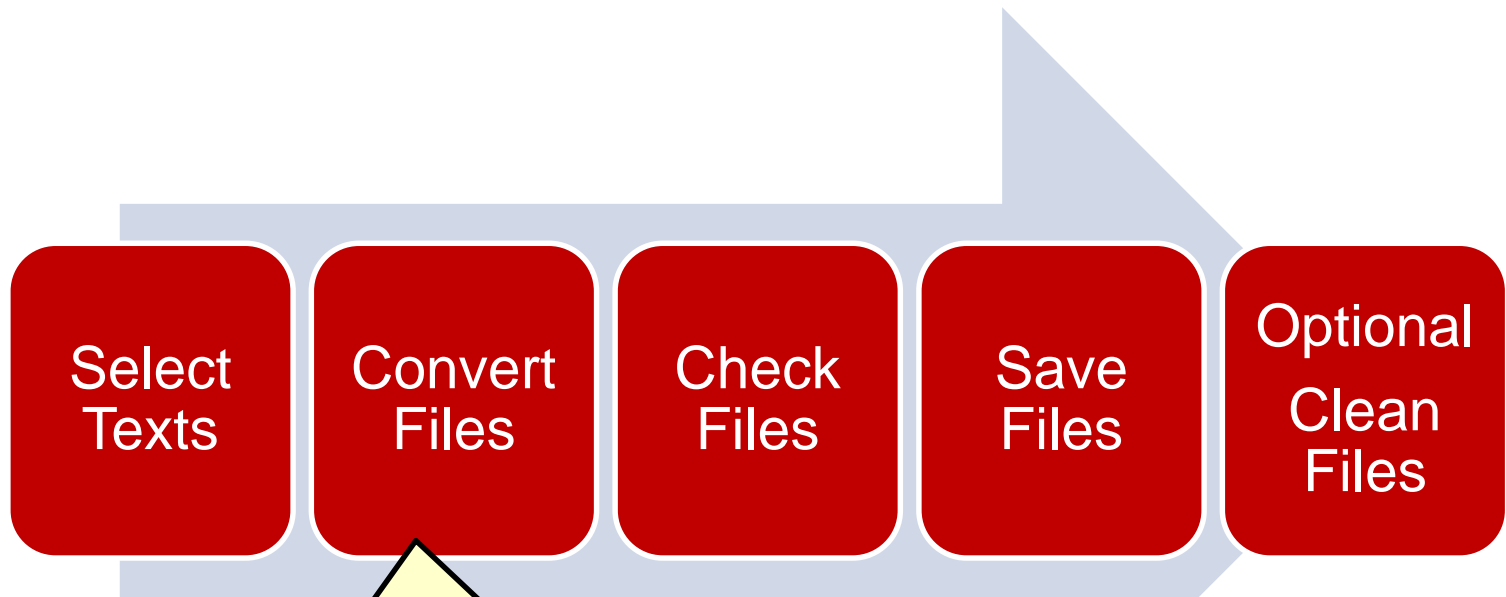


4(5) Steps for Building a Corpus



Choose texts that represent an appropriate genre

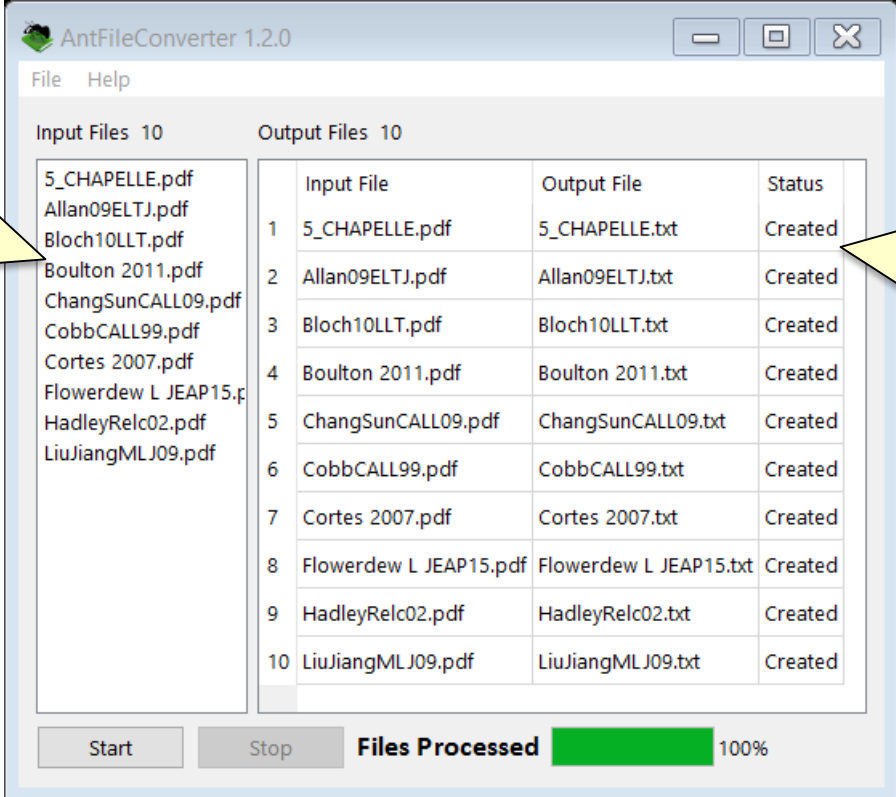
4(5) Steps for Building a Corpus



- Corpus files must be in plain text (.txt) format
- AntFileConverter converts multiple files simultaneously

AntFileConverter


Input several pdf or Word files here



File Help

Input Files 10 Output Files 10

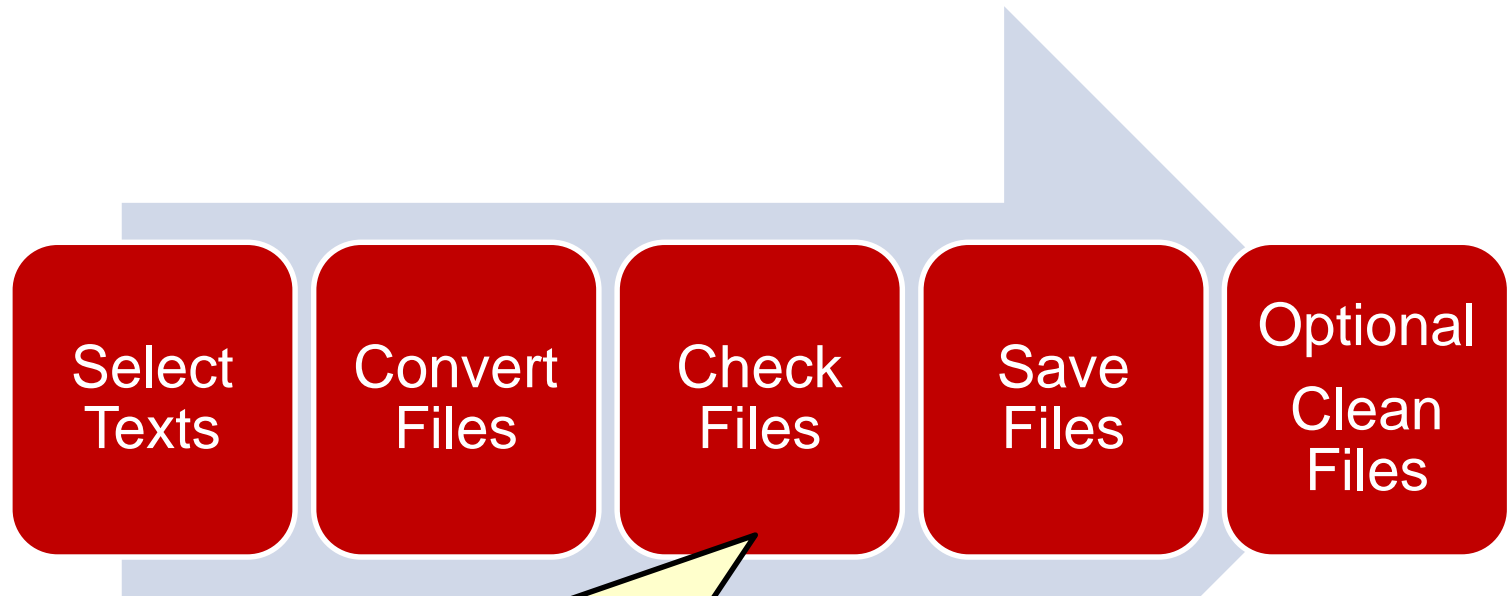
	Input File	Output File	Status
1	5_CHAPELLE.pdf	5_CHAPELLE.txt	Created
2	Allan09ELTJ.pdf	Allan09ELTJ.txt	Created
3	Bloch10LLT.pdf	Bloch10LLT.txt	Created
4	Boulton 2011.pdf	Boulton 2011.txt	Created
5	ChangSunCALL09.pdf	ChangSunCALL09.txt	Created
6	CobbCALL99.pdf	CobbCALL99.txt	Created
7	Cortes 2007.pdf	Cortes 2007.txt	Created
8	Flowerdew L JEAP15.pdf	Flowerdew L JEAP15.txt	Created
9	HadleyRelc02.pdf	HadleyRelc02.txt	Created
10	LiuJiangMLJ09.pdf	LiuJiangMLJ09.txt	Created

Start Stop Files Processed  100%

File conversions shown here

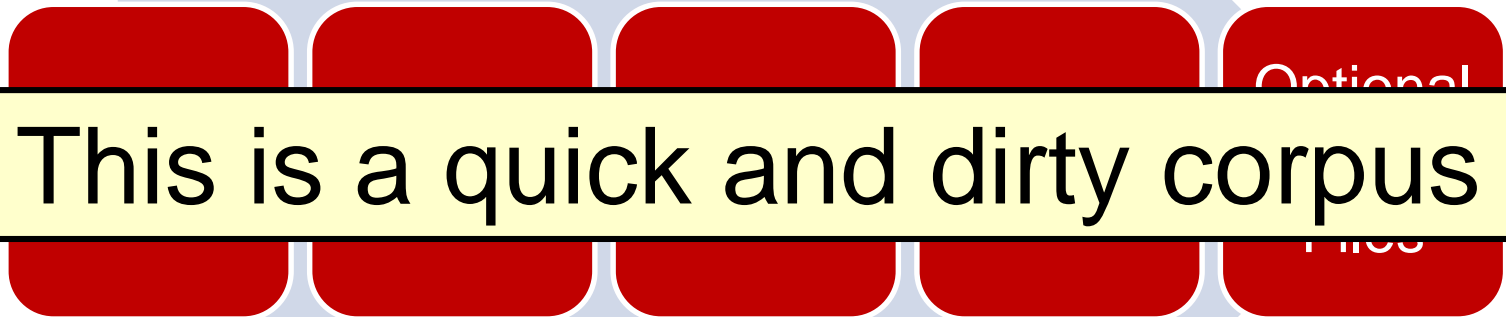
(Anthony 2015)

4(5) Steps for Building a Corpus



- Has the **whole** text converted?
- Have **line/word breaks** and **individual letters** converted correctly?

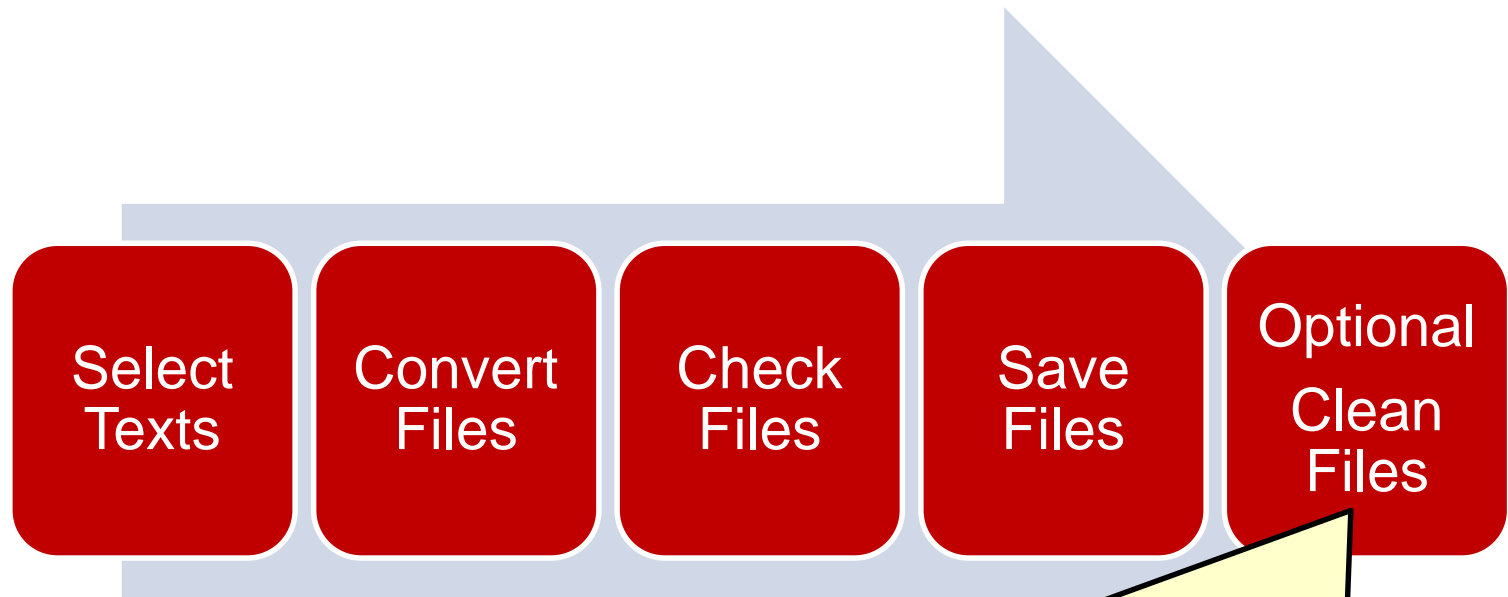
4(5) Steps for Building a Corpus



This is a quick and dirty corpus

- Save files to a corpus folder

4(5) Steps for Building a Corpus



- To improve results, delete everything that is not part of the running text (e.g. author, title, graphics)
- Try the dirty corpus first!

Summarising Process and Tools

Process

- Quick
- Easy
- Low-tech

Tools

- Free
- Available
- User-friendly

Part 3

Tools and Examples of Corpus-Assisted Editing

The Concordancer

an example is [NiFe]-hydrogenase, and inhibition by agent X is shown to be essentially cycle of this enzyme as well as in the inhibition by carbon monoxide or molecular oxygen cycle of this enzyme as well as in the inhibition by carbon monoxide or molecular oxygen hydrogenases are susceptible to reversible inhibition by CO and more devastating attack by hydrogenases are noted for their strong inhibition by CO, but unlike Pt-based catalysts, it is 20-fold more sensitive than Hyd-1 to inhibition by CO during H₂ oxidation. Fig. 6C shows the effect of O₂, and it is completely resistant to inhibition by CO even when a large excess of H₂ and O₂ is present, and the redox-state selective inhibition by CO. 39. || Foerster S, Stein M, B... Summary of all the constants for the inhibition by CO of H₂ oxidation and H⁺ reduction.

- searches the corpus for **every instance** of a word/phrase you choose
- presents each one **with its context** in a line on screen
- shows **search item in the centre**, with about 5/6 words on either side

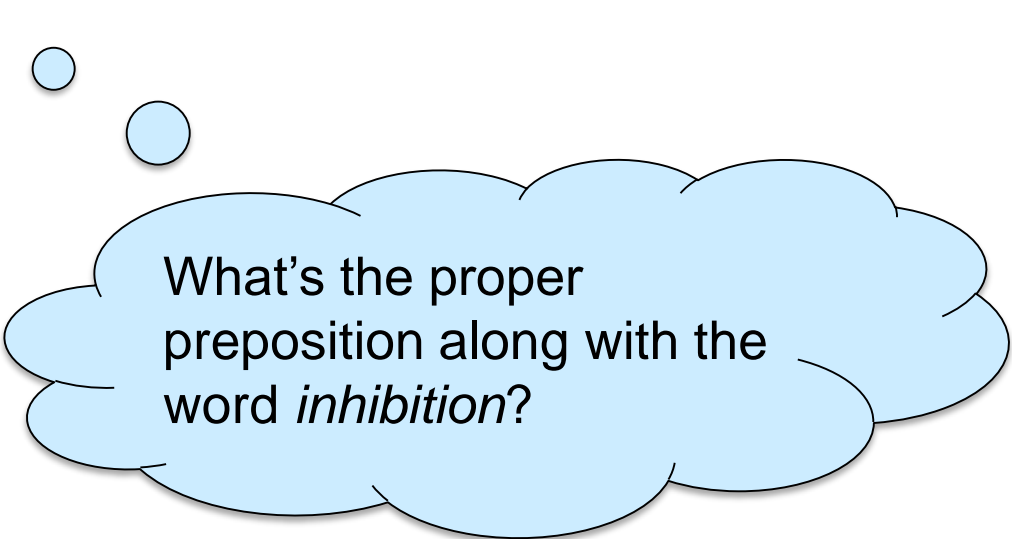
The Concordancer in Use: Siyu

Siyu: Chinese doctoral student in Chemistry

Corpus: 50 research articles; 394,000 words

Issue: Preposition use in specialized text

Siyu's Question



What's the proper preposition along with the word *inhibition*?

Corpus Files

- 01.activation and
- 02.investigating a
- 03.a metal-metal k
- 04.chem.rev.2007-c
- 05.direct electroc
- 06.an infrared spe
- 07.hydlhyd2.txt
- 08.curr. opin. che
- 09.current opinion
- 10.the interconver
- 11.Lubitz 2006 spe
- 12.Aa Hase I weekl
- 13.hy-d-1 crystal s
- 14.hy-d-1 EPR.txt
- 15.Pandelia 2010 F
- 16.JACS 2013.txt
- 17.Electrontransfe
- 18.Siegbahn 2007 C
- 19.chemphyscnem-[N
- 20.PCCP-2001.txt
- 21.CarbonFilmElect
- 22.Attenuated tota
- 23.Armstrong Phil.
- 24.NRVS-H2nase.txt
- 25.A Functional [N
- 26.Copying Biology
- 27.JACS 2005 The M
- 28.Siegbahn-D.giga
- 29.biochemistry200
- 30.IR spectroelect
- 31.SH-Review-2012.
- 32.Combining Spect
- 33.RH high concent
- 34.Silakov_2009_2E
- 35.The Auxiliary I
- 36.Nature Chemical
- 37.X3LYP.txt
- 38.Aa Hase I weekl
- 39.In situ step-sc
- 40.nature microbic
- 41.gromacs-MD-NiFe
- 42.Chem Com 2002-I
- 43.FAA-book_chapte
- 44.JACS-2005-Elect
- 45.2010-Inhibitor
- 46.Michael.Hall-JA
- 47.PCCP-2010-minir
- 48.Effective Core
- 49.Pickett 2003 Ch
- 50.Siegbahn_2009.t

Concordance	Concordance Plot	File View	Clusters	Collocates	Word List	Keyword List
Hit	KWIC					File
15	electronic and coordination changes during catalysis or inhibition brought about by light triggers, gas exchange or electroc					22.Attenuated tota
16	sitive potential, an example is [NiFe]Vhydrogenase, and inhibition by agent X is shown to be essentially instantaneous.					43.FAA-book_chapte
17	ic Inactivation by Increased Redox Potential43175.1.3. Inhibition by Carbon Monoxide43185.1.4.Other Inhibitors43195.2.					01.activation and
18	the active site equivalent to the Ni-B state.156 5.1.3. Inhibition by Carbon Monoxide CO is a competitive inhibitor of mos					01.activation and
19	n and cata- lytic cycle of this enzyme as well as in the inhibition by carbon monoxide or molecular oxygen and the light-sen					15.Pandelia 2010 F
20	n and cata- lytic cycle of this enzyme as well as in the inhibition by carbon monoxide or molecular oxygen and the light-sen					19.chemphyscnem-[N
21	n and cata- lytic cycle of this enzyme as well as in the inhibition by carbon monox					ght-sen 47.PCCP-2010-mini
22]- and [NiFe]-hydrogenases are susceptible to reversible inhibition by CO and more d					h [FeFe 02.investigating a
23	s and Discussion Hydrogenases are noted for their strong inhibition by CO, but unlik					restore 33.RH high concent
24	-1 to O2, is almost 20-fold more sensitive than Hyd-1 to inhibition by CO during H2					d-1 to 07.hydlhyd2.txt
25	mospheric level of O2, and it is completely resistant to inhibition by CO even when					oreover 02.investigating a
26	g simultaneous transfer of two electrons.156,283 5.2.3. Inhibition by CO Fe-Fe					d by CO 01.activation and
27	en proton, hydride and H2, and the redox-state selective inhibition by CO. 39					H, Hig 09.current opinion
28	4405 Figure 41. Summary of all the constants for the inhibition by CO					ctosoVo 02.investigating a
29	ion by O243195.2.2.Anaerobic Inactivation43195.2.3. Inhibition by CO43195.2.2					c Cyle 01.activation and
30	their assertion that Ni-SI is the target for competitive inhibition by CO presents a					ted the 02.investigating a
31	very reducing conditions (below -500 mV). ;@The issue of inhibition by molecular oxy					conseq 15.Pandelia 2010 F
32	very reducing conditions (below -500 mV). ;@The issue of inhibition by molecular oxygen is significant and has several conseq					19.chemphyscnem-[N
33	very reducing conditions (below -500 mV). ;@The issue of inhibition by molecular oxygen is significant and has several conseq					47.PCCP-2010-mini
34	interest, owing to their stability and insensitivity to inhibition by O2 and CO. So far, there is no crystallographic struct					01.activation and
35	foVibrio reveal that hydrogenases are more sensitive to inhibition by O2, CO, and NO than are the Ni-Fe and Ni-Fe-Se hydroge					01.activation and
36	y to zero for the AV and Dg enzymes, indicating complete inhibition by O2. In contrast, the current drops by only 30% when Re					44.JACS-2005-Elect
37	lems concerning the effectiveness of hydrogenases is the inhibition by O2 [6]. [NiFe] hydrogenases are generally considered m					41.gromacs-MD-NiFe
38	r circulator (Neslab). Kinetic analyses of hydrogenase inhibition by product and CO are described under supplemental data					07.hydlhyd2.txt
39	he limit of ;Winfinite rotation rate; where there is no inhibition by product. For each pH, these values can now be used to					43.FAA-book_chapte
40	uitable for applications because in these enzymes the O2 inhibition can be reversed, whereas in the [FeFe] hydrogenases expo					41.gromacs-MD-NiFe
41	tepiece from 0 to 50%. To dem- onstrate reversibility of inhibition, CO was then flushed from solution; in the case of bot					07.hydlhyd2.txt
42	alysts is CO (54). Fig. 6 shows experiments in which the inhibition constant Ki(CO/H2) was reactivation of Ni-A, with a ra					07.hydlhyd2.txt
43	ition of the H2 pro- H2 duction activity of Hyd-2, the inhibition constant Ki was mea- sured to be 210 ± 19 J.M at pH 6.					07.hydlhyd2.txt
44	F [NiFe]-hydrogenase, and the potential dependence of CO inhibition constants for D. fructosoVorans [NiFe]-hydrogenase. A vo					02.investigating a
45	r [NiFe]-hydrogenases lie in measure the Michaelis and inhibition constants of hydroge- nases for H2 and other gaseous reac					02.investigating a
46	edox potentials than those for Ni-A. Recently, this Na2S inhibition effect and reductive reactivation has been shown for seve					01.activation and
47	739 mV the hydrogenase appears to have recovered from CO inhibition. enzyme at ?639 mV, Ni;VR(1) becomes the species with					45.2010-Inhibitor
48	19, which arise from standard equations for competitive inhibition.66,94 Equation 18 shows the relationship between activity					02.investigating a
49	sensitivity to oxidative inactivation and small molecule inhibition, even for enzymes with very high sequence similarity. Onl					02.investigating a
50	Cycle Chemical Reviews, 2007, Vol. 107, No. 10 4319 inhibition even though O2 can reach the active site.251 For this r					01.activation and
51	ion. In the present work spectroscopic aspects of the CO inhibition for this bacterial organism are reported for the ?rst tim					45.2010-Inhibitor
52	hydrogenases and CO 4404 7.2.2. [FeFe]-hydrogenase: CO Inhibition 4405 3.1.2. Gas Supply and Gas Purity 4376 3.1.3. Ligh					02.investigating a
53	as active cul-de-sacs. 7.2.2. [FeFe]-hydrogenase: CO Inhibition Hatchikian and co-workers reported a value of KI) 0.16 µM					02.investigating a
54	mutant enzyme. This was in agreement with the lack of CO inhibition in activity measurements (data not shown). ;@Chemical Det					35.The Auxiliary I

24 hits for inhibition by

Total No. 50

Files Processed

Reset

Level 1 1R Level 2 2R Level 3 3R

Search Term Words Case Regex

Concordance Hits 119

Search Window Size 60

Corpus Files

- 01.activation and
- 02.investigating a
- 03.a metal-metal b
- 04.chem.rev.2007-c
- 05.direct electroc
- 06.an infrared spe
- 07.hydlhyd2.txt
- 08.curr. opin. che
- 09.current opinion
- 10.the interconver
- 11.Lubitz 2006 spe
- 12.Aa Hase I weekl
- 13.hyd-1 crystal s
- 14.hyd-1 EPR.txt
- 15.Pandelia 2010 F
- 16.JACS 2013.txt
- 17.Electrontransfe
- 18.Siegbahn 2007 C
- 19.chemphyscnem-[N
- 20.PCCP-2001.txt
- 21.CarbonFilmElect
- 22.Attenuated tota
- 23.Armstrong Phil.
- 24.NRV5-H2nase.txt
- 25.A Functional [N
- 26.Copying Biology
- 27.JACS 2005 The M
- 28.Siegbahn-D.giga
- 29.biochemistry200
- 30.IR spectroelect
- 31.SH-Review-2012.
- 32.Combining Spect
- 33.RH high concen
- 34.Silakov_2009_25
- 35.The Auxiliary F
- 36.Nature Chemical
- 37.X3LYP.txt
- 38.Aa Hase I weekl
- 39.In situ step-sc
- 40.nature microbic
- 41.gromacs-MD-NiFe
- 42.Chem Com 2002-I
- 43.FAA-book chapte
- 44.JACS-2005-Elect
- 45.2010-Inhibitor
- 46.Michael.Hall-JA
- 47.PCCP-2010-mini
- 48.Effective Core
- 49.Pickett 2003 Ch
- 50.Siegbahn_2009.t

- Concordance
- Concordance Plot
- File View
- Clusters
- Collocates
- Word List
- Keyword List

Hit	KWIC	File
70	CO is absent; red lines show the reactivation and rapid	10.the interconver
71	as obtained by experiments by Thauer et al. in which the	01.activation and
72	onation of SI-CO is blocked. Therefore, it seems that CO	30.IR spectroelect
73	0% H2/90% N2 to 10% H2/90% CO results in almost complete	02.investigating a
74	carried out various voltammetry experiments to study the	02.investigating a
75	to those obtained experimentally.122 The kinetics of CO	01.activation and
76	negative current oc- curred, due solely to specific CO	07.hydlhyd2.txt
77	04 Le'ger et al. have estimated the constant, KI, for H2	02.investigating a
78	The reduction current increases during this step, as H2	44.JACS-2005-Elect
79	tial results in about 40% inhi- bition of Hyd-1 and 100%	07.hydlhyd2.txt
80	tion band at 2060 cm ⁻¹ was ascribed to the extrinsic CO.	45.2010-Inhibitor
81	4402 4403 3.1. Electrochemical Equipment 4374 7.2.	02.investigating a
82	g CO) and CH KM app(CO/H)) K (CO/H) 1 + (20) 7.2.	02.investigating a
83	ummary: Similarities and Differences between CO and O2	02.investigating a
84	ummary: Similarities and Differences between CO and O2	02.investigating a
85	rt. Figure 9 shows the mechanism we propose for CO	30.IR spectroelect
86	cular reactions with redox partners. The strong	09.current opinion
87	consecutive scans. The peak reflects cyanide binding and	43.FAA-book chapte
88	d out of solution (Figure 40A). Similarly, reversible CO	02.investigating a
89	CO in argon resulted in a rapid loss of current, due to	07.hydlhyd2.txt
90	structural schemes is most likely a hydride ligand. 4.	15.Pandelia 2010 F
91	. The Catalytic Mechanism of [NiFe] Hydrogenases and the	15.Pandelia 2010 F
92	Ni2 + in the mechanism remains to be investi- gated. The	15.Pandelia 2010 F
93	structural schemes is most likely a hydride ligand. 4.	19.chemphyscnem-[N
94	. The Catalytic Mechanism of [NiFe] Hydrogenases and the	19.chemphyscnem-[N
95	Ni2 + in the mechanism remains to be investi- gated. The	19.chemphyscnem-[N
96	ite at the distal iron is blocked by CO, resulting in an	34.Silakov_2009_25
97	el center [18]. In the presence of oxygen, reversible	45.2010-Inhibitor
98	structural schemes is most likely a hydride ligand. 4.	47.PCCP-2010-mini
99	. The Catalytic Mechanism of [NiFe] Hydrogenases and the	47.PCCP-2010-mini
100	Ni2 + in the mechanism remains to be investi- gated. The	47.PCCP-2010-mini
101	Fig. S2). From sep- arate experiments to examine product	07.hydlhyd2.txt
102	jour nal homepage: ww w. elsevier.com/locate/bbabcio	45.2010-Inhibitor
103	.D. Peck, J. LeGall, P.A. Lespinat, The carbon- monoxide	45.2010-Inhibitor
104	ro H+ reduction activity under H2. This is mainly due to	02.investigating a
105	-558 mV to ensure lack of activity is mainly due to	44.JACS-2005-Elect
106	atalytic constants, activity, inactivation/reactivation,	02.investigating a
107	which are potent inhibitors.1-6 Consequently, overcoming	10.the interconver
108	enzymes opportunity to catalyze H2 oxidation before CO	02.investigating a
109	er, voltammetric cycles recorded during recovery from CO	10.the interconver
110	inhibition revealed quite large changes in waveshape as the CO level	10.the interconver

35 hits for inhibition of

Search Term Words Case Regex

Total No. 50

Files Processed

Reset

Concordance Hits 119

Search Window Size 60

Advanced

Start Stop Sort

Kwic Sort

Level 1 1R Level 2 2R Level 3 3R

Save Window

Exit

Concordance Outcomes

Siyu's Examples and Conclusions

*inhibition **by** CO or product* (24 hits)

By is used for describing which species impose this effect

*inhibition **of** Ni-Fe hydrogenases* (35 hits)

Of is followed by species having this behaviour.

Why use the Concordancer?

- To check/find ***collocations, phraseology, patterns***
- To see ***many examples*** of a word/phrase at the same time
- To find ***specialist information not available elsewhere***

Concordance Plot

HIT FILE: 1 FILE: Chapter1.3.txt



No. of Hits = 84
File Length (in chars) = 125747

HIT FILE: 2 FILE: Chapter2.2.txt



No. of Hits = 133
File Length (in chars) = 128932

HIT FILE: 3 FILE: Chapter3.1.txt



No. of Hits = 55
File Length (in chars) = 73378

HIT FILE: 4 FILE: Chapter4.2.txt



No. of Hits = 18
File Length (in chars) = 55333

- provides a ***graphic display***
- shows where the search item occurs ***within a file***
- gives a ***simultaneous overview of all files*** in a corpus

Concordance Plot in Use: Andrea

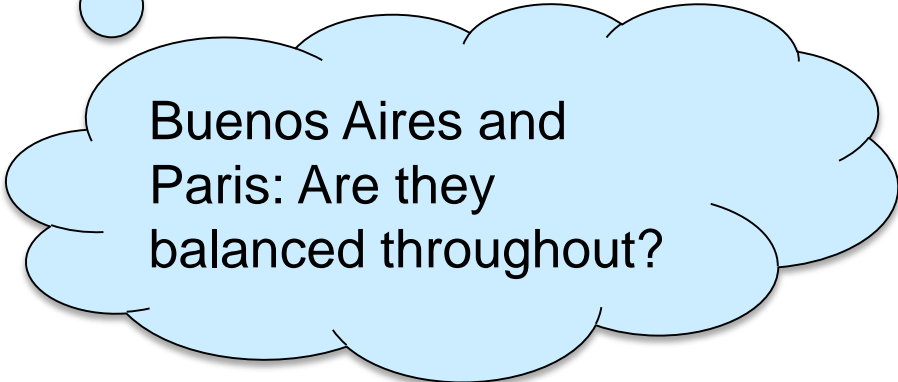
Andrea: Dominican doctoral student in Modern Languages

Corpus: 4 chapters of her thesis (64,000 words)

Thesis: Compares Buenos Aires and Paris in work by Borges and Réda

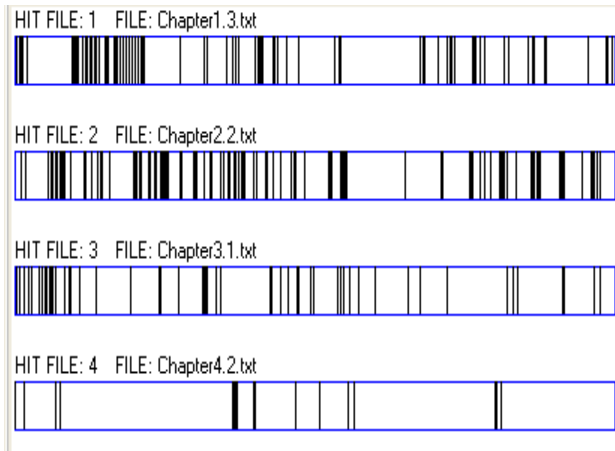
Issue: Checking the development of the topic

Andrea's Question



Buenos Aires and
Paris: Are they
balanced throughout?

Comparison: *Buenos Aires, Paris*



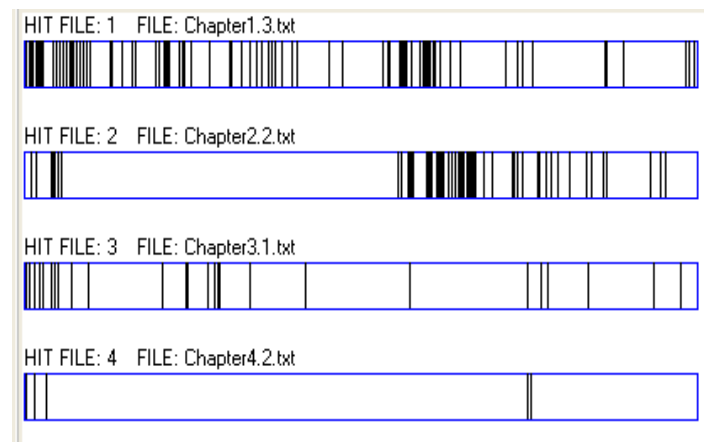
Buenos Aires

Chapter 1: **84** hits

Chapter 2: **133** hits

Chapter 3: **55** hits

Chapter 4: **18** hits



Paris

Chapter 1: **102** hits

Chapter 2: **65** hits

Chapter 3: **27** hits

Chapter 4: **5** hits

Concordance Plot Outcomes

Andrea's Conclusions

Chapter 2: *Balance the Buenos Aires and Paris sections.*

Chapter 3: *Investigate the structure of the chapter.*

Chapter 4: *Very few hits for both cities. Is another theme emerging that needs to appear throughout the thesis (i.e. imminence)?*

Why use Concordance Plot?

- To track **content, ideas, terms** in a single file
- To **compare usage across files**
- To **check content issues** in a long text

The N-Grams Tool

Rank	Freq	Range	N-gram
Total No. of N-Gram Types 6235			
Total No. of N-Gram Tokens 21551			
1	95	33	the use of
2	90	36	in order to
3	55	24	the fact that
4	45	13	men and women
5	43	26	as well as
6	42	23	it is not
7	42	25	there is a
8	40	20	a number of
9	40	22	in terms of
10	39	28	one of the
11	36	21	it has been
12	34	23	that there is
13	33	17	on the other
14	33	22	that it is
15	31	12	as a result
16	31	9	males and females
17	31	16	the other hand
18	31	24	there is no
19	30	17	due to the
20	30	21	in the same

- shows a list of ***all word sequences*** of a length n you choose
- presents them as a ***list*** and gives their ***frequency***
- the procedure is ***automatic***

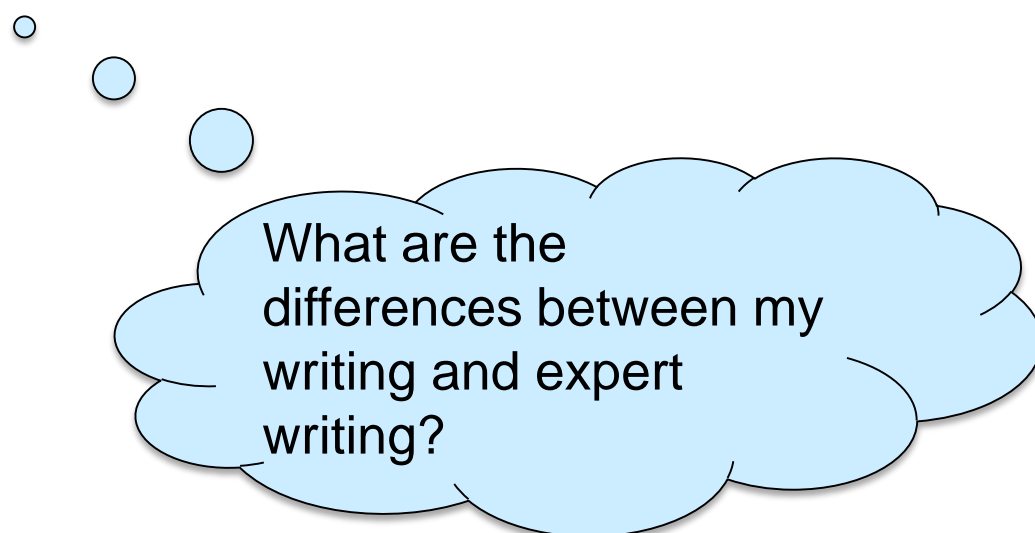
N-Grams in Use: Hiromi

Hiromi: Japanese doctoral student in sociology

Corpora: 52 research articles; 523,427 words
4 thesis chapters; 18,945 words

Thesis: Integration of immigrants in Japan

Hiromi's Question:



What are the differences between my writing and expert writing?

Hiromi's Top Five 3-grams

Research Article Corpus

1. of national identity (192)
2. **as well as (150)**
3. of the nation (135)
4. **in terms of (119)**
5. **there is a (90)**

Thesis Corpus

1. of national identity (55)
2. national identity and (46)
3. civic national identity (34)
4. ethnic national identity (31)
5. and attitude toward (27)

Hiromi's research article corpus contains **2 referential expressions** and **1 discourse organizer** (Simpson-Vlach & Ellis (2010))

Her own writing contains only **content-related** 3-grams

N-grams Outcomes

Hiromi's conclusions

- *I should check if I can write more sentences using the **general expressions**.*
- *It may be that I need more **interpretations of the results**.*
- *How is '**there is a**' used in my research article corpus?*
- *It is used to **reference the previous research** and to explain the **gap in the field** of study, as well as to **explain the results of the statistical analysis**.*

Why use the N-grams Tool?

- to **identify frequent expressions**
- to **explore the difference** between **student writing** and **expert text**

The Keywords Tool

Corpus Files		Concordance	Concordance Plot	File View	Clusters/N-Grams	Collocates	Word List	Keyword List
Chapter 2 October 20:		Types Before Cut: 1936		Types After Cut: 1936		Search Hits: 1		
Rank	Freq	Keyness	Keyword					
434	21	1.825	these					
435	5	1.789	eighth					
436	5	1.789	fourth					
437	5	1.789	settlement					
438	8	1.774	reared					
439	13	1.737	iron					
440	13	1.737	thames					
441	34	1.729	but					
442	6	1.654	before					
443	6	1.654	der					
444	6	1.654	pattern					
445	4	1.646	apparently					
446	4	1.646	culture					

- Identifies words which are *unusually frequent* or *infrequent* in one corpus when compared to a reference corpus
- Gives insight into the *content of individual chapters* compared to the whole thesis

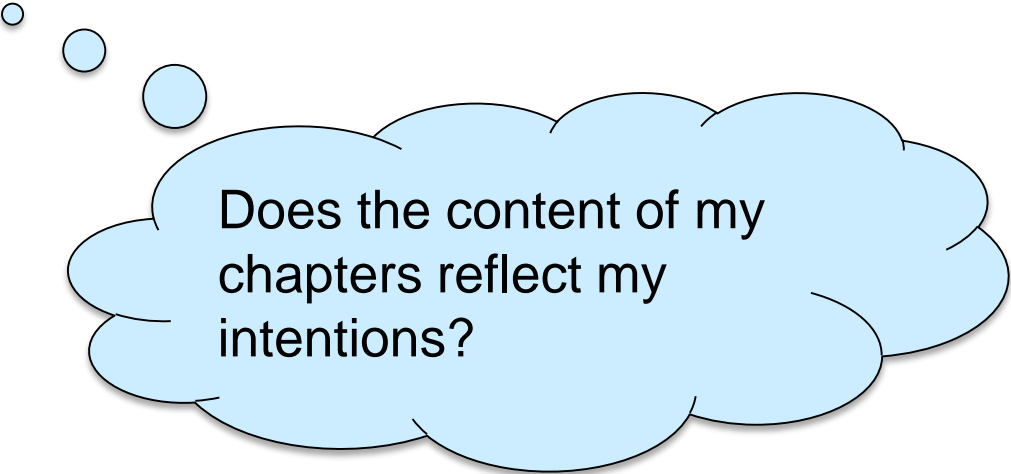
Keywords in Use: Keiko

Keiko: Japanese doctoral student in archaeological science

Corpus: 7 thesis chapters; 57,492 words

Thesis: Transition from the Roman period to the Anglo-Saxon period in the Upper Thames Valley: Analysis using stable isotope data

Keiko's Question:



Does the content of my chapters reflect my intentions?

Keiko's Keywords

Literature Review

AntConc 3.4.4w (Windows) 2014

File Global Settings Tool Preferences Help

Corpus Files
Chapter 2 October 20:

Types Before Cut:	1936	Types After Cut:	1936	Search Hits:	1
Rank	Freq	Keyness	Keyword		
434	21	1.825	these		
435	5	1.789	eighth		
436	5	1.789	fourth		
437	5	1.789	settlement		
438	8	1.774	reared		
439	13	1.737	iron		
440	13	1.737	thames		
441	34	1.729	but		
442	6	1.654	before		
443	6	1.654	der		
444	6	1.654	pattern		
445	4	1.646	apparently		
446	4	1.646	culture		

Search Term Words Case Regex Hit Location Search Only 1

Total No. 1

Files Processed

Sort by Invert Order Sort by Keyness

Discussion

AntConc 3.4.4w (Windows) 2014

File Global Settings Tool Preferences Help

Corpus Files
Chapter 6 October 20:

Types Before Cut:	1763	Types After Cut:	1763	Search Hits:	1
Rank	Freq	Keyness	Keyword		
94	31	3.382	seasons		
95	27	3.333	no		
96	18	3.267	does		
97	18	3.267	fed		
98	96	3.224	this		
99	41	3.210	however		
100	41	3.210	neolithic		
101	11	3.208	consumption		
102	5	3.153	brittany		
103	5	3.153	english		
104	5	3.153	prunus		
105	5	3.153	unlikely		
106	5	3.153	vetch		
107	5	3.153	waterhole		

Search Term Words Case Regex Hit Location Search Only 1

Total No. 1

Files Processed

Sort by Invert Order Sort by Keyness

Keywords Outcomes

Keiko's conclusions

Chapter 2 Literature Review: *iron* positive keyword

Chapter 6 Discussion: *neolithic* positive keyword

'I talk about Iron Age more in Chapter 2 (Literature Review) and Neolithic period more in Chapter 6 (Discussion), but my main focus is in the Roman and Anglo Saxon period. References to Iron Age and Neolithic should be reduced'.

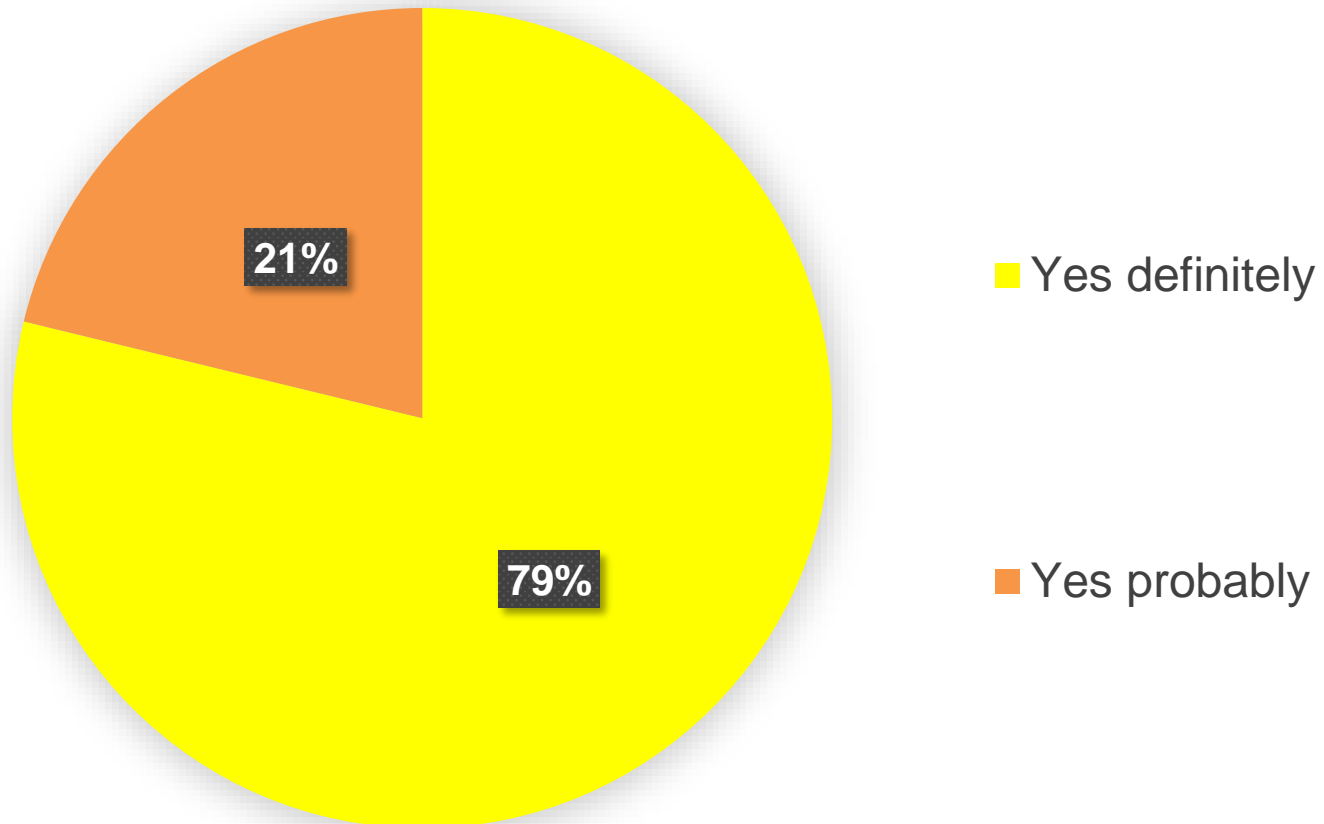
Why use the Keywords Tool?

- to allow ***aspects of content*** to emerge
- to ***identify content issues*** the student is not aware of

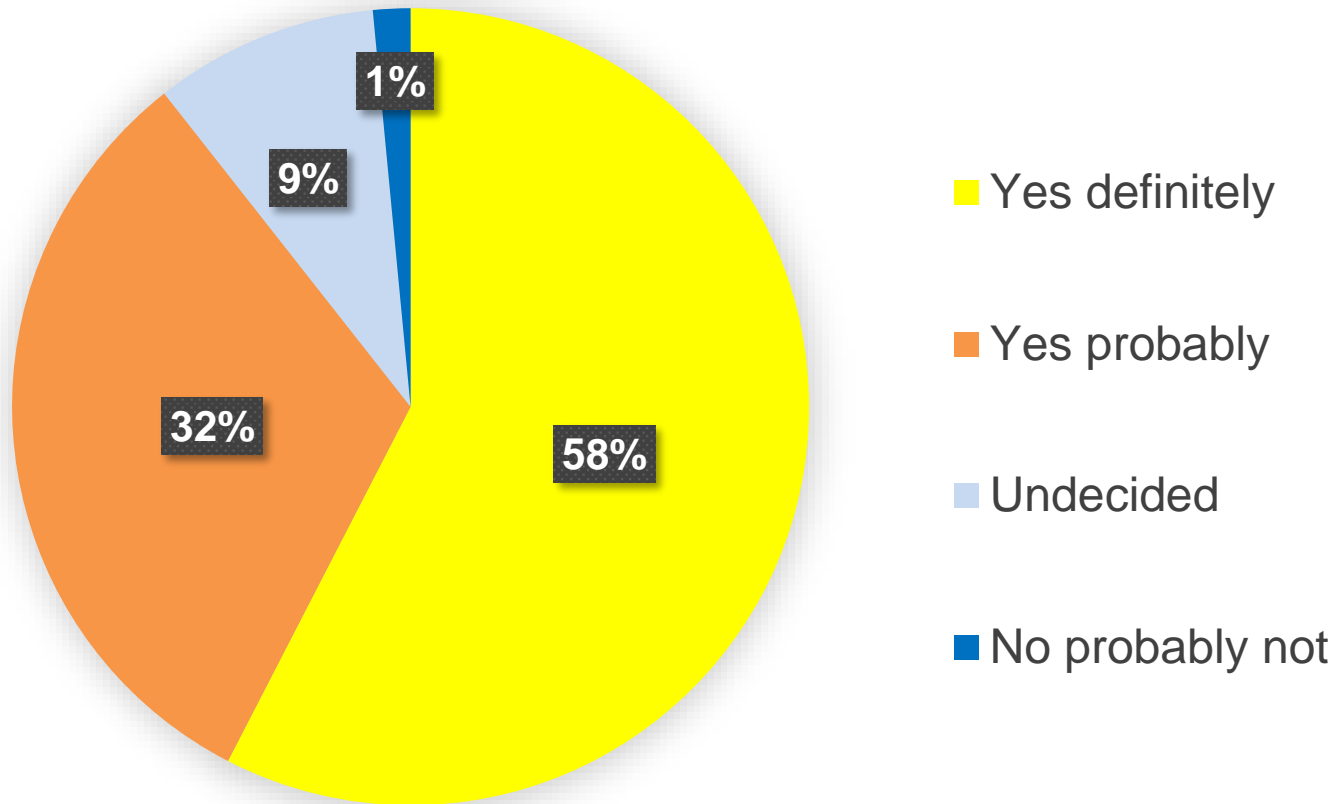
Part 4

Evaluation of the Approach

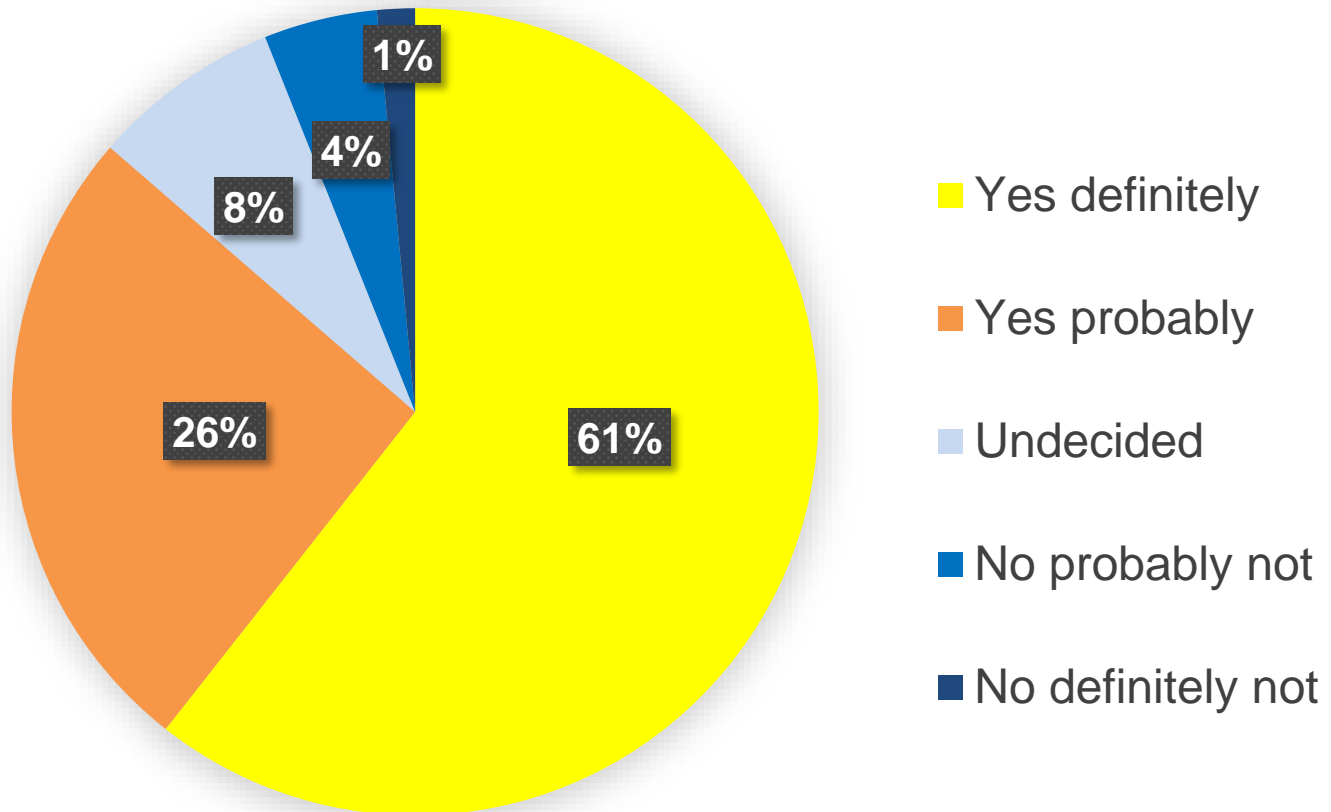
Is it helpful to use your corpus and AntConc for editing?



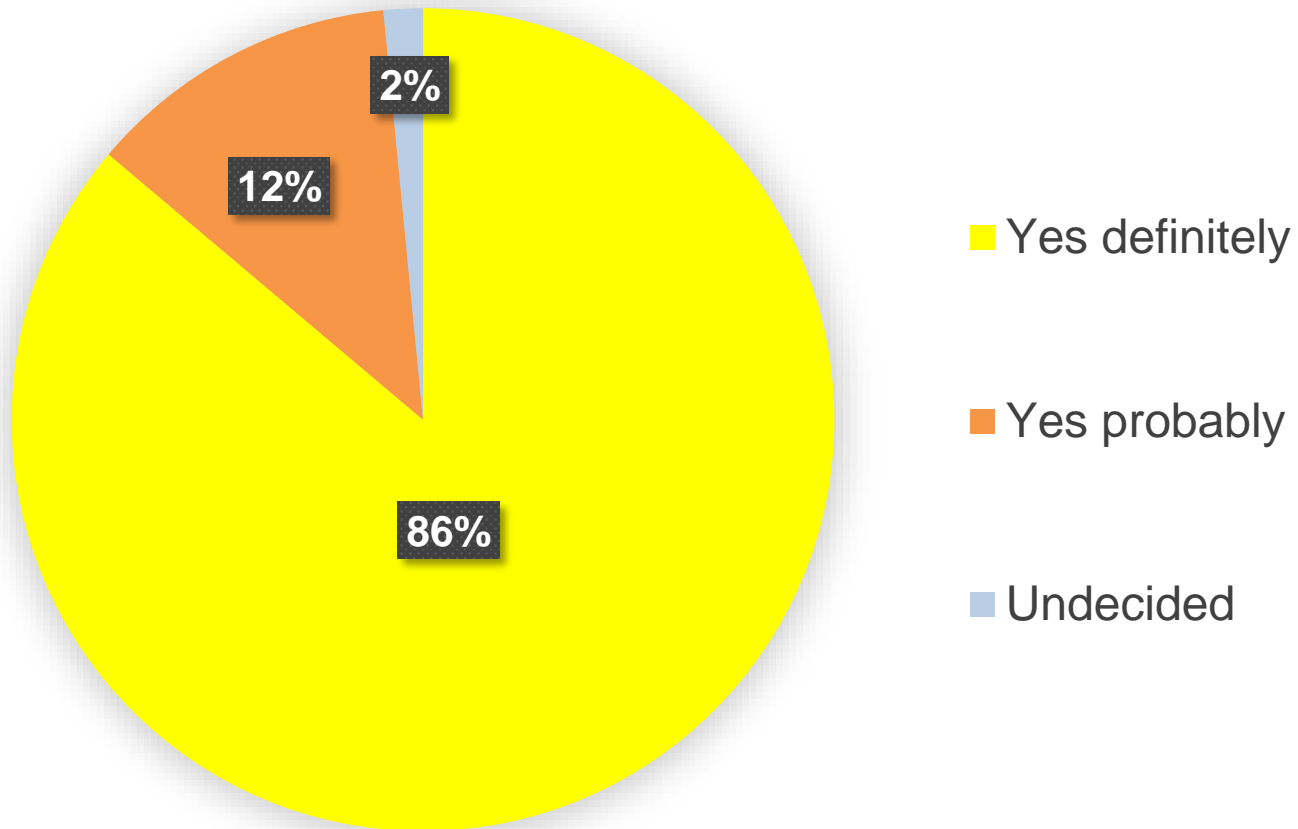
Is it easy to use the AntConc software?



Was it easy to build your corpus?



Do you intend to use your corpus and AntConc for editing in the future?



Part 5

In Conclusion

Affordances of Corpus Tools for Editing

- enable ***comparisons*** of student writing e.g. with expert texts or between chapters
- facilitate ***a focus on language, content and organisation separately***
- show ***issues in language, content and organisation*** that are not evident when reading linearly
- ***de-familiarise*** an over-familiar text



A bird's eye
view from
above

A bug's
eye view
from below



References (1)

- Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/>
- Anthony, L. (2015). AntFileConverter (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/>
- Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, 31(2), 93–102.
- Charles, M. (2015a). Same task, different corpus: The role of personal corpora in EAP classes. In A. Boulton & A. Leńko-Szymańska (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning* (pp. 131–153). Amsterdam: Benjamins.
- Charles, M. (2015b). Genre, corpus and discourse: Enriching EAP pedagogy. In P. Thompson & G. Diani (Eds.), *English for Academic Purposes: Approaches and Implications*. Newcastle upon Tyne: Cambridge Scholars.
- Charles, M. (2017). Do-it-yourself corpora in the classroom: Views of students and teachers. In K. Hyland & L. Wong, (Eds.), *Faces of English education: Students, teachers and pedagogy* (pp. 107–123). Abingdon: Routledge.

References (2)

- Kübler, N. (2011). Working with corpora for translation teaching in a French-speaking setting. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New Trends in Corpora and Language Learning* (pp. 62–80). London: Continuum.
- Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25(1), 56–75.
- Seidlhofer, B. (2000). Operationalizing intertextuality: Using learner corpora for learning. In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 207–223). Frankfurt: Peter Lang.
- Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487–512.