Creating a small corpus to inform materials design in an ongoing English for Specialist
Purposes (ESP) course for Orthodontists and Orthodontic Assistants

Marc Jones

University of Portsmouth

marc.jones@myport.ac.uk

Abstract

In order to inform materials design for an ESP course for orthodontists and orthodontic assistants, the teacher-researcher assembled a small corpus of practitioner-to-patient written texts by using a search engine and copying websites using ICEWeb software (Weisser 2014-2016), cleaning the text manually and then analyzing the text with AntConc software (Anthony, 2014).

Findings include a balance between nominal and adjectival modification, tendency toward attributive processes in first and second person while third-person singular verbs have attributive and also material and mental processes. In addition, collocations of 'bite', 'teeth' and 'braces' are detailed.

Throughout the article, the role of corpus compiler-as-researcher (Koester, 2010) is discussed as an advantage in understanding the corpus by locating findings within a deeper contextual use. Limitations are discussed regarding software use, cleaning and use of written discourse to inform spoken discourse.

Key words: ESP, pedagogy, orthodontics, modification, transitivity

1. Introduction

In the author's work as a language teacher in Japan, a group of orthodontists and orthodontic treatment assistants were clients, seeking to improve communication with English-speaking interlocutors encountered daily. These interlocutors are either patients or parents and guardians of patients. There were no materials available for this particular language-teaching context, and the clients informed the author that previous language teaching providers had attempted to sell textbooks that did not meet their needs.

A search for orthodontic corpora within Sketch Engine (Lexical Computing, 20??) and within the Google Plus Corpus Linguistics Community (2014?-2017) returned no results. The client required no need for written communication at any length because they already have several forms with checkboxes providing overviews to patients and caregivers about treatment. One reiterated need was to develop speaking and listening skills for communication.

The exact needs of the client expressed were explanation of treatment procedures and providing information about orthodontic device maintenance and cleaning procedures. This was required to be clear and concise, with many of the clients' patients using English as L2 due to low proficiency in Japanese, and the clients' practitioners and other employees speaking no other foreign languages than English. While simple communication could be achieved, comprehension and comprehensibility caused anxieties among the practitioners. Additionally, there is a lack of knowledge about what does and does not constitute jargon; as Dahm (2011) states about international medical graduates in Australia: "IMGs may have trouble in identifying what is likely to be 'medical jargon' from the perspective of the patient and this, in turn, will create additional complications in deciding when and how to avoid or explain jargon." (p.188) Jargon, in this case, is defined as technical terminology not likely to be understood by the general public at large. Thus, in constructing a corpus of orthodontic patient-facing discourse, jargon should largely be avoided and any examples found would, through common sense assumption, be situated in comprehensible contexts. Using such a corpus to inform language instruction was assumed to should result in greater comprehensible practitioner-patient communication.

The current study's purpose is highly applied and was undertaken by a language teacher as researcher, as opposed to a researcher working with a language teacher as informant/participant. The reason for this was practical: the expense in hiring researchers to construct a corpus to support small-scale language teaching projects is obviously untenable. Additionally, the corpus construction and analysis were time sensitive projects due to the need to inform an ongoing language course. Therefore the author constructed the corpus, with some methodological and data-handling limitations and errors, which are outlined in section 4. However, the resultant corpus was sufficient to meet the author's purposes which were to gain data on language patterns in the target context (i.e. orthodontic practitioner-patient communication), in order to provide English language instruction to meet communicative norms in that context.

Therefore, the purpose of the study is not the development of theory in corpus linguistics, nor the development of corpus analysis, but as a case study of one corpus assembled for the purposes of language teaching for specific purposes. The contribution to the community may be seen in the need for clearer guidelines for corpus construction in part 6, as well as the findings of the corpus analysis in part 5, which may inform similar contexts. The work may also inform language teachers beginning work with their own corpus construction regarding the possible outcomes of early-stage work. Additionally, for orthodontic practitioners using English as a second language or lingua franca, it may form a basic model for collocations and contextual language use.

2. Literature Review

There is a clear rationale for using a corpus to inform the materials (including lesson planning) for learners. In fact, Long (2015) advocates the use of discourse analysis in order to teach language that is authentic. Furthermore, Hou (2014) comments that teachers and materials writers can provide greater affordances for target language comprehension if they understand which words and their meanings are most likely to be encountered in professional contexts. Corpus linguistics provides a method enabling researchers to take a larger sample of such words than may ordinarily be handled in a non-corpus based discourse analysis. Handford (2012, p. 23) supports this position bemoaning “the contrived and potentially inappropriate language examples and concepts to be found in certain textbooks and training manuals”. Bearing in mind that it may be impossible to gain access to practitioner-patient discourse without breaking the law, or even obvious ethical considerations, corpus linguistics provides the most practical way to do this.

While it would have been advantageous for the fields of both ESP and also orthodontic practitioners using English as L2 to address the difference between spoken and written practitioner-patient discourse, it would perhaps be overly ambitious. Gaining access to a community as a non-healthcare practitioner while accessing private healthcare details is extremely difficult due to patient reluctance. Additionally, for practitioners, it may even be illegal to provide such access. Even for contexts other than medical discourse, professional discourse research is difficult due to permission required and time for recording and transcription (Handford, 2012). For teachers with large workloads and/or without institutional support, an approach of recording and transcription may be impossible. In the current study, with no funding resources, yet with existent access issues and limited time, a spoken-discourse corpus was untenable.

With a sample of patient-facing texts, it is possible to select language that practitioners would ordinarily expect patients to understand. However, without access to significant resources (of time, money or human resources), there is little other option but to use existing corpora (of which none were suitable for the needs of the client) or to assemble a small corpus assembled. However, the limits of this study may also provide benefits, due to the corpus user/analyst being the corpus compiler, quantitative and qualitative data from and about the corpus may be more readily available to complement one another (Koester, 2010, citing Flowerdew, 2004, and O’Keeffe, 2007)

It is also argued by Boulton et al. (2012, p.3) that the more specialised the discourse to be analysed, the smaller the corpus can be. This may be mitigated by preparing a corpus where the corpus compiler had access to the client who was able to provide information regarding needs and typical scenarios encountered with patients and/or patients’ guardians. Therefore in spite of the researcher’s limited familiarity with the context, the client’s emic knowledge could be leveraged. This is supported by Shortall (2007), who recommends that a corpus represents domain-specific language by the target group. Therefore, it is clear that a written corpus with the same target interlocutionary relationship is the nearest useful solution in assessing linguistic features that are likely to form much of the content in the discourse

between orthodontic practitioners and their patients or their guardians. Looking at rhetoric features in texts is advocated as good practice by Mona Baker (1988).

While it is easy to run a simple survey of the corpus for words by frequency, frequency itself may not be enough. A reason for this is that grammatical words, particularly determiners, appear at the top of such word frequency lists and other words that are less frequent but more markedly frequent than in their use by other discourse communities and contexts, may lie further down and be easily overlooked. Hence the importance of keyness as opposed to simple frequency (Koester, 2010). This necessitates the production of a keyword list and using this to inform the lexis to be taught due to it being more likely to be a part of the typical practitioner-patient discourse.

However, corpus informed materials should not be limited to word lists. Additional ways that corpora may inform teaching materials are through collocation and colligation, which may be found in the clusters and/or N-Gram functions of concordance software such as AntConc (Anthony, 2014). There are clear advantages to this for ESP courses: the language that is most commonly used, or else is key, should be prioritized in order for mutual comprehensibility to be maximized. The use of chunks (here meaning multi-word items) “provide a lexicon as useful as, (and often more frequent than) the single-word lists that commonly form the basis of the vocabulary syllabus of a language course.” (McCarthy & McCarten, 2012, p. 230)

Beyond lists of items, Jabbour (1997) sees corpus linguistics as improving teaching at the syllabus level. “Corpus linguistics should be seen as the instigator of a better theory of syllabus design... In ESP contexts, the course designer can use corpora for the investigation of the linguistic elements used in restricted languages and the type of patternings such languages are likely to use.” (p.17.) Jabbour is clear that corpus linguistics should be the instigator only. It is of utmost importance that corpus analysis remain merely a tool at the course designer’s disposal, not something that dictates a course with neither mediation nor mitigation.

3. Research Questions

Based upon the information in the introduction regarding the client’s context and the findings in the literature the following research questions were decided upon.

What language features arise significantly in the patient-facing websites?

The way this question was operationalized was to consider expectations regarding practitioner-patient communication and then examine words which are key compared to a reference corpus word list. Nouns and verbs are particularly useful in medical fields, being as nouns state the objects of treatment and the agent and verbs stating the method of treatment.

How should this inform which collocations and syntactic features to prioritize in form-focused lesson stages?

While nouns and verbs are examined in the first research question, how they are used is of importance, due to the different contextual nuances. While the corpus constructed was a written corpus, it would be expected that the collocation and other syntactic features in the patient-practitioner discourse would be expected to be understood by the orthodontic practitioners, even if this may not generalize to all of the patients communicating in L2 English.

4. Methodology

The corpus was assembled by scraping the search results of ‘orthodontist’, ‘orthodontic’ and ‘orthodontics’, pasting several of the top fifty search results from Google (chosen for relevance after being assessed by me as a source of patient-practitioner discourse). Unfortunately, at the time a list of websites scraped was not taken, which may result in any attempts to replicate this study varying, although this may be slight, given the search terms and the criterion of practitioner-patient communication. The website URLs were then pasted into ICE Web (Weisser, 2014-2016), as detailed in Navaratnam (2013). This software stripped the HTML tags from the content and exported it to a single text file. The resultant text file was then cleaned manually. One clear limitation of this method is that word counts of single texts and the number of single texts downloaded are unknown.

The cleaned text file was then tagged. Options included using the CLAWS tagging service from the University of Lancaster (Rayson, n.d.) or using TagAnt (Anthony, 2015). The CLAWS tagging service is limited to a corpus of up to 50,000 for free use therefore it was more beneficial to use TagAnt because it is free to use with an unlimited word count.

A keyword list was then generated by running an automated comparison in AntConc (Anthony, 2014) with a word frequency list from the BE06 corpus (Baker, 2009). The BE06 word list is from a written corpus with a median sampling point of 2006 and, due to it being a corpus of British English, should mark any non-British usages as more key than British usages in the corpus than British usages. The sampling strategy was to take samples of between 1,950 and 2,050 words from press, general prose, learned writing and fiction (Baker, 2009. p.317) and made up of “37% beginnings of texts, 26% middles of texts and 5% ends of texts. Additionally, 32% of the files contained full texts.” (Baker, 2009. p. 318) composing a corpus of 500 texts (Baker, 2009. p.316). The rationale behind this is to ensure that any non-British usages that may have been overlooked were also covered in the curriculum.

Additional exploratory analysis included an N-Gram search was run with the settings 5L 5R; clusters analyzed on both left and right sides up to both 5L and 5R for salient pedagogical words; collocations analyzed for salient pedagogical words at 5L 5R; searches

were also conducted for adjective-noun, noun-noun, and all verb tags and results checked manually.

An exploratory rather than experimental process was followed due to the corpus being for personal use and for pedagogical information only. Due to teaching commitments of the teacher-researcher an in-depth statistical analysis of corpus information was considered beyond the needs of the study, which was unfunded. Additionally, a list of websites downloaded and copied was not kept after construction of the corpus, nor was a word count for each text copied from each site possible from the ICE Web output, which was dumped into a single text file. This is an issue for replicability, yet the priority of the current study was pedagogy-focused research rather than empirical replicability.

5. Results

5.1 General Corpus Information

The resultant corpus consists of 126,366 tokens with 6847 word types, of which 61,871 are keyword tokens of 765 keyword types when run in AntConc's default settings. The top 50 keywords are provided in table 1. Some artifacts in the corpus are the terms 'x', 'xa', '(name)' and 'cobb' which arise from cleaning issues and boilerplate text.

Table 1: Top 50 key words. Keyness all positive values to 2 decimal places.

Rank	Word	Keyness	Rank	Word	Keyness
1	teeth	7457.3	26	aligners	579.46
2	braces	6277.97	27	adults	572.69
3	treatment	4800.51	28	x	569.22
4	orthodontic	4633.72	29	appliances	567.86
5	your	4594.52	30	xa	561.89
6	dental	2657.99	31	gum	535.81
7	tooth	2438.66	32	options	521.58
8	orthodontist	2314.9	33	removable	513.6
9	orthodontics	2011.59	34	brace	485.86
10	smile	1702.71	35	patient	476.86
11	can	1402.72	36	metal	465.86
12	orthodontists	1102.05	37	problems	453.39
13	you	1093.15	38	gums	442.87
14	invisalign	1084.48	39	cobb	434.57
15	bite	947.2	40	wear	430.31
16	jaw	895.21	41	adult	425.02
17	oral	791.83	42	decay	414.46
18	brackets	788.82	43	retainers	410.49
19	dr	770.76	44	wires	406.69
20	mouth	763.06	45	(name)	399.45

21	dentist	758.45	46	jaws	369.91
22	or	756.47	47	appliance	360.61
23	patients	742.12	48	dentistry	358.08
24	are	728.45	49	clear	355.94
25	may	651.53	50	upper	330.09

5.2 Noun modification

Prior to undertaking the project, I had anticipated the majority of nouns being subject to nominal rather than adjectival premodification of nouns, yet here was a lot less nominal premodification of nouns, with nominal premodification occurring at slightly over half the frequency of adjectival premodification.

Fig. 1. Proportions of noun premodification

However, nominal premodification is used within the corpus, and it tends to be used for orthodontic or dental terms such as ‘bite issues’, ‘bite problems’, ‘bite irregularity/~ies’, ‘tooth enamel’, ‘tooth movement’, and ‘jaw alignment’. However, where adjectival modification is used, it also tends to be used to talk about technicalities; ‘orthodontic’, ‘dental’, ‘clear’, ‘best’ and ‘oral’ are the five most key adjectives in the corpus; ‘orthodontic’, ‘dental’, ‘used’, ‘front’ and ‘lower’ are the five most frequent adjectives in the raw frequency list. This lack of markedness, while surprising, is logical, due to the corpus being one of patient-facing discourse. Patients may not have a wide orthodontic technical vocabulary, particularly patients who are children, therefore the use of adjectival modification is likely adopted in order for practitioners to be more easily understood, rather than using compound nouns which may appear to be more technical. When considering that many orthodontic patients are in early-mid adolescence, this becomes yet more important. Additionally, the corpus was assembled by searching for ‘orthodontic’ as well as ‘orthodontist’ as search terms; therefore this may have caused skew in the results.

In addition to the marked nouns, there are also uses of nominal premodification that are not technical but highly discourse appropriate, such as ‘treatment option(s)’ (frequency

65), and also ‘treatment plan’. These should be talked about with patients in the clinic. However, both are already know and also are not in the occupational duties of my learners. Another, irrelevant item is ‘insurance coverage’ due to Japanese orthodontics not being covered by national health insurance nor private providers.

5.2 Verb use

Present tense verbs and verb phrases abound in the corpus, and are far fewer than past tenses, other than past participles used in passive voice constructions, which are discussed in greater detail below. An exploratory approach was used with the verbs in order to find patterns of meaning. While a strict semantic analysis may offer interesting insights for the lay person, it is likely that the most proficient practitioners working for the client understand most of the senses of given verbs in general. Therefore, after consideration, it was decided that a situational use of particular verbs and transition processes may lead to more adaptability with language output. When examining the transitivity processes of verbs in the corpus, the processes used were those operationalized by Thompson (2004).

Table 2: verbs in the corpus, ranked by frequency

Verb	Frequency
is	28
are	24
be	21
been	4
s	3
was	3
were	3
am	1
being	1
re	1

When non-third-person singular present tense verbs are used, there is a greater tendency for them to be used for attributive processes, particularly with ‘need’, or mental processes such as ‘think’. The attributive processes are used to describe negative conditions, such as the teeth requiring treatment. This is unsurprising, due to the texts being sourced from websites hosted by orthodontists or professional organizations of orthodontists. The mental process ‘think’ is used by the orthodontic practitioners to correct patient/general public misconceptions.

Third-person singular verbs are used primarily for attributive, material and verbal processes. The attributive processes are mainly used to describe treatment and technology used by practitioners, which is again unsurprising. The material processes however, are rather interesting in that they tend to use the verbs ‘occurs’ (frequency 35), ‘happens’ (frequency

22) and ‘causes’ (frequency 21). These are processes relating to the teeth, many relating to natural processes occurring to untreated or neglected teeth. The verbal processes within the corpus are used for varied functions: ‘offers’ is used to inform patients about services and products available while ‘recommends’ is used to provide more persuasive information regarding treatment and examination; ‘says’ is used mainly for testimonials from patients; and ‘means’ is used to signal greater detail in an explanation. The first two are fairly obvious, while the latter two are relatively novel in their use, particularly with ‘says’ being almost exclusively to do with testimonials and ‘means’ being used as a discourse marker to signal greater depth.

The passive voice is used mainly for attributive processes, which are mainly used to explain and define tools and devices used for orthodontic treatment. This is primarily done by using ‘is used’ and ‘is called’; ‘is designed’ is also used to explain orthodontic appliances, in particular those which are custom made for individual patients.

5.3 Collocation

‘Teeth’ is the most key and seventh most frequent word in the corpus, and the second most key and frequent word after ‘you’. The key clusters that were present are ‘your teeth’, with a frequency of 287 occurring more often than the 48th most frequent word, ‘about’. This tends to be prefaced often with a base-form verb (89 incidences). Other relative common incidences with ‘your teeth’ are with prepositions, particularly ‘of’, in order to talk about the state of teeth, actions or processes toward the teeth, and locations or parts of patients’ teeth. The most common adjectives to modify teeth are ‘front’ and ‘lower’, which would relate to the state of patients’ bites more than other teeth. However, ‘upper teeth’ are mentioned with less than half the frequency of ‘lower teeth’, which would suggest that these are relatively unimportant in practitioner-patient communications. Some other highly frequent modifiers used are ‘permanent’ and ‘baby’, which have ramifications for the orthodontic treatment of children.

The word ‘bite’ has a specialist meaning in dentistry and orthodontics - being the position of the upper and lower teeth in relation to one another when the jaws are closed. The way that ‘bite’ is explained to patients in the corpus texts tends to be with collocates, particularly ‘correct bite’ and ‘bite problems’. The former is actually evenly spread between a verb-noun chunk and adjective-noun chunk. There are also occurrences of ‘correct bite problems’ in the text. Other collocates of ‘bite’ tend to be relatively negative, which leads to the semantic prosody (general connotation of meaning based upon collocation [Louw, 2000; Xiao & McEnery, 2006]) of ‘bite’ displaying negative and/or abnormal connotations, with a requirement for treating it.

Table 3: Top 20 lexical collocations of ‘bite’ taken at 3L3R

Rank	Total frequency	Frequency to left of bite	Frequency to right of bite	Word
------	-----------------	---------------------------	----------------------------	------

1	45	22	23	teeth
2	44	30	14	your
3	29	27	2	open
4	16	9	7	you
5	16	14	2	correct
6	15	2	13	problems
7	14	13	1	bad
8	13	5	8	can
9	12	4	8	jaw
10	12	6	6	bite
11	11	2	9	upper
12	10	5	5	underbite
13	10	3	7	together
14	9	1	8	chew
15	8	0	8	it
16	7	1	6	problem
17	7	1	6	occurs
18	7	5	2	healthy
19	7	6	1	deep
20	7	5	2	crossbite

Unsurprisingly, one of the most common N-Grams in the corpus was the combination ‘orthodontic treatment’, due to it containing one of the original search terms that seeded the corpus. However, there is more of interest here, in that it is common at the end of sentences but does also come mid-sentence, followed by ‘is’. When ‘orthodontic treatment’ is the theme of the sentence, it is generally followed by a sentence rheme that seeks to provide justification for orthodontic treatment or otherwise a rheme that defines what orthodontic treatment is. This provides an opportunity to see why the practitioner-as-compiler has greater insight than a third party using a small corpus (Koester, 2010; previously cited). There is very little need for my learners to define orthodontic treatment to their patients, because most of their patients are referred by their dentists or are already seeking continuing orthodontic treatment.

‘Teeth and gums’ as opposed to ‘gums and teeth’, confirms typical intuitive structure of this binomial through frequency in the corpus. The former occurs 29 times whereas there is only 1 occurrence of the latter. While the structure of binomials is not a high-stakes cause of misunderstanding, it may be considered to be advantageous to ensure correction in order for my learners to be more easily understood and for a better impression to be made by conforming to a norm.

Braces are an important part of orthodontic treatment and ‘braces’ has some interesting collocation that should be borne in mind by ESP teachers. ‘Clear braces’ comes higher than ‘(traditional) metal braces’, which I was advised is due to clear braces being more suitable for patients in North America due to the shape of Asian teeth being less suitable, and also due to the fact that the company supplying clear braces has an internet ordering service. While these collocations do not challenge my learners, the surrounding language is somewhat different, with there being a greater amount of explicit sales-orientated talk in KWIC lines for ‘clear braces’, whereas the KWIC lines for ‘metal braces’ tend to explain the benefits or mitigate the possible perceived disadvantages.

The three verbs that collocate with braces are ‘wear(ing)’, ‘get’, ‘have’ and ‘need’. The first three verbs connote possession to some degree, while ‘need’ connotes treatment. Therefore, it may be assumed that common-sense predictions are correct and that braces are not highly desired therefore the orthodontic practitioners are using broad language: ‘install’ does not feature at all; ‘set’ is only used as a noun with braces as ‘a set of braces’; and ‘fit’ is not a strong collocate of braces and is mainly used to describe parts of the braces and how they fit in the mouth, not that they are installed. ‘Fitted’ only collocates quite weakly in the corpus, and then quite distant at relatively low frequency (frequency of 5 where the position is up to 5L/R of fitted).

6. Conclusion and Limitations

The outcomes of the corpus research for pedagogy is positive but relatively underwhelming. The reasons for this is that the output from the corpus heavily reflects what had already arisen in Focus on Form (Long, 2015) in lessons, which does at least confirm that time has not been spent on low-priority language. However, the main findings of the project are that the client’s students should spend more time on explaining the benefits of orthodontic treatment and that their patients’ bite conditions will largely require correction. When one considers that they face patients who arrive at the clinic in order to request at least a consultation and probably a course of treatment, this becomes a case of stating the obvious. This is also discussed by Gavioli and Aston (2001) and Shortall (2007) who advises consideration of pedagogical factors in addition to corpus data.

Other than the above semantic prosody findings, there are insights into grammar, particularly in regard to attributive processes, material processes, and to a lesser extent verbal processes. Additionally, it appears to be the case that passive voice, when used, should

primarily be used for attributive processes to explain treatment processes and appliances used in treatment. Third-person-singular present tense active verbs should be prioritized mainly for talking about potential changes to the teeth and non-third-person singular verbs should be used for attributive processes with the teeth as the subject or for mental processes regarding preconceptions of orthodontic treatment.

The limitations of the corpus are clear, in that as a written corpus it can only be used as one source of lexicogrammar and must be heavily mediated by teachers regarding their knowledge of spoken lexicogrammar required for the orthodontic ESP context. This could be mitigated by the addition of transcription from orthodontic practitioner websites. However, this would not be a spoken sub-corpus but merely a written-to-be-spoken one. If access were available to authentic practitioner-patient discourse then this would provide a considerably useful albeit highly niche resource for the ESP community.

A cleaner corpus would be more advantageous due to the greater clarity of frequency that it would provide. Should a further iteration of this corpus be produced, it would be advisable to use software such as SarAnt (Anthony, 2016) or one of the tools on JusText (Pomikálek, 2011), or the more reliable version of BootCaT (Baroni & Bernardini, 2004) which now runs without problems on the author's computer unlike at the time this study was conducted in 2016. While ICE Web (Weisser, 2014-2016) was sufficient to construct a rudimentary corpus, there are issues with replicability as well as data transparency and length of texts scraped remains unknown. However, for the limited use to be made of this corpus, the author's knowledge of the corpus as compiler, cleaner and user affords the knowledge to avoid painstaking close work that may be unrewarded later.

7. References

- (2014?-2017) Google + Corpus Linguistics Community [Internet forum]. Retrieved from <https://plus.google.com/communities/101266284417587206243> December 18th, 2017

Anthony, L. (2014). AntConc (Version 3.4.4w) [Computer software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/>

Anthony, L. (2015). TagAnt (Version 1.2.0w) [Computer software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/>

Anthony, L. (2016). SarAnt (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/>

Baker, M. (1988) Sub-Technical Vocabulary and the ESP Teacher: An Analysis of Some Rhetorical Items in Medical Journal Articles. *Reading in a Foreign Language* 4 (2).

- Baker, P. (2009) 'The BE06 Corpus of British English and recent language change.' *International Journal of Corpus Linguistics*. 14:3 312-337.
- M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*.
- Boulton, A., Carter-Thomas, S., & Rowley-Jolivet, E. (2012) "Issues in corpus-informed research and learning in ESP" in Boulton, A., Carter-Thomas, S., & Rowley-Jolivet, E. (ed.) (2012) *Corpus-Informed Research and Learning in ESP: Issues and applications*. Amsterdam, John Benjamins.
- Dahm, M. L. (2011) Exploring perception and use of everyday language and medical terminology among international medical graduates in a medical ESP course in Australia. *English for Specific Purposes* 30 (2011) 186–197
- Gavioli, L. & Aston, G. (2001) "Enriching reality: language corpora in language pedagogy" in Seidlhofer, B. (ed.) *Controversies in Applied Linguistics*. Oxford: OUP.
- Handford, M. (2012) "Professional Communication and Corpus Linguistics" in Hyland, K., Huat, C.M., & Handford, M. (eds.) *Corpus Applications in Applied Linguistics*. London: Bloomsbury.
- Hou, H. (2014). Teaching Specialized Vocabulary by Integrating a Corpus-Design at the University: Implications for ESP Course Design at the University Level. *English Language Teaching*, 7(5), 26-37. doi:10.5539/elt.v7n5p26 Retrieved April 4th 2017 from <https://eric.ed.gov/?id=EJ1075729>
- Jabbour, G. (1997). Corpus linguistics, contextual collocation and ESP syllabus creation : a text-analysis approach to the study of medical research articles. University of Birmingham. Retrieved April 4th 2017 from <http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.342109>
- Koester, A. (2010) "Building small specialised corpora" in O'Keeffe, A. & McCarthy, M. (eds.) *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge.
- Long, M. H. (2015). *Second Language Acquisition and Task-Based Language Teaching*. Oxford: Wiley.
- Louw, B (2000) "Contextual Prosodic Theory: Bringing Semantic Prosodies To Life" in Heffer, C. and Sauntson, H. *Words in Context: A Tribute to John Sinclair on His Retirement*. Birmingham, University of Birmingham.
- McCarthy & McCarten (2012) "Corpora and Materials Design" in Hyland, K., Huat, C.M., & Handford, M. (eds.) *Corpus Applications in Applied Linguistics*. London: Bloomsbury.

Navaratnam, M. (2013, September 4) Building Your Own Corpus - ICEweb. [Blog post] Retrieved from <https://eflnotes.wordpress.com/2013/09/04/building-your-own-corpus-iceweb/> December 23rd, 2016.

Pomikálek, J. *Removing boilerplate and duplicate content from web corpora*. PhD thesis, Masaryk University, Faculty of informatics, Brno, Czech Republic (2011).

Rayson, P. (-) Free CLAWS WWW Tagger. Available from <http://ucrel.lancs.ac.uk/claws/trial.html>

Shortall, T. (2007) The L2 syllabus: corpus or contrivance? *Corpora* 2(2): 157–185

Thompson, G. (2004) *Introducing Functional Grammar*. London: Hodder Education.

Weisser, M (2014-2016) ICEWeb (Windows 32-bit). [Computer software]. Retrieved from http://martinweisser.org/ling_soft.html December 23rd, 2016.

Xiao, R. & McEnery, T. (2006) Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics* 27/1: 103–129