Exploring Register with *The Prime Machine*: Promoting register awareness through a module for English Majors at a Sino-British University in China

Stephen Jeaco
Xi'an Jiaotong-Liverpool University

The published Version of Record is:
Jeaco, Stephen. 2021. Exploring register with The Prime Machine : Promoting register awareness through a module for English majors at a Sino-British university in China. *Register Studies* 3:2 pp. 279–298.
https://doi.org/10.1075/rs.20015.jea
Please access the figures from the published version: https://www.jbe-platform.com/content/journals/10.1075/rs.20015.jea#dataandmedia

Abstract

Corpus approaches underpin a range of postgraduate studies and professional work in language, linguistics, translation and beyond. Awareness of the influences of contextual features on language choice is important for many activities: exploring new text varieties; finding relationships between social factors and language patterning; considering choices for post-editing machine translation; and understanding the very nature of language. Work on register relies on corpus methods, but more support and direction could be offered to help undergraduates gain earlier insights into the power of such corpus analysis. This paper introduces some ways register differences can be revealed through *The Prime Machine* corpus tool (Jeaco 2017a) and describes the design of a practically-oriented undergraduate module which uses this concordancer. Software features include the organization of texts and presentation of source information for readymade corpora, and methods which can be used to reveal useful starting points for register analysis of do-it-yourself corpora.

Key words: Register analysis, corpus tools, data driven learning, undergraduate corpus projects

1. Introduction

With growing interest in corpus methods for language learning, stylistics, sociolinguistics and translation, as well as Digital Humanities, new generations of language and linguistics graduates will need to apply these methods for a variety of work. Fostering an awareness of the need to consider

register differences when exploring and producing texts in specific domains, as well as scaffolding the skills with which to uncover register differences through their own research are key priorities for development of these competencies. However, most students in China (and the region) only typically encounter opportunities to develop corpus skills towards the end of their undergraduate degree, or at postgraduate level if at all. Earlier exposure to and adoption of corpus techniques could mean undergraduate linguistic and translation work could be better underpinned by quantitative data; these data also offer more power and authority for such students when working on language patterns in a foreign language (Mair 2002). From a language learning perspective, Data Driven Learning (DDL) has been shown to be effective (Boulton & Cobb 2017) and it engages students fruitfully (Flowerdew 2015). Pioneering work with postgraduates (Charles 2012; Johns 1991) and undergraduates (Cheng, Warren, & Xu 2003; Fligelstone 1993) serve as inspiration for hands-on concordancing work. If DDL is adopted for exploring linguistic differences between registers, undergraduate students should also benefit. Comparing linguistic features across two (or more) registers makes the individual characteristics of a register very much clearer (Biber & Conrad 2009). However, corpus tools do not typically offer much help for introductory undergraduate projects.

This paper introduces how *The Prime Machine* (tPM) (Jeaco 2017a) was designed to provide ways to highlight and explore register differences across readymade corpora and how theory on register influenced the development of tools for Do-It-Yourself (DIY) corpora. First, the ways the organization of corpus texts in *tPM*'s readymade corpora facilitate concordancing from a register perspective are presented, including divisions and subdivisions of corpora such as the *British National Corpus* (BNC; BNC Consortium 2007; Lee 2001), as well as functions like Key Labels (Jeaco 2020a). Then the paper describes how specialized text collections can be compared with these readymade corpora using *tPM*'s DIY tools. Finally, the paper reports on the use of *tPM* in an undergraduate corpus linguistics module for English majors studying at a Sino-British university in China, providing an overview of the course design and assessment. The paper explains the functionality of *tPM* from a register perspective and describes how these functions can promote register awareness in the foreign language classroom.

2. Background

This section provides an overview of the foundations for the approach proposed in this paper: the register perspective, Lexical Priming (Hoey 2005), and DDL. This section also summarizes some of the features for register analysis available in other corpus tools.

2.1 Register

In accordance with this special issue, register analysis is defined as the combined examination of common linguistic characteristics and situations of use. As Biber and Conrad characterize:

> The underlying assumption… is that core linguistic features like pronouns and verbs are functional, and, as a result, particular features are commonly used in association with the communicative purposes and situational contexts of texts.
>
> (Biber & Conrad 2009, p. 2)

While the central role of function and situational context is explicit and obvious in register work by Ferguson (Ferguson 1983), Biber (1991), and Biber & Conrad (2009), Conrad (2019) explains that this term is used differently by different authors and in different contexts, especially in terms of the relationships between linguistic features, function and social context. Conrad explains:

> Register analysis is described as having three components: the situation of use, including all aspects of the context of production or reception; the linguistic features; and the functional associations between the situational characteristics and the linguistic features.
>
> (Conrad 2019, p. 170)

She goes on to explain that since register analysis does not usually focus on forms which are exclusive to a register, the whole approach relies on considering relative measures – the quantitative differences in frequencies of linguistic features that emerge through comparisons with other text types.

Therefore, two important fundamentals of register work are the need for detailing the situational contexts of text production and the use of other corpora to bring out quantitative contrasts. The role computer software plays in register analysis should be to facilitate discovery and provide summary data of the first two components (details about the situations and quantitative data on linguistic features), so the researcher can interpret functional associations. When using a pre-existing corpus for register analysis (a readymade corpus), corpus tools need to provide access to the metadata or labelling related to the texts, and authors or speakers. To a large extent, the limitations for detailing the situational contexts are determined by the original corpus design. If the researcher constructs corpora specifically for register analysis work (DIY corpora), the details of situational contexts can usually be determined more fully as background information about texts can be noted as texts are collected.

A comprehensive framework for the systematic analysis of situational contexts is provided by Biber and Conrad, covering details of participants, their relationships, channel, production circumstances, setting, communicative purposes and topic (Biber & Conrad 2009, p. 40). Regarding use of other

corpora, Biber and Conrad explain, "… the characteristics of any individual register become much more apparent when it is compared to other registers" (Biber & Conrad 2009, p. 9). Indeed, Biber (2012) criticises some corpus linguistic work and associated reference works because the possibility of variation due to register differences has been overlooked. He argues that corpus studies should begin expecting a variation effect from register, and only disregard this if empirical data proves otherwise. Further, exploration of register can take place at different levels of text classification, from studies looking across major groupings of texts, down to more nuanced analysis within text collections – sub-variation within registers.

The second two components of register analysis – linguistic analysis and functional associations – depend on a linguistic framework, and several linguistic theories are generally compatible with register analysis. *tPM* was developed to build on Lexical Priming theory (Hoey 2005), so it is important to note how this theory relates to register analysis. Hoey presents his theory using corpus data as 'proof' for explaining the patterns of language use that must therefore exist in the minds of language users. His theory brings together collocation, colligation and semantic association, through the presentation of ten claims about how language is primed in its users. The importance of situations of use and the need for specialized corpora are clear in his notes below these claims:

> Very importantly, all these claims are in the first place constrained by domain and/or genre. They are claims about the way language is acquired and used in specific situations… corpus linguists have characteristically worked with general corpora. But certain kinds of feature only become apparent when one looks at more specialised data.
>
> (Hoey 2005, p. 13)

Although Hoey does not actually use the word *register*, it can readily be recognised that some kinds of relationships between contexts and use that form inputs for these primings are those described in this paper as register. For example, Berber Sardinha (2017) has explored patterns of collocations from a Lexical Priming perspective using a register approach. Within concordancing software, while basic search queries may not be heavily dependent on linguistic frameworks, more sophisticated techniques rely on programming decisions related to the patterning to be measured in software algorithms and the units of study the software designer considers important.

2.2 Data Driven Learning

Having established the importance corpora for register work, there remains a question about the desirability and effectiveness of engaging students in hands-on corpus activities as opposed to drawing on corpus-based descriptive works. The hands-on use of corpora in language learning

classrooms is known as Data Driven Learning (DDL). Work with postgraduate doctoral candidates includes early work by Johns (1991) with texts being selected by students and fed into the computer to allow the analysis of pairs of synonyms. More recently, Charles (2012) has demonstrated that doctoral students can create their own DIY corpora based on their reading texts, and use this as an effective database for editing their thesis. Over the last two decades, DDL has flourished and Boulton and Cobb (2017) have demonstrated it is an effective approach through their metaanalysis of published articles on DDL and language learning.

DDL is an approach that can be characterized as active learning, inductive learning or discovery learning (Bernardini 2004; Flowerdew 2015). It not only provides insights into the use of specific words and phrases; it also provides insights into language itself with concepts such as collocation being readily exhibited through engaging with corpus data. The strengths of DDL for language learning should also be clearly applicable to English majors who are studying about language using English as a second/foreign language. For such students, the DDL approach offers access to important linguistic concepts including register, and also provides multitudes of examples to further improve their knowledge of English. Fligelstone (1993) reported on three aspects of corpora and teaching: teaching about corpus linguistics, teaching students how corpus data can be exploited, and exploiting corpus data in order to teach. His paper described initiatives in these areas, including teaching undergraduates corpus methods through a general course on computing and language. Cheng et al. (2003) drew on DDL to give English majors hands-on tasks, merging material on discourse analysis and computer technology to create a successful module on corpus linguistics. For English majors studying translation, corpus skills could also be a powerful tool in post-editing translation activities. The question remains to what extent corpus tools facilitate all these activities.

2.3 Software for register analysis

Having introduced register, the link between register and Lexical Priming, and the potential of DDL for teaching linguistics, it is now important to consider the ways register can be analysed in other software. Since corpus tools generally provide frequency data and concordance lines, register analysis can be conducted using a variety of packages. Stand-alone software such as *WordSmith Tools* (Scott 2020), *AntConc* (Anthony 2019a) and *LancsBox* (Brezina, McEnery, & Wattam 2015) provide a variety of functions that can be applied to register analysis if texts are first divided by register and separate corpora are built using these groupings. Similarly, online corpus tools can provide ways of exploring register, with tools such as *CQPWeb* (Hardie 2012), *Sketch Engine* (Kilgarriff, Rychly, Smrz, & Tugwell 2004) and *English Corpora Online* (Davies 2008-) having functions allowing the user to specify advanced filters to create sub-corpora of a register out of

readymade corpora by specifying the domain or other metadata. In some ways, the question of what can be done in register analysis with these tools comes down to how a corpus is compiled and what register-relevant information about the texts are available as metadata. It could be argued, however, that many of the readymade online corpora are not in the most convenient form for register analyses. Similarly, tools for working with DIY corpora may provide a range of functions for suitably prepared folders of raw texts, but comparing subsections with each-other or with subsections of reference corpora is not very easy. From this perspective, the software design needs greater consideration for the possible role of register, and perhaps the effort involved in comparing registers using these tools is one reason why variation due to register is sometimes overlooked.

One online tool which allows for more straightforward comparison of major registers is *English Corpora Online* and its interface for COCA. Since its release in 2008, it has provided a range of corpus functions to compare frequency data across its major registers: Spoken, Fiction, Popular Magazines, Newspapers and Academic Journals (Davies 2009). The sampling method of 20% from each register meant that comparisons were intuitive, and simple frequency charts and tables could be easily generated to show the frequencies of search terms across these five registers.

It can be seen that researchers now have access to a range of corpora and tools which make some aspects of register analysis more manageable. Nevertheless, with the exception of the graph data from COCA, few register-related functions seem to have been designed with less sophisticated users (such as undergraduate students) in mind. In contrast, *tPM* was initially designed for English language learning, and has been further developed to provide corpus functions for undergraduate corpus research projects. The remainder of this paper describes the features related to register for readymade and DIY corpora as well as the learning activities designed for undergraduate students using this tool at a Sino-British university in China. The features of the software and the description of the learning activities provide an example of how register awareness can be promoted in the classroom.

3. *tPM*'s Readymade Corpora

After launching *tPM* and connecting to the server, the user is presented with the main search screen for readymade corpora as shown in Figure 1. The screen provides boxes for a single query or for comparing two expressions, and suggestion boxes appear showing auto-complete, collocations, other word forms and words with similar meanings. The corpus selection menu provides access to a range of readymade corpora which are stored remotely on the server, including the BNC, some academic corpora of different disciplines, and literary and non-literary corpora from 19th Century. Having entered a query, one click retrieves concordance lines, frequency data, collocations and other summary data for patterns across the whole corpus.

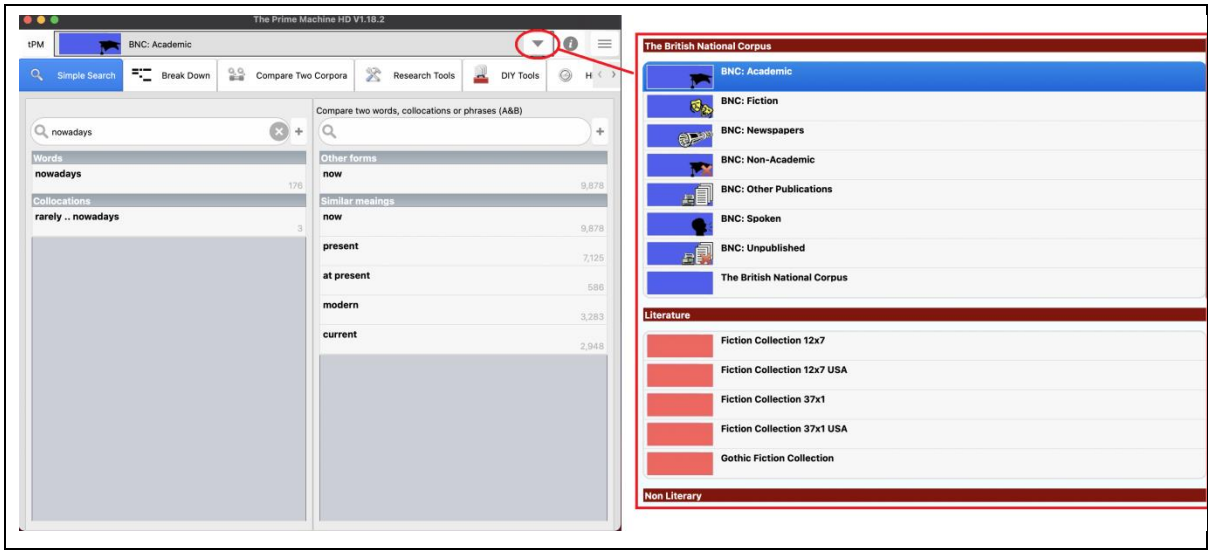

Figure 1: The main search screen with the drop-down menu for readymade corpora

When examining concordance lines from a register perspective, a less sophisticated user not only needs access to information about the source text for each hit; such information needs to be prominent and clear. Scott (2008) describes how familiarity with search engines helps students grasp more quickly how to understand the snippets from different sources which appear as concordance lines. However, because search engine users actively consider each source as a potential destination for web browsing and getting to a suitable destination was the purpose of making the query, the sense that each website listed is a separate potential source for information is much more prominent for search engines than for concordancers. In the design of the Cards and Lines views for *tPM*, the question of how best to facilitate clearer information about the source of each concordance line was considered carefully (Jeaco 2017b). As shown in Figure 2, one of the striking features of the concordance cards display is the text category and citation displayed near the top, which was specifically designed to help students become more aware of differences across registers. On its Frequency Tab, users can see the normalised frequency of the search item and they can also click a button for a chart as shown in Figures 3 and 4 where the proportion of lines occurring in the different major categories of the corpus is displayed with the proportion of words from each major category superimposed on top. The need for some kind of superimposing was because most readymade corpora in *tPM* are not like COCA in that they do not have equal portions for each major category. The arrows therefore give an indication of what would be equal distribution across all major categories of the corpus. Figure 2 shows concordance cards from a search for *nowadays* in the BNC, and Figure 3 shows the Frequency Tab graphs for its distribution across this sub-corpus' major categories.

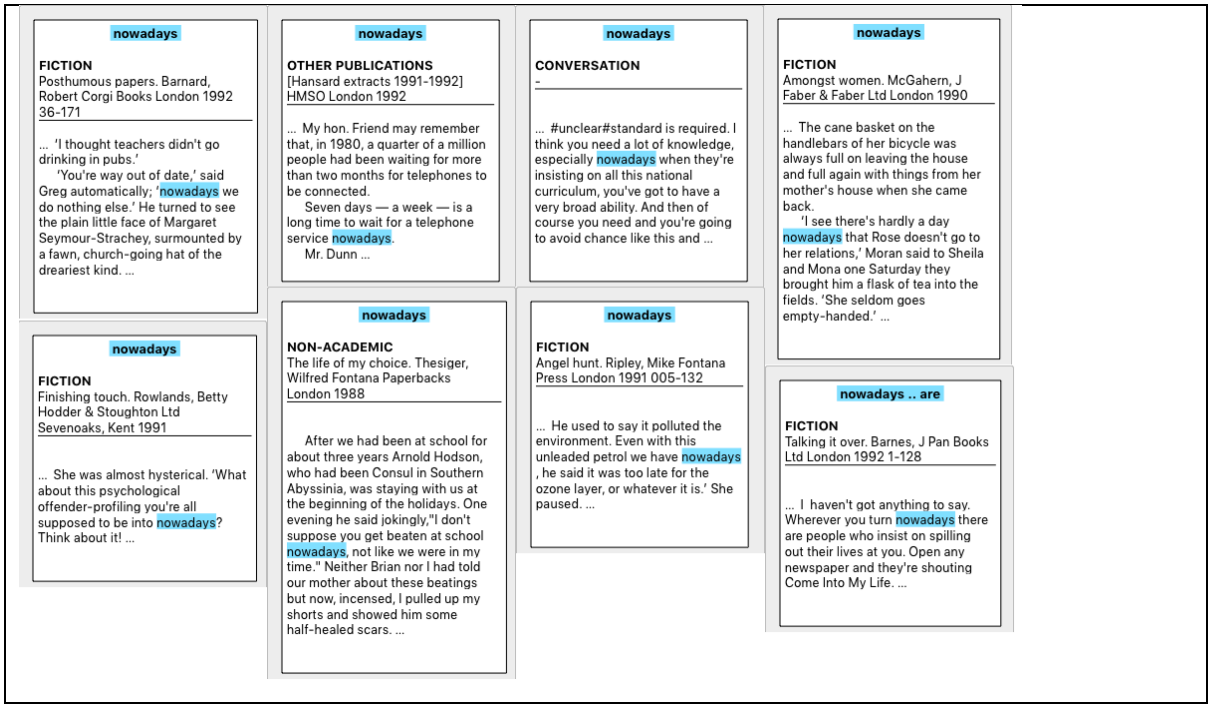

Figure 2: Concordance cards for the word *nowadays* in the BNC

Figure 3: The distribution of *nowadays* across major categories of the BNC

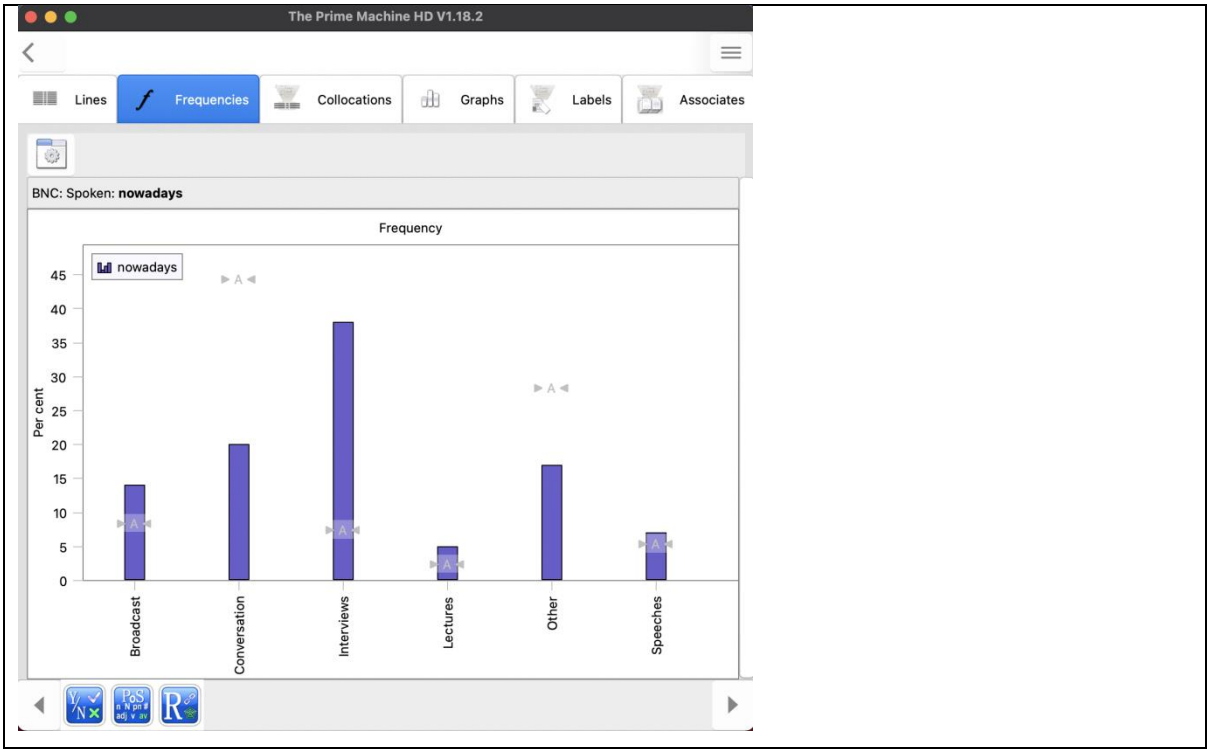

Figure 4: The distribution of *nowadays* across major categories of the BNC: Spoken Sub-Corpus

Corpora such as the BNC have additional text categories within the XML documents, such as those based on Lee's (2001) categories. The complete BNC can be accessed through *tPM*, with major categories of Academic, Conversation, Fiction, Newspapers, Non-Academic, Other Publications, Other Spoken and Unpublished. However, linguistics students are often particularly interested in one of these major categories, and *tPM* also has each of the major categories loaded as a separate sub-corpus, giving a second level of categorization. For some of these sub-corpora it seemed most appropriate to organise these major categories by topic or domain: for example with the BNC Academic sub-corpus. However the BNC Spoken sub-corpus in *tPM* is comprised of the conversation portion of the BNC and the other spoken texts as categorised by Lee (2001), showing the distribution across registers of Broadcast, Conversation, Interviews, Lectures, Speeches or other. Figure 4 shows the Frequency Tab for the distribution of *nowadays* across categories when the BNC: Spoken sub-corpus has been selected. While commerical corpus providers have extended and developed large web-harvested corpora, the approach in *tPM* demonstrates that there are some advantages in providing access to more carefully crafted text collections even if the size of the corpus is in the millions rather than the billions. Major categories for some of the corpora on *tPM* are more geared towards analysis of style or domain, with Victorian literature organized by author names and some academic corpora organised by field. However, corpora which are more register-oriented in organisation include a non-literary collection of texts from roughly the Victorian era and several sub-corpora of the BNC.

While most corpus tools allow filtering of results by metadata, *tPM* offers an easy means to divide corpora at the major text category level. However, corpora often also have many potential levels of sub-categorization if metadata can be used to group texts. As described by Jeaco (2020a) *tPM* provides a means to explore the possible importance of metadata and other labels in another way; when a search is conducted, one of the tabs of results that is returned provides clouds or tables of results known as Key Labels. The metadata and text labels shown there are *key* in the sense that if each was used as the basis for re-organising texts into sub-corpora, the search term would become a keyword. Instead of answering the question which words are key in a subcorpus divided manually by the user, it shows which text labels could be used to identify potential divisions into subcorpora in which a specific word is key. This function was developed in response to the point made by Kreyer (2008) who stated that students may not be aware of the sub-varieties contained within a corpus; they may note top level differences in register such as between spoken and written modes, but sub-corpora divisions may not be so obvious. The example he gives is where all the results in the written mode actually come from correspondence texts. An expert user may be able to see immediately that the type of writing seems to be limited to letters, but a language learner may assume that the word is equally common in all kinds of writing. Key Labels provide information about typical uses in terms of the major text categories, other text metadata as well as results from MAT (Nini 2014) which is used to pre-process all the individual texts within every readymade corpus. Figure 5 shows the text level Key Labels for *nowadays* in the BNC, indicating the associations between this phrase and the oral history interviews, as well as for texts associated with high "Involved" dimension scores.

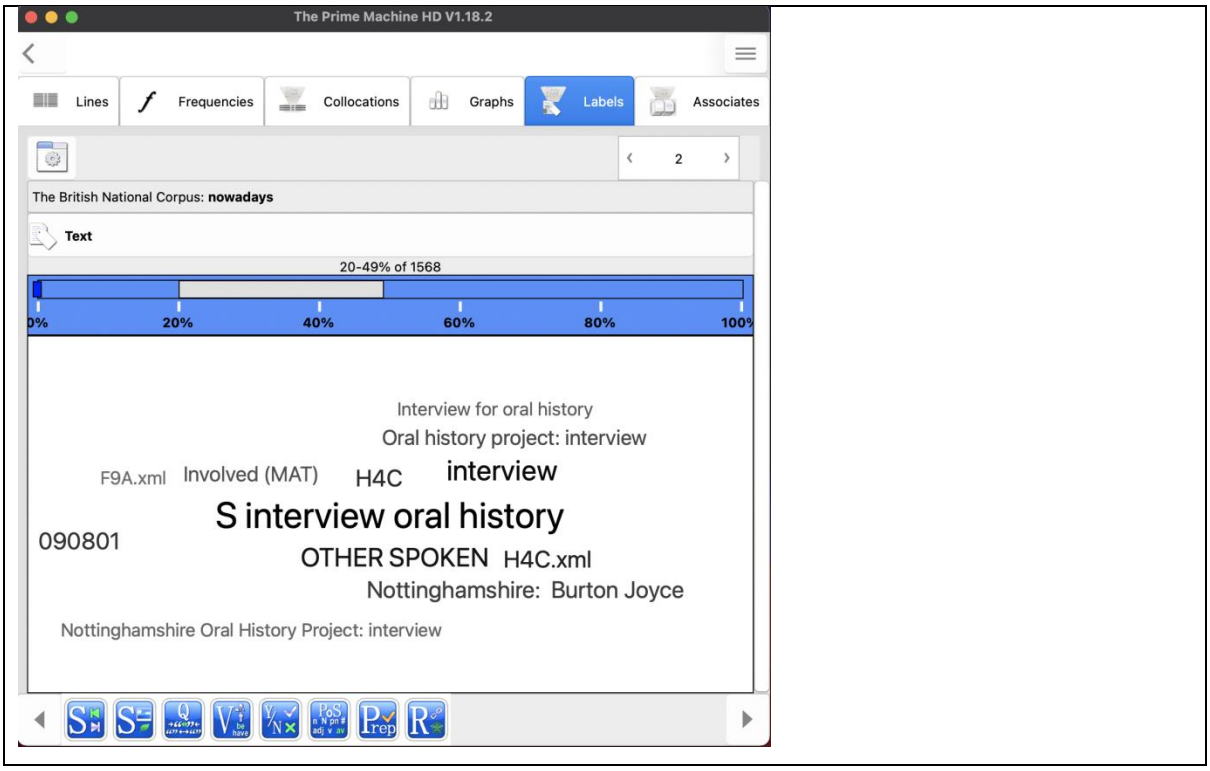

Figure 5: Key Text Labels for *nowadays* in the BNC

4. *tPM* tools for DIY corpora

tPM operates as a client-server application meaning that the application itself has a relatively small footprint and it is not necessary to download the large readymade corpora to the user's own computer. The DIY corpora, on the other hand, reside on the user's computer, having been imported through a system designed to optimise storage and retrieval of text. Data from DIY corpora can be transmitted for comparison with the server's readymade corpora. It is relatively simple, therefore, for students to import small collections of corpus texts with tokens in the tens or hundreds of thousands and then select one of the readymade corpora as reference for various operations. As well as generating concordance lines, cards and frequency data just for their own corpus, users can use the compare functions to display results from DIY and readymade corpora side by side. Keyword results and other functions inspired by *WordSmith Tools* such as Key Keywords, Key Associates and Clusters can be easily obtained using the readymade corpus as reference.

One function of the DIY tool which is particularly useful in terms of probing a DIY corpus for starting points for register analysis is the wordlist statistics function. This was designed to make stepping stones between some of the simple and intuitive information available in tools such as Lex Tutor's Vocab Profiler (Cobb 2000, 2020) for vocabulary wordlists and the wide range of register linguistic features that are measured in multidimensional analysis. Residing on the server are a number of pre-prepared wordlists covering some aspects of vocabulary profiling and features which are often strong predictors of register. Some wordlists are drawn from sources such as the Academic Word List (Coxhead 2000) while others contain simple lists such as first and second person personal pronouns. With the click of one button users are able to send the word frequencies from their own DIY corpus to the server and obtain the number of hits in each pre-prepared wordlist. Using the formula from keyword analysis, these numbers are also processed using a log-likelihood contingency table to provide a ranking for the differences between the DIY corpus and the reference corpus according to the extent to which coverage seems to be diverging from the reference corpus. For example, by building two collections of texts based on subsections of corporate annual reports, students can easily discover that a collection of texts built from Letters to Shareholders would typically have higher proportions of modal verbs and first and second personal pronouns than might be expected in comparison with typical business texts. Students wanting to look at two collections of their own texts can either compare one DIY corpus against the other, or use a readymade corpus to see how each DIY corpora differs from this common reference. Figure 6 and Figure 7 show wordlist statistics for two DIY corpora compared against one-another, and Figure 8 shows a comparison of one of these corpora against BNC: Newspapers. These examples show how the registers of movie critic

reviews differ from those of the general public, but also how some features are shared when compared to a more general newspaper corpus.

	Wordlist	Study Freq.	Study Per Thousand	Ref. Freq.	Ref. Per Thousand	Arrows	LL	Bayes
1	1st & 2nd Pers. Pronouns	1,866	38.12	490	10.21	≥ 3x ↑	829.74	Very strong evidence in favour
2	General Service List 1	34,471	704.18	30,038	626.01	↑	222.69	Very strong evidence in favour
3	Function Words	23,433	478.69	20,167	420.29	↑	183.94	Very strong evidence in favour
4	Personal Pronouns	3,861	78.87	2,716	56.60	↑	178.12	Very strong evidence in favour
5	Punctuation	3,004	61.37	2,316	48.27	↑	76.00	Very strong evidence in favour
6	Modals	538	10.99	346	7.21	↑	38.29	Very strong evidence in favour
7	Modals Subgroup 1	266	5.43	166	3.46	↑	21.40	Strong evidence in favour
8	Modals Subgroup 2	204	4.17	135	2.81	↑	12.80	Weak evidence in favour
9	Modals Subgroup 3	68	1.39	45	0.94	↑	4.27	Weak evidence in favour
10	Archaic Pronouns	0	0.00	0	0.00	=	0.00	
11	Academic Word List	1,243	25.39	1,532	31.93	↓		
12	General Service List 2	2,110	43.10	2,117	44.12	↓		
13	Positive words	779	15.91	774	16.13	↓		
14	Negative words	771	15.75	846	17.63	↓		

Figure 6: Wordlist Statistics for Customer Reviews compared against Professional Reviews

	Wordlist	Study Freq.	Study Per Thousand	Ref. Freq.	Ref. Per Thousand	Arrows	LL	Bayes
1	Academic Word List	1,532	31.93	1,243	25.39	↑	36.21	Very strong evidence in favou
2	Negative words	846	17.63	771	15.75	↑	5.14	Weak evidence in favour
3	General Service List 2	2,117	44.12	2,110	43.10	↑	0.57	Weak evidence in favour
4	Positive words	774	16.13	779	15.91	↑	0.07	Weak evidence in favour
5	Archaic Pronouns	0	0.00	0	0.00	=	0.00	
6	Modals	346	7.21	538	10.99	↓		
7	1st & 2nd Pers. Pronouns	490	10.21	1,866	38.12	≥ 3x ↓		
8	Personal Pronouns	2,716	56.60	3,861	78.87	↓		
9	Function Words	20,167	420.29	23,433	478.69	↓		
10	Modals Subgroup 3	45	0.94	68	1.39	↓		
11	Modals Subgroup 2	135	2.81	204	4.17	↓		
12	Modals Subgroup 1	166	3.46	266	5.43	↓		
13	Punctuation	2,316	48.27	3,004	61.37	↓		
14	General Service List 1	30,038	626.01	34,471	704.18	↓		

Figure 7: Wordlist Statistics for Professional Reviews compared against Customer Reviews

	Wordlist	Study Freq.	Study Per Thousand	Ref. Freq.	Ref. Per Thousand	Arrows	LL	Bayes
1	1st & 2nd Pers. Pronouns	1,866	38.12	125,026	11.57	≈ 3x ↑	1837.82	Very strong evidence in favour
2	Personal Pronouns	3,861	78.87	458,167	42.39	↑	1216.41	Very strong evidence in favour
3	Function Words	23,433	478.69	4,371,103	404.38	↑	628.02	Very strong evidence in favour
4	General Service List 1	34,471	704.18	6,682,938	618.25	↑	556.70	Very strong evidence in favour
5	Punctuation	3,004	61.37	545,390	50.45	↑	107.47	Very strong evidence in favour
6	Positive words	779	15.91	142,831	13.21	↑	25.22	Strong evidence in favour
7	Modals Subgroup 2	204	4.17	40,195	3.72	↑	2.54	Weak evidence in favour
8	Academic Word List	1,243	25.39	383,128	35.44	↓		
9	General Service List 2	2,110	43.10	561,598	51.95	↓		
10	Modals Subgroup 3	68	1.39	16,876	1.56	↓		
11	Modals	538	10.99	121,851	11.27	↓		
12	Modals Subgroup 1	266	5.43	64,502	5.97	↓		
13	Archaic Pronouns	0	0.00	74	0.01	↓		
14	Negative words	771	15.75	197,576	18.28	↓		

Figure 8: Wordlist Statistics for Customer Reviews compared against the BNC: Newspapers sub-corpus

5. Introducing register to undergraduates using tPM

In order to demonstrate the strengths of *tPM's* tools for promoting register in the classroom, a description of a course using this tool will be presented.

5.1 Teaching context

A module introducing corpus linguistics was developed at a Sino-British university in China as part of English programmes delivered in English as the Medium of Instruction (EMI). Students studying this module have typically taken English for Academic Purposes modules for eighteen months, alongside other EMI content modules. As such, they are usually well aware of some of the main differences between academic conventions in writing and less formal text types they have encountered prior to entering university. Nevertheless, they usually lack sensitivity to the rainbow of different linguistic features associated with different text types; for most students the assumption seems to be that there is this rather formal but awkward kind of language used in academic settings, while other domains offer few restrictions. Therefore, rather than focusing on historical developments of corpus linguistics or corpus linguistic concepts for their own sake, the module gives practical experience of encountering and handling language data from different registers, introducing corpus tools as a way of demonstrating linguistic differences. Almost all the students view English as a foreign language, and the module also includes aims related to analysing their own language output.

5.2 Course outline

The main software used is *tPM*, which was specifically developed for such students and offers a simple interface for searching and making comparisons. It is used to demonstrate differences between

registers and/or text types from a highly practical standpoint, while foundations for future corpus research projects are laid down via explanations about more advanced features of other software.

The assessment of the module takes the form of two coursework projects which are split into smaller guided tasks, offering students great flexibility in choosing their topic areas, but also providing ongoing support. The first assignment centres around the students' examination of their own written or spoken English, with their own writing or transcripts of their speech used as a launch platform for the selection of linguistic features for analysis. Unsurprisingly, many students choose extracts from their academic essays for at least one of these tasks, with comparisons between their own language choices and those in the BNC: Academic sub-corpus forming a major basis for their analysis. *tPM's* simple search screen allows the students to quickly start exploring words and collocations in one or more readymade corpora, and to start considering the appropriacy of specific language choices for different contexts. The tasks provide a stimulus for exploring patterns across registers though the lens of a specific communicative need in a specific context; the nature of the task demands critical thinking about how form relates to function and how different situations call for different language choices. It also provides plenty of practice using *tPM*'s search screens and becoming familiar with different corpus data results.

The second assignment requires students to build DIY corpora – either two of at least 40,000 tokens or one corpus of 80,000 tokens. From experience, these sizes offer sufficient evidence to make tentative generalizations without overwhelming students in the text pre-processing and corpus building stages. Students complete an analysis of the situational context (following Biber & Conrad 2009), and present quantitative analysis. Because *tPM* offers a good range of different registers in its readymade corpora, students can quickly generate useful results comparing their texts with readymade corpora. This allows them to concentrate more on drawing conclusions about how situational contexts influence linguistic choices and to consider possible reasons for any mismatches. For example, if they find that a corpus of Letters to Shareholders despite being produced in a setting with low interactivity seems to share many features with more interactive registers, they might consider whether the writers are trying to promote a sense of closeness or trust. As shown in Figure 9, the smaller tasks provide scaffolding to what would form the major elements of a full corpus research report. These tasks take students through the steps of (Task 1) identifying research questions (albeit tentatively); outlining the methodology and tools used; (Task 2) systematically making notes on situational contexts as they collect texts; (Task 3) generating basic corpus statistics and wordlist data; (Task 4) generating keywords; (Task 5) using concordance line evidence to describe differences in the use of one word or phrase across two corpora; (Task 6) writing up results based on other data and discussing how the results connect (or otherwise) to the situational contexts; and (Task 7) reflecting on the whole process in the form of a reflective "executive summary".

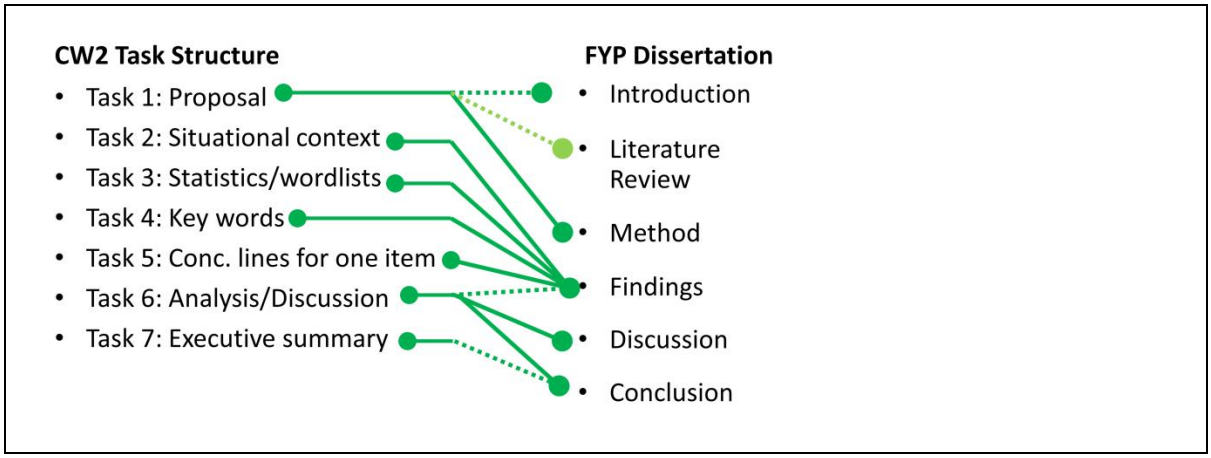

Figure 9: Possible links between coursework tasks and chapters for a dissertation project

5.3 Reflections on trouble-spots and solutions

Jeaco (2020b) reported on a favourable response from students taking this module, with reflections in student assignments and module evaluations indicating the usefulness of the approach and its potential in future research and language learning. While the overall outcome was highly positive, reflections and comments also indicate difficulties students faced. While initial fears that only some text varieties will generate meaningful results tend to dissipate as students start working with sub-corpora and discover how register is concerned with the frequent, the prevalent and the pervasive, technical issues can still present problems. In the author's own setting, students seem to be reluctant to concentrate on issues like file encoding until corpus results include strange characters or highly unexpected results. Issues with the encoding of smart quotes and issues from working on more than one machine can cause problems displaying or even searching through text. These can be rectified using *WordSmith Tools File Utilities* (Scott 2020), *EncodeAnt* (Anthony 2019b), or *tPM*'s *tPMCrafty*. With default file encoding for *WordSmith Tools*, *AntConc*, *tPM* and *MAT* in mind, *tPMCrafty* was developed to save files into two folders in two formats: UTF-16 and UTF-8. It also includes several options and filters to automatically process features like paragraphing, spacing and smart quotes. For projects looking at customer reviews or other documents which may have been obtained from one long webpage or e-book, it also includes options to split files into separate texts or chapters. Figure 10 shows a text being processed in *tPMCrafty,* where a newspaper article has been transformed for smart quotes, paragraphing and spacing.

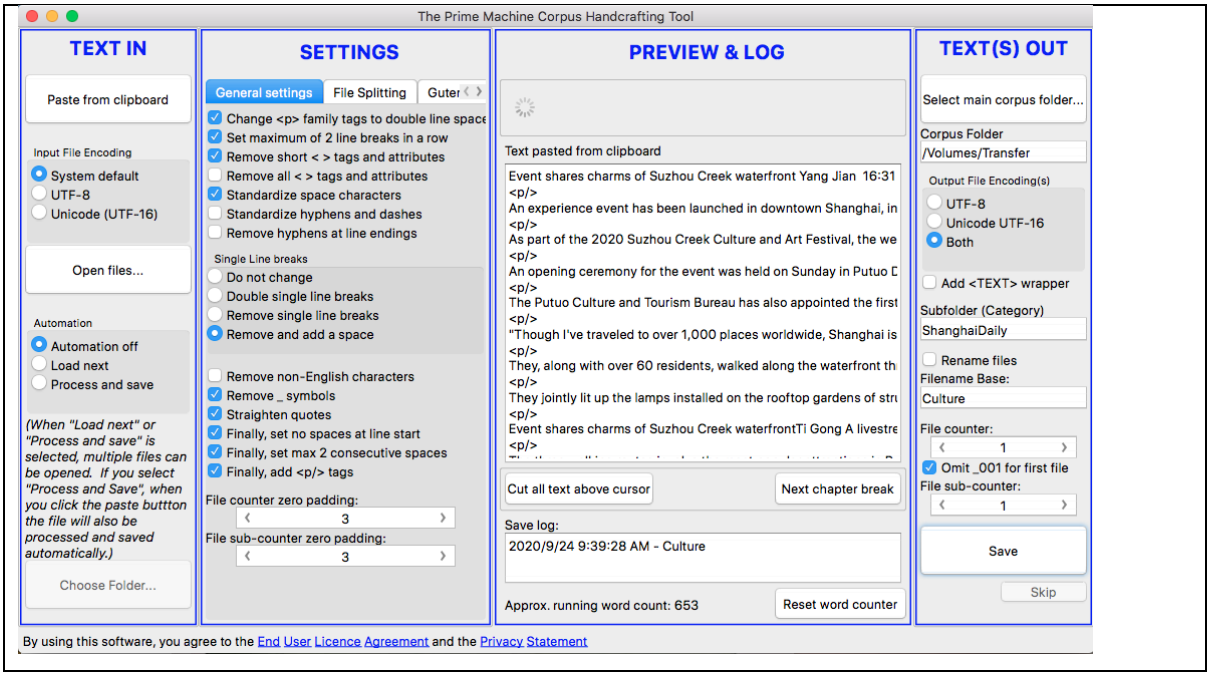

Figure 10: Corpus text processing using tPMCrafty, with incidental data from an article from the Shanghai Daily.

6. Conclusion

This paper has described how a user-friendly corpus tool can be used to help raise second language learners' awareness of the importance of variations in language use across registers. It has described some of the special features of *The Prime Machine* in terms of readymade and DIY corpora. It has also described an undergraduate linguistics module that was designed in tandem with this corpus tool, explaining how moving from analysis of students' own language choices for specific contexts can lead into more sophisticated register work. It is hoped that other linguistics lecturers and English majors will also find this tool useful either as a stepping stone to more advanced corpus work, or as a useful software application in its own right.

The Prime Machine and *tPMCrafty* are free tools and are available for Windows and Mac OSX from www.theprimemachine.net.

References

Anthony, L. (2019a). AntConc (Version 3.5.8). Tokyo, Japan: Waseda University. Retrieved from https://www.laurenceanthony.net/software/antconc/

Anthony, L. (2019b). EncodeAnt (Version 1.2.1). Tokyo, Japan: Waseda University. Retrieved from https://www.laurenceanthony.net/software/encodeant/

Berber Sardinha, T. (2017). Lexical Priming and Register Variation. In M. Pace-Sigge & K. J. Patterson (Eds.), *Lexical Priming: Applications and Advances* (pp. 189-229). Amsterdam: John Benjamins.

Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J. M. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 15-36). Amsterdam: John Benjamins.

Biber, D. (1991). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2012). Register as a Predictor of Linguistic Variation. *Corpus Linguistics and Linguistic Theory, 8*(1), 9-37. doi: 10.1515/cllt-2012-0002

Biber, D., & Conrad, S. M. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

BNC Consortium. (2007). The British National Corpus (Version 3 BNC XML ed.): Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning, 67*(2), 348-393.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics, 20*(2), 139-173.

Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes, 31*, 93-102. doi: 10.1016/j.esp.2011.12.003

Cheng, W., Warren, M., & Xu, X.-f. (2003). The language learner as language researcher: putting corpus linguistics on the timetable. *System, 31*(2), 173-186. doi: 10.1016/s0346-251x(03)00019-8

Cobb, T. (2000, 2020). The Compleat Lexical Tutor (Version 8.3), from http://www.lextutor.ca

Conrad, S. (2019). Register in English for Academic Purposes and English for Specific Purposes. *Register Studies, 1*(1), 168-198.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 600 million words, 1990-present. Retrieved 25 February, 2020, from https://www.english-corpora.org/coca/

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159-190.

Ferguson, C. (1983). Sports announcer talk: Syntactic aspects of register variation. *Language in Society, 12*(2), 153-172.

Fligelstone, S. (1993). Some reflections on the question of teaching, from a corpus linguistics perspective. *ICAME, 17*, 97-109.

Flowerdew, L. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins.

Hardie, A. (2012). CQPweb: Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics, 17*(3), 380-409.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Jeaco, S. (2017a). Concordancing Lexical Primings: The rationale and design of a user-friendly corpus tool for English language teaching and self-tutoring based on the Lexical Priming theory of language. In M. Pace-Sigge & K. J. Patterson (Eds.), *Lexical Priming: Applications and Advances* (pp. 273-296). Amsterdam: John Benjamins.

Jeaco, S. (2017b). Helping Language Learners Put Concordance Data in Context: Concordance Cards in The Prime Machine. *International Journal of Computer-Assisted Language Learning and Teaching, 7*(2), 22-39.

Jeaco, S. (2020a). Calculating and Displaying Key Labels: The texts, sections, authors and neighbourhoods where words and collocations are likely to be prominent. *Corpora, 15*(2).
Jeaco, S. (2020b). DIY needs analysis and specific text types: Using The Prime Machine to explore vocabulary in readymade and homemade English corpora. In M. Dodigovic & M. P. Agustín-Llach (Eds.), *Vocabulary in Curriculum Planning: Needs, Strategies and Tools*: Palgrave Macmillan.
Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom Concordancing* (Vol. 4, pp. 1-13). Birmingham: Centre for English Language Studies, University of Birmingham.
Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). *The Sketch Engine.* Paper presented at the 2003 International Conference on Natural Language Processing and Knowledge Engineering, Beijing.
Kreyer, R. (2008). Corpora in the classroom and beyond. In B. Barber & F. Zhang (Eds.), *Handbook of Research on Computer-Enhanced Language Acquisition and Learning* (pp. 422-437).
Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology, 5*(3), 37-72.
Mair, C. (2002). Empowering non-native speakers: the hidden surplus value of corpora in Continental English departments. In B. Kettemann, G. Marko & T. McEnery (Eds.), *Teaching and Learning by Doing Corpus Analysis* (pp. 119-130). Amsterdam: Rodopi.
Nini, A. (2014). Multidimensional Analysis Tagger 1.1 - Manual. Retrieved from http://sites.google.com/site/multidimensionaltagger
Scott, M. (2008). Developing WordSmith. *International Journal of English Studies, 8*(1), 95-106.
Scott, M. (2020). WordSmith Tools (Version 8). Oxford: Oxford University Press.