Accepted Manuscript version [submitted September 2016] of the following published title:
Fitzgerald, A., Marin. M.J., Wu, S. & Witten, I.H. (2017). Evaluating the Efficacy of the Digital Commons for Scaling Data-Driven Learning. In M. Carrier, R. M. Damerow, & K. M. Bailey (Eds.), *Digital Language Learning and Teaching: Research, Theory, and Practice* (pp. 38 – 51). New York, NY: Routledge & TIRF.

Evaluating the Efficacy of the Digital Commons for Scaling Data-Driven Learning

Alannah Fitzgerald, Maria José Marín, Shaoqun Wu, Ian H. Witten

Motivation for the Research

The Growing Digital Commons and Open Educational Resources

This chapter presents the open-source FLAX project (Flexible Language Acquisition, flax.nzdl.org), an automated digital library scheme, which has developed and tested an extraction method that identifies typical lexico-grammatical features of any word or phrase in a corpus for data-driven learning. Here in this study, FLAX will be described and discussed in relation to the reuse of openly licensed content available in the digital commons. Typically, the digital commons involves the creation and distribution of informational resources and technologies that have been designed to stay in the digital commons using various open licenses, including the GNU Public License and the Creative Commons suite of licenses (Wikipedia, 2016; see also the chapter by Stranger-Johannessen, this volume). One of the most widely used informational resources developed by and for the digital commons is Wikipedia. In response to the growing digital commons, we will provide insights into design considerations for the reuse of transcribed video lectures from Massive Open Online Courses (MOOCs) that have been licensed with Creative Commons as Open Educational Resources (OERs). We will demonstrate how OERs can be remixed with open corpora and tools in the FLAX system to support English for Specific Academic Purposes (ESAP) in classroom-based language education contexts.

This research arose largely in response to the open education movement having recently gained traction in formal higher education and in the popular press with the advent of the MOOC phenomenon. The OpenCourseWare movement, which began in the late 1990s, preceded MOOCs with the release of free teaching and learning content onto the Internet by well-known universities, most notably the Massachusetts Institute of Technology. Indeed, MOOCs are the latest in a long line of innovations in open and distance education.

This chapter also draws attention to the OER movement, where the emphasis on 'open' signifies more than freely available teaching and learning resources for philanthropic purposes (open gratis). Here, we focus on the truly open affordance of flexible and customizable resources that can be retained, revised, repurposed, remixed, and redistributed by multiple stakeholders for educational purposes (open libre). In the present research with the FLAX project, open resources are specifically employed in the design and development of domain-specific language corpora for scaling *data-driven learning* (DDL, discussed below) approaches across informal MOOCs and formal language learning classrooms.

The mainstreaming of open content, including OERs and open access publications, came swiftly on the back of the development of the Creative Commons suite of licenses by copyright lawyer, Larry Lessig, in collaboration with Internet activist and open education advocate, Aaron Swartz. Their collaboration resulted in six Creative Commons licenses that were released in 2002 to retain the copyright of authors for enabling 'Some Rights Reserved' in a movement away from the default 'All Rights Reserved' restrictions of licensed creations. An estimated one billion Creative Commons-licensed works now reside in the digital commons (Creative Commons, 2015). This growing movement provides evidence that the read-only culture of analogue content developed by commercial publishers and broadcasters for passive consumers is being eclipsed by the read-write digital culture of remix, with an increasing number of active creators electing to share content online with free culture licenses (Lessig, 2008). According to Wiley (n.d.), Creative Commons licenses enable the following permissions to the education community by means of defining the affordances of OERs:

1. Retain: the right to make, own, and control copies of the content (e.g., download, duplicate, store, and manage).

2. Reuse: the right to use the content in a wide range of ways (e.g., in a class, in a study group, on a website, in a video).
3. Revise: the right to adapt, adjust, modify, or alter the content itself (e.g., translate the content into another language).
4. Remix: the right to combine the original or revised content with other open content to create something new (e.g., incorporate the content into a mash-up).
5. Redistribute: the right to share copies of the original content, your revisions, or your remixes with others (e.g., give a copy of the content to a friend). (Wiley, n.d.)

Open Data-driven Learning Systems in Specialized Language Education

Concerning the use of corpus-based language teaching materials in language instruction, Tim Johns is often regarded as the pioneer in the field, coining the term DDL to refer to the method of inferring the rules of language by directly observing them in corpora using text analysis tools. He affirmed that by discovering the rules of language underlying real samples extracted from corpora, learners become “language detectives” (Johns, 1997, p. 101). The term *DDL* was revisited by Boulton (2011), who considers Johns’ definition of DDL as too broad to be systematized. Boulton also offers some of the most comprehensive overviews of research carried out in DDL and identifies the number of experiments in the field of legal English as quite reduced (Boulton, 2011).

This identifiable lack was a motivating factor for conducting the experiment described below in response to the following research questions. They arose from the planning, implementation, and analysis of the data obtained from our experiment:

1. To what extent can the digital commons of open and authentic content enrich data-driven learning across formal and informal language learning?

2. What effect does the application of DDL methods for querying open and authentic content have on the acquisition of specialized terminology, as opposed to accessing non-DDL-based online resources?

Throughout this chapter, we will refer to the *Law Collections* in FLAX, which are derived from openly licensed pedagogic texts and open access publications from law education and research, along with legal code and judicial hearings from case law available in the public domain. In the area of legal English, as with many areas of ESAP, corpora and published language learning resources are too scarce, too small, too generic, and in most cases inaccessible due to licensing restrictions or cost. The *Law Collections* in FLAX demonstrate the potential for engagement with diverse higher education audiences by drawing attention to the growing digital commons of openly available and high quality authentic texts, which can be mined by DDL approaches to render them linguistically accessible, discoverable, and adaptable for further remixing in ESAP education.

This inquiry is directly concerned with the scalability of DDL applications and their potential to take root across both informal online learning and formal classroom-based language learning. (See the de Groot chapter, this volume.) We also contend that our open research and development methodology enables critique by relevant stakeholders within the fields of language education, applied corpus linguistics, and now open and distance education.

Digital Tools Used in this Study

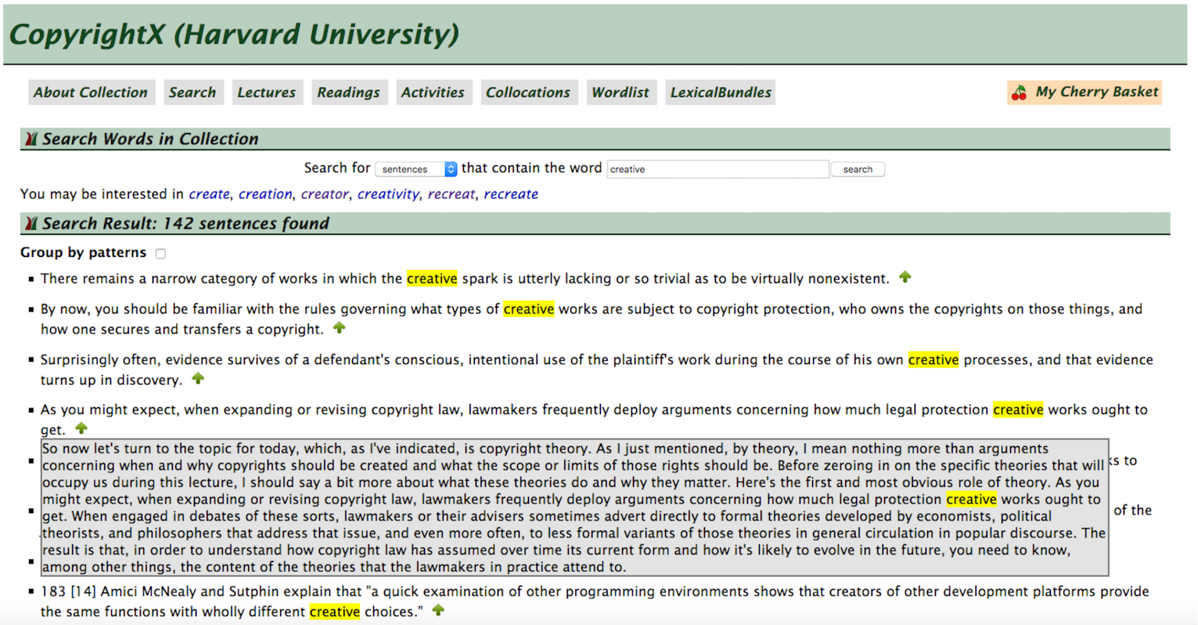
Transcending Concordance: Augmenting Academic Text with FLAX

Many language learners consult concordancers. Although successful outcomes are widely reported, learners face challenges when using such tools to seek lexico-grammatical patterns. Concordancers are popular tools for supporting language learning. They allow learners to access, analyze, and discover linguistic patterns in a particular corpus, which can be chosen to match the task at hand. Researchers report positive responses from students using concordance data for checking grammatical errors, seeking vocabulary usage, and retrieving collocations (Gaskell & Cobb, 2004; O'Sullivan & Chambers, 2006; Varley, 2009; Yoon & Hirvela, 2004).

However, these tools were originally designed for linguistic analysis by professionals, and not all their facilities can be easily navigated and investigated by language learners. Learners are often overwhelmed by the vast amount of data returned. The presentation of concordance lines appears random, with no discernable ordering. It is challenging and time-consuming to go through lines of text to identify patterns. Learners may pick up a rare exceptional case for a rule and over-generalize it. Advanced search options, for example, seeking the verb collocates of a word, are sometimes provided but expressed in a syntax that requires specialized knowledge and varies among concordance providers.

Some researchers suggest that concordance data be screened before being presented to students (Varley, 2009). Others ask for commonly used linguistic patterns to be made more accessible (Coxhead & Byrd, 2007), perhaps through a simple interface for retrieving collocations (Chen, 2011). Consequently, the tool described in this chapter was conceived as a solution to these shortcomings, making it easier for language learners to seek language patterns by going far beyond simply returning concordance lines. The FLAX system supports the following functions and presents a design departure from traditional concordancer interfaces for (1) checking vocabulary usage, (2) seeking grammatical patterns, (3) looking up collocations and lexical bundles, and (4) glossing and augmenting full-text documents with additional open and multi-media resources.

By way of introduction, FLAX is an automated scheme that extracts salient linguistic features from text and presents them in an interface designed specifically for language learners. An extraction method was developed to build the *Law Collections*, which identifies typical lexico-grammatical features of any word or phrase in the corpora. For example, as shown in Figure 1, learners can search at the article, paragraph, sentential, or collocational level, highlighting search terms in color. Clicking on the color arrows at the end of the sentences enables learners to move up a resource granularity level, for example, to the paragraph level, to enable the inspection of search terms along with their contextual information.



Search Words in Collection

Search for sentences that contain the word creative search

You may be interested in *create*, *creation*, *creator*, *creativity*, *recreat*, *recreate*

Search Result: 142 sentences found

Group by patterns □

- There remains a narrow category of works in which the creative spark is utterly lacking or so trivial as to be virtually nonexistent.
- By now, you should be familiar with the rules governing what types of creative works are subject to copyright protection, who owns the copyrights on those things, and how one secures and transfers a copyright.
- Surprisingly often, evidence survives of a defendant's conscious, intentional use of the plaintiff's work during the course of his own creative processes, and that evidence turns up in discovery.
- As you might expect, when expanding or revising copyright law, lawmakers frequently deploy arguments concerning how much legal protection creative works ought to get.

So now let's turn to the topic for today, which, as I've indicated, is copyright theory. As I just mentioned, by theory, I mean nothing more than arguments concerning when and why copyrights should be created and what the scope or limits of those rights should be. Before zeroing in on the specific theories that will occupy us during this lecture, I should say a bit more about what these theories do and why they matter. Here's the first and most obvious role of theory. As you might expect, when expanding or revising copyright law, lawmakers frequently deploy arguments concerning how much legal protection creative works ought to get. When engaged in debates of these sorts, lawmakers or their advisers sometimes advert directly to formal theories developed by economists, political theorists, and philosophers that address that issue, and even more often, to less formal variants of those theories in general circulation in popular discourse. The result is that, in order to understand how copyright law has assumed over time its current form and how it's likely to evolve in the future, you need to know, among other things, the content of the theories that the lawmakers in practice attend to.

ks to

of the

- 183 [14] Amici McNealy and Sutphin explain that "a quick examination of other programming environments shows that creators of other development platforms provide the same functions with wholly different creative choices."

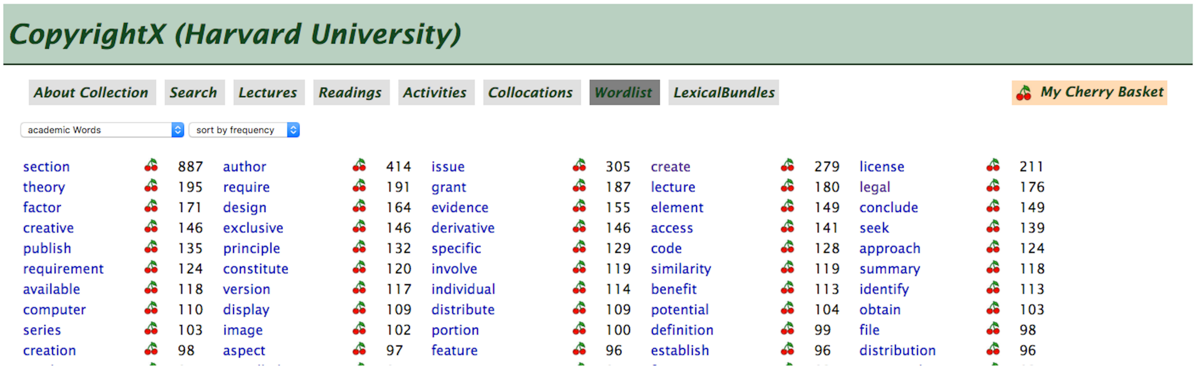
Figure 1. Keyword search for *creative* in the *CopyrightX* collection

FLAX first facilitates the retrieval of typical words or phrases by grouping concordance data and sorting search results to show the most common patterns first. Second, it incorporates grammar rules involving prepositions, word inflection, and articles, and it makes common patterns stand out. Third, it retrieves collocations and lexical bundles according to part-of-speech tags—for example, all adjectives associated with a particular noun—without using any special syntax. Fourth, it links texts to larger corpora, such as the *Learning Collocations* collection in FLAX and Wikipedia to provide further examples of collocates and to gloss key terms. FLAX is available on the web for anyone to use. Its design, with regard to the *Law Collections* in FLAX, is illustrated in this chapter. However, this method can be applied to any specialized corpus, including samples of writing collected by an individual teacher (provided they are available electronically for reuse) or writing completed by students.

Research on Academic Text

Academic text has considerable value for supporting ESAP, and many pedagogical implications have arisen from studies of academic corpora. Although specificity in academic text has received much attention in the research literature, the findings have not been fully exploited in language teaching and learning practice. Suggestions from the research literature, for example, for bridging the gap between expert and novice academic English language proficiency include helping students appreciate the importance of common collocates and recurring lexical and grammatical patterns in different contexts (Coxhead & Byrd, 2007), making commonly used lexical bundles more accessible (Hafner & Candlin, 2007), and providing more realistic models for students (Hyland, 2008a). Emphasis in this study has therefore been placed on supporting the acquisition of specialized terminology from academic text. Also highlighted in this research, are the affordances of open and authentic texts for increased uptake by practitioners in the design and application of DDL methods in teaching and language materials development, for imparting the learning of specialized terminology in ESAP.

Words and wordlists. Great emphasis has been placed on identifying the language features of academic texts. Coxhead (2000) developed the Academic Word List (AWL), a list of 570 academic word families from a 3.5 million-word corpus of academic writing, which has become a widely used resource for teachers and students. Computer tools, such as the Vocab profiler available at the Compleat Lexical Tutor website, help teachers and learners analyze the vocabulary in a text with reference to the AWL and other wordlists. Certainly, learning vocabulary involves far more than simply memorizing words in lists or looking them up in dictionaries. Users can explore the most frequent one to two thousand words from the general service list, and academic words from the AWL. Clicking the *Wordlist* tab in the *CopyrightX* collection menu, as shown in Figure 2, yields the different wordlist options.



CopyrightX (Harvard University)

academic Words | sort by frequency

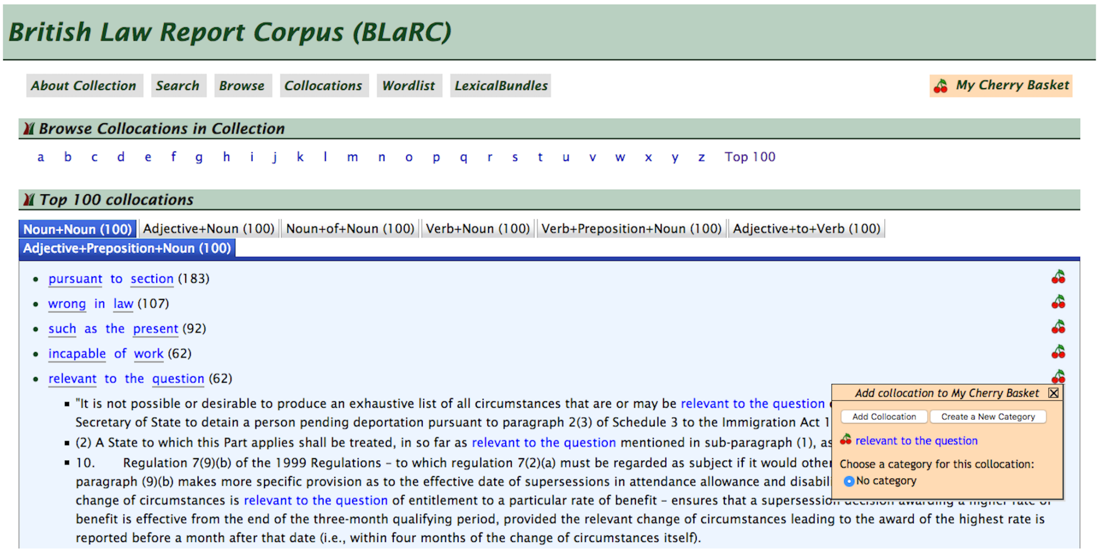
section	🍒	887	author	🍒	414	issue	🍒	305	create	🍒	279	license	🍒	211
theory	🍒	195	require	🍒	191	grant	🍒	187	lecture	🍒	180	legal	🍒	176
factor	🍒	171	design	🍒	164	evidence	🍒	155	element	🍒	149	conclude	🍒	149
creative	🍒	146	exclusive	🍒	146	derivative	🍒	146	access	🍒	141	seek	🍒	139
publish	🍒	135	principle	🍒	132	specific	🍒	129	code	🍒	128	approach	🍒	124
requirement	🍒	124	constitute	🍒	120	involve	🍒	119	similarity	🍒	119	summary	🍒	118
available	🍒	118	version	🍒	117	individual	🍒	114	benefit	🍒	113	identify	🍒	113
computer	🍒	110	display	🍒	109	distribute	🍒	109	potential	🍒	104	obtain	🍒	103
series	🍒	103	image	🍒	102	portion	🍒	100	definition	🍒	99	file	🍒	98
creation	🍒	98	aspect	🍒	97	feature	🍒	96	establish	🍒	96	distribution	🍒	96

Figure 2. Most frequent Academic Wordlist items in the *CopyrightX* collection

Collocations. The importance of collocation knowledge in academic writing has been widely recognized. Hill (1999) observes that students with good ideas often lose marks because they do not know the four or five most important collocations of a keyword that is central to what they are writing about. Topic-specific corpora are therefore valuable resources that help learners build up collocation knowledge within the areas that concern them.

With FLAX, learners can browse as well as search collocations. Figure 3 shows some of the Top 100 collocations in the *British Law Reports Corpus (BLaRC)* to enable ready identification of useful patterns in the corpus by users. They are grouped under tabs that reflect the syntactic roles of the associated word or words, of which the first four can be seen here grouped under the "Adjective + Preposition + Noun" tab, along with their contexts. The "cherries" icon links to the collocations associated with particular a word, enabling learners to harvest and save collocations to "My Cherry Basket".

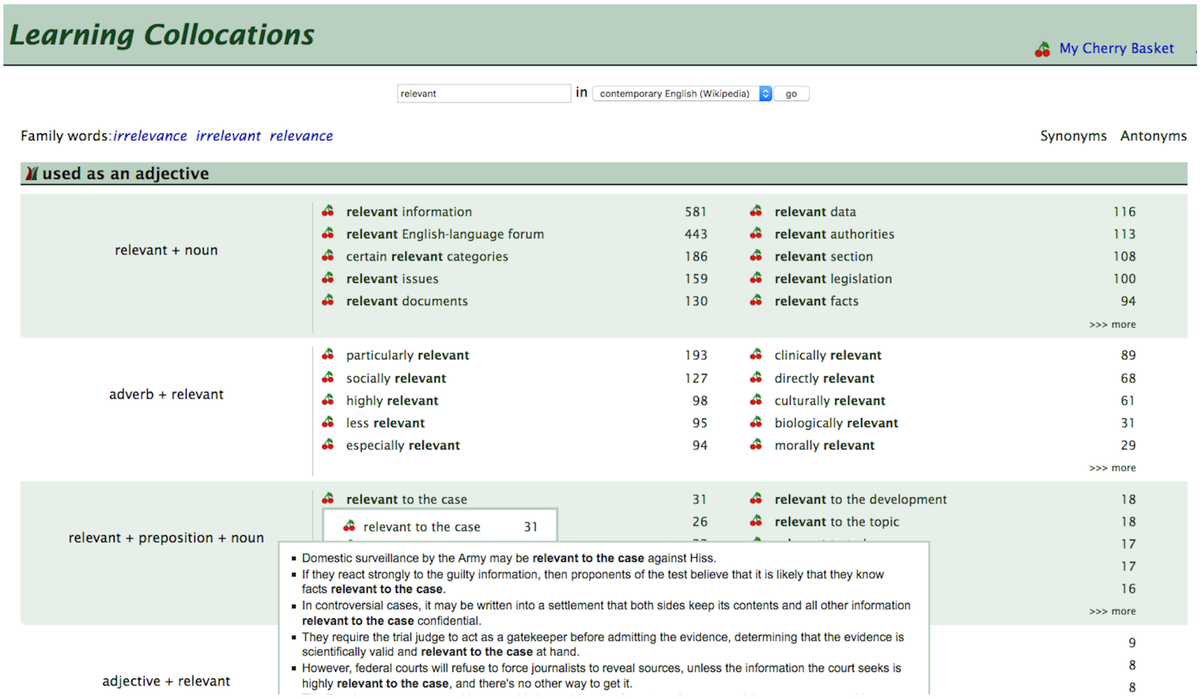
The underlined words in Figure 3, for example *relevant to the question*, are also hyperlinked to entries for those words in an external collocation database built from all the written texts in the *British National Corpus (BNC)*. For example, clicking *relevant* in Figure 3 generates a further popup, shown in Figure 4, that lists *relevant to the case*, *relevant to the* *needs*, *relevant to the study*, etc., along with their frequency in that corpus. Furthermore, samples of these collocations in context can be seen by clicking on them in Figure 4, which displays relevant extracts from a choice of three corpora in the FLAX *Learning Collocations* collection: the *BNC*, the *British Academic Written English* corpus, and a Wikipedia corpus. For example, clicking *relevant to the study* brings up 22 sentences that use this phrase.


British Law Report Corpus (BLaRC)
About Collection
Search
Browse
Collocations
Wordlist
LexicalBundles
My Cherry Basket
Browse Collocations in Collection
a b c d e f g h i j k l m n o p q r s t u v w x y z Top 100
Top 100 collocations
Noun+Noun (100)
Adjective+Noun (100)
Noun+of+Noun (100)
Verb+Noun (100)
Verb+Preposition+Noun (100)
Adjective+to+Verb (100)
Adjective+Preposition+Noun (100)
Add collocation to My Cherry Basket ☒
Add Collocation
Create a New Category
Choose a category for this collocation:
● No category

- pursuant to section (183)
- wrong in law (107)
- such as the present (92)
- incapable of work (62)
- relevant to the question (62)
  - "It is not possible or desirable to produce an exhaustive list of all circumstances that are or may be relevant to the question Secretary of State to detain a person pending deportation pursuant to paragraph 2(3) of Schedule 3 to the Immigration Act 1
  - (2) A State to which this Part applies shall be treated, in so far as relevant to the question mentioned in sub-paragraph (1), as
  - 10. Regulation 7(9)(b) of the 1999 Regulations – to which regulation 7(2)(a) must be regarded as subject if it would other paragraph (9)(b) makes more specific provision as to the effective date of supersessions in attendance allowance and disabil change of circumstances is relevant to the question of entitlement to a particular rate of benefit – ensures that a supersessio benefit is effective from the end of the three-month qualifying period, provided the relevant change of circumstances leading to the award of the highest rate is reported before a month after that date (i.e., within four months of the change of circumstances itself).

relevant to the question

Figure 3. Preview of some of the top 100 collocations in the *British Law Report Corpus (BLaRC)* displaying *relevant to the question*



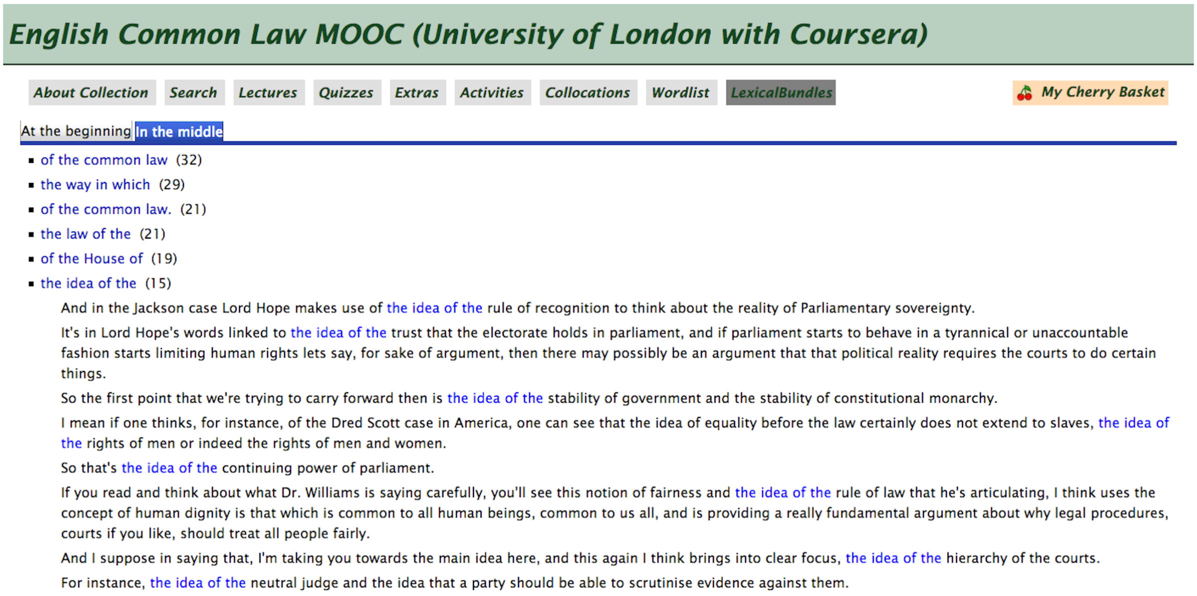
relevant + noun	**relevant** information	581	**relevant** data	116
	relevant English-language forum	443	**relevant** authorities	113
	certain **relevant** categories	186	**relevant** section	108
	relevant issues	159	**relevant** legislation	100
	relevant documents	130	**relevant** facts	94
			>>> more	
adverb + relevant	particularly **relevant**	193	clinically **relevant**	89
	socially **relevant**	127	directly **relevant**	68
	highly **relevant**	98	culturally **relevant**	61
	less **relevant**	95	biologically **relevant**	31
	especially **relevant**	94	morally **relevant**	29
			>>> more	
relevant + preposition + noun	**relevant** to the case	31	**relevant** to the development	18
		26	**relevant** to the topic	18
				17
				17
				16
			>>> more	
adjective + relevant				9
				8
				8

relevant to the case 31

- Domestic surveillance by the Army may be **relevant to the case** against Hiss.
- If they react strongly to the guilty information, then proponents of the test believe that it is likely that they know facts **relevant to the case**.
- In controversial cases, it may be written into a settlement that both sides keep its contents and all other information **relevant to the case** confidential.
- They require the trial judge to act as a gatekeeper before admitting the evidence, determining that the evidence is scientifically valid and **relevant to the case** at hand.
- However, federal courts will refuse to force journalists to reveal sources, unless the information the court seeks is highly **relevant to the case**, and there's no other way to get it.

Figure 4. Related collocations for the word *relevant* linked in from the *Learning Collocations* collection

Lexical bundles. To become proficient in ESAP, learners need to develop a repertoire of discipline-specific phrases. Recently, Biber and his colleagues developed the notion of "lexical bundles," which are multi-word sequences with distinctive syntactic patterns and discourse functions commonly used in academic prose (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004). Typical patterns include noun phrase + *of* (*the end of the*, *the idea of the*, as shown in Figure 5), prepositional phrase + *of* (*as a result of, as a part of*), *it* + verb/adjective phrase (*it is possible to, it is necessary to*), *be* + noun/adjective phrase (*is one of the*, *is due to the*), and verb phrase + *that* (*can be seen that, studies have shown that*). Such phrases fulfill discourse functions such as referential expression (framing, quantifying, and place/time/text-deictic), stance indicators (epistemic, directive, ability) and discourse organization (topic introduction/elaboration, inference, and identification). Hyland's (2008b) follow-up study compared the most frequent 50 four-word bundles in texts on biology, electrical engineering, applied linguistics, and business studies, and discovered substantial variation between the disciplines. This variation suggests the need for learners to understand relevant discourse features in their subject domains.



- of the common law (32)
- the way in which (29)
- of the common law. (21)
- the law of the (21)
- of the House of (19)
- the idea of the (15)

And in the Jackson case Lord Hope makes use of the idea of the rule of recognition to think about the reality of Parliamentary sovereignty.

It's in Lord Hope's words linked to the idea of the trust that the electorate holds in parliament, and if parliament starts to behave in a tyrannical or unaccountable fashion starts limiting human rights lets say, for sake of argument, then there may possibly be an argument that that political reality requires the courts to do certain things.

So the first point that we're trying to carry forward then is the idea of the stability of government and the stability of constitutional monarchy.

I mean if one thinks, for instance, of the Dred Scott case in America, one can see that the idea of equality before the law certainly does not extend to slaves, the idea of the rights of men or indeed the rights of men and women.

So that's the idea of the continuing power of parliament.

If you read and think about what Dr. Williams is saying carefully, you'll see this notion of fairness and the idea of the rule of law that he's articulating, I think uses the concept of human dignity is that which is common to all human beings, common to us all, and is providing a really fundamental argument about why legal procedures, courts if you like, should treat all people fairly.

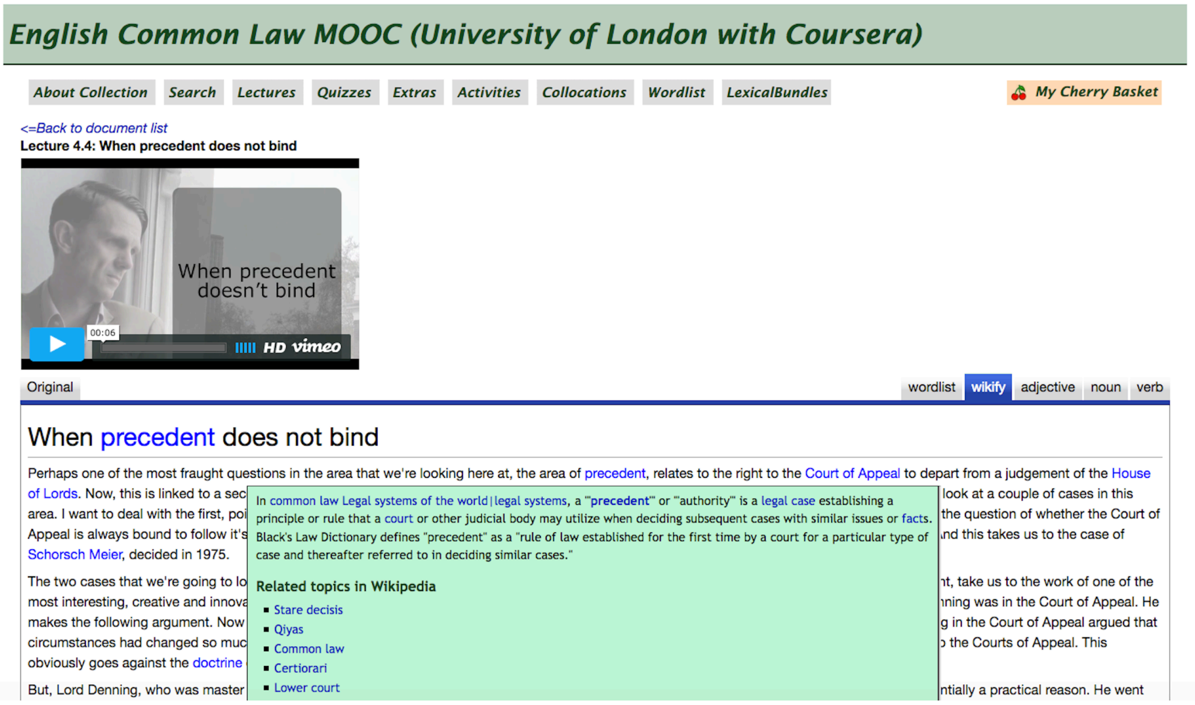
And I suppose in saying that, I'm taking you towards the main idea here, and this again I think brings into clear focus, the idea of the hierarchy of the courts.

For instance, the idea of the neutral judge and the idea that a party should be able to scrutinise evidence against them.

Figure 5. Lexical Bundles feature in the *English Common Law MOOC* collection

Augmenting text with Wikification. FLAX also interfaces with the Wikipedia Miner tool (Milne & Witten, 2013) to extract key concepts and their definitions from Wikipedia articles. Wikification in FLAX acts as a glossary tool for learners, promoting reading and vocabulary acquisition in domain-specific areas, as seen in Figure 6 with the wikify function.

The wikification process goes as follows. First, sequences of words in the text that may correspond with Wikipedia articles are identified using the names of the articles, as well as their redirects and every referring anchor text used anywhere in Wikipedia. Second, situations where multiple articles correspond to a single word or phrase are disambiguated. Third, the most salient linked (and disambiguated) concepts are selected to include in the output. For example, *Stare decisis, Qiyas, Common law, Certiorari,* and *Lower court* in the lecture document in Figure 6 are identified in FLAX as Wikipedia concepts. A definition for *precedent* is also extracted by the Wikipedia Miner, as shown in Figure 6, within the *English Common Law MOOC* collection.



Lecture 4.4: When precedent does not bind

When precedent does not bind

Perhaps one of the most fraught questions in the area that we're looking here at, the area of precedent, relates to the right to the Court of Appeal to depart from a judgement of the House
of Lords. Now, this is linked to a sec | look at a couple of cases in this
area. I want to deal with the first, poi | the question of whether the Court of
Appeal is always bound to follow it's | nd this takes us to the case of
Schorsch Meier, decided in 1975.

The two cases that we're going to lo | nt, take us to the work of one of the
most interesting, creative and innova | nning was in the Court of Appeal. He
makes the following argument. Now | g in the Court of Appeal argued that
circumstances had changed so muc | the Courts of Appeal. This
obviously goes against the doctrine

But, Lord Denning, who was master | ntially a practical reason. He went

In common law Legal systems of the world|legal systems, a '"precedent'" or '"authority'" is a legal case establishing a principle or rule that a court or other judicial body may utilize when deciding subsequent cases with similar issues or facts. Black's Law Dictionary defines "precedent" as a "rule of law established for the first time by a court for a particular type of case and thereafter referred to in deciding similar cases."

Related topics in Wikipedia

- Stare decisis
- Qiyas
- Common law
- Certiorari
- Lower court

Figure 6. Wikify feature in the *English Common Law MOOC* collection

Data Collection Procedure

The experiment described herein was conceived as a method to measure quantitatively the usefulness and effectiveness of employing a corpus-based online learning platform, FLAX, in the teaching of legal English. To that end, a group of 52 students in the fourth year of the Translation Degree program at the University of Murcia (Spain) were selected as informants. All the students' linguistic competence level complied with the Common European Framework of Reference for Languages requirements for the B2 level. Our initial intention was to incorporate FLAX as part of the course methodology itself, trying not to alter the original syllabus of the subject in its essence.

The informants were asked to write an essay on a given set of legal English topics, defined by the subject instructor as part of their final assessment. They were then divided into two groups. The experimental group (16 informants organized into four sub-groups) were requested to only consult the FLAX *English Common Law MOOC* collection as the single source of information to draft their essays. The remaining 36 students (divided into nine different sub-groups) would act as the control group, following the traditional method for the design and drafting of essays before this experiment was carried out, that is, using any information source available.

The students' essays provided the database for two small learner corpora. The difference in the number of students in the control and experimental groups resulted from the fact that only two-thirds of the essay topics suggested by the subject instructor prior to the experiment were covered by the content of the *English Common Law MOOC* collection in FLAX.

Analysis and Discussion of Findings

The quantitative analysis of the two corpora yielded results which reinforced our belief that the use of a corpus-based learning platform like FLAX may be a good methodological choice for the legal English instructor to complement more traditional teaching methods employed in the ESAP classroom.

Corpora Description and Methods of Analysis

Once the essays were completed, they were divided into two small learner corpora whose size differed considerably. The FLAX-based corpus contained 16,939 tokens, while those texts not based on consulting FLAX amounted to 55,030. The term "type" refers to every different word in a corpus, whereas "token" stands for the number of repetitions of the same word within it. The former corpus was articulated into four texts, whereas the latter comprised nine. (Each of these texts corresponds with the essays assigned to the experimental and control groups respectively.) Both corpora were processed automatically using Scott's (2008) *Wordsmith Tools* software, with the aim of extracting information that could allow us to measure the degree of effectiveness in the use of FLAX as an experimental learning method. The texts were analyzed quantitatively by applying Corpus Linguistics techniques for the exploration of the lexical level of the language, focusing on specialized term usage.

Results and Discussion: Specialized Term Usage

On a lexical level, the parameter that was measured as part of the quantitative analysis was term usage. To that end, both corpora were analyzed using Drouin's (2003) *TermoStat,* an online Automatic Term Recognition method (ATR henceforth). According to Marín (2014), this method turned out to be the most efficient method in the extraction of legal terms from an 8.85 million-word legal corpus, the *BLaRC*, reaching a peak precision rate of 88% for the top 200 candidate terms. Automatic identification of terms from the *BLaRC* employing the ATR method confirmed them as true terms after comparing them with a legal English glossary.

TermoStat mined 226 specialized terms from the learner corpus based in FLAX and 405 from those texts not using FLAX as reference. The difference in size between the two corpora, and the fact that the number of topics covered by the non-FLAX based corpus was twice as big as the other corpus, led to a size reduction of the former corpus (non-FLAX) with the aim of making the results comparable. Applying a normalization procedure such as dividing the number of terms by the number of tokens in each corpus would have sufficed for the comparison. However, the greater number of topics in the non-FLAX corpus would have caused the results to be skewed. The higher the number of different topics in a specialized corpus (as illustrated by Table 1), the higher the number of technical terms employed in it (there are more areas and sub-areas to be covered). Therefore, this variable also had to be taken into consideration in the calculations applied in each case. In order to try to compensate for that fact, the results were divided by the number of topics, four for the FLAX texts and nine for the non-FLAX ones.

As Table 1 shows, the term average obtained for those essays written using FLAX as a resource was 13.73 points higher than the same parameter for the non-FLAX-based corpus. It could therefore be argued that those students resorting to the FLAX *English Common Law MOOC* collection as an information source for the drafting of their essays displayed a greater command in the use of legal terms than those who did not. The different possibilities offered by the platform, such as the "wikify" option (allowing search for definitions or related topics to a given term) or the activities aimed at fostering the acquisition of specialized terminology, may have contributed to the greater command of employing legal terms by the experimental group.

	FLAX Corpus	Non-FLAX Corpus
Terms Identified by *Themostat* (A) (Drouin, 2003)	226	385
Corpus Size After Reduction	16,939	16,264
Number of Topics (B)	4	9
Term Average (A/B)	56.5	42.77
Standardized type/token ratio	35.3	38.63

Table 1: Term Average in Each Corpus

Furthermore, Drouin's (2003) ATR method allows for the ranking of terms according to their level of specialization, which is calculated using such values as term frequency or distribution in the general and specialized fields. The average value of this parameter also turned out to be higher for the FLAX-based corpus, reaching 14.68 against 13.37 for the non-FLAX text collection. This difference could be interpreted as a greater capacity on the part of the experimental group to express themselves more accurately through more specific terms than those in the control group. However, the difference is not substantial enough for us to be able to state this conclusion with absolute certainty. Therefore, a larger sample would thus be required to confirm our observations. Furthermore, a qualitative study of a corpus sample (instead of an automatic analysis of the whole text collection) —examining text excerpts with regard to term usage—would also be helpful to reinforce this perception.

According to the data, the members of the experimental group appear to have acquired the specialized terminology of the area better than those in the control group, as attested by the higher term average obtained by the texts in the FLAX-based corpus (56.5) as opposed to the non-FLAX-based text collection, at 13.73 points below (see Table 1). This result goes some way toward answering our second research question on the effect of DDL methods using open and authentic content on the acquisition of specialized terminology, as opposed to using non-DDL-based online resources. Employing Drouin's (2003) *TermoStat* ATR method as a reference, the terms identified in the former corpus are more specialized than those in the latter; that is, they are assigned a higher specificity average value based on such data as their frequency or distribution.

However, the standardized type/token ratio assigned to each set of texts, which is often indicative of the richness of the vocabulary (the higher, the richer), is lower for the FLAX-based texts, standing at 3 points below the texts written by the control group (as shown in table 1. Although the difference is not substantial, the proportion of different types is greater in the latter corpus and hence the greater diversity of its lexicon.

Policy Implications

Formal language teacher qualifications are still predominantly concerned with training teachers in how to adapt authentic linguistic content for classroom use with minimal attention to copyright and licensing. This training extends to the adaptation of All Rights Reserved proprietary language course books and their free supplementary resources, also intended for classroom use. A notable gap in formal language teacher education arises, however, when teachers wish to share their teaching materials, which they have developed using third-party content, on the Internet beyond the secret garden of the classroom. This gap in formal teacher education also extends to developing and sharing language corpora on the Internet where issues around copyright infringement and enforcement are more likely to arise than in schools.

Policy implications for language teacher education include the need for increased awareness of the digital commons and open licensing for developing digital literacies in online language materials development and distribution. Imparting understanding of the difference between free proprietary resources and OERs licensed with Creative Commons that afford reuse and remix is also essential for redressing the current shortfall in formal language teacher training where understanding copyright is concerned. Indeed, we are already witnessing a growing awareness of OERs among educators outside of formal teacher training channels, and the advent of Amazon Inspire —a free service for the search, discovery, and sharing of digital OERs— will further increase this awareness especially in the K-12 sector. We are also witnessing changes in, for example, university policy on open education and in government regulation where publicly funded education initiatives for developing learning resources require open licensing with Creative Commons.

In this chapter, we have also illustrated a novel corpus-based tool, FLAX, that identifies useful lexico-grammatical patterns and extracts academic words, collocations, and lexical bundles in academic text. All these features are made easily accessible through a unified searching and browsing interface. Our goal is to make current corpus technology suitable for L2 learners, helping them seek salient language samples in academic texts during writing and editing. The design was guided by outcomes and findings recorded in the research literature, and the process is entirely automatic. It should be emphasized again that although for illustrative purposes our description has focused on a particular corpora, the *Law Collections* in FLAX, it is certainly not restricted to those ESAP resources.

The versatility of the approach we have presented here also has wide-ranging implications regarding the adoption of open education policy across formal and informal learning contexts. The implementation of policy to encourage the practice of licensing pedagogic and academic texts with Creative Commons will ensure that high quality authentic texts are openly accessible to language teaching and research professionals for educational and research purposes. It is widely understood that English is the academic lingua franca of research and teaching. Open licensing will, therefore, have the positive effect of rendering pedagogic and academic texts as remixable for the development of authentic ESAP materials to support specialized language learning, both online and offline.

References

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purpose, 26*, 263–286.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*, 371–405.

Boulton, A. (2011). Data driven learning: The perpetual enigma. In S. Roszkowski & B. Lewandowska-Tomaszczyk (Eds.), *Explorations across languages and corpora* (pp. 563-580). Frankfurt, Germany: Peter Lang.

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

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Second Language Writing*, *16*, 129–147.

Chen, H. H. (2011) Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. *Computer Assisted Language Learning*, *24*(1), 59–76.

Creative Commons. (2015). *State of the Commons*. Retrieved from https://stateof.creativecommons.org/2015/cc-sotc-2015-xx12.html

Digital commons (economics). (2016, March 13). In *Wikipedia, the free encyclopedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Digital_commons_(economics)&oldid=709892273

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology, 9*(1), 99–117.

Hafner, C. A., & Candlin, C. N. (2007). Corpus tools as an affordance to learning in professional legal education. *English for Academic Purposes*, *6*(4), 303–318.

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62.

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21.

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, *32*(3), 301-319.

Hill, J. (1999). Collocational competence. *ETP*, *11,* 1-6.

Johns, T. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Figelston, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100-115). London, UK: Longman.

Lessig, L. (2008). *Remix: Making art and commerce thrive in the hybrid economy*. New York, NY: Penguin Press.

Marín, M. J. (2014). Evaluation of five single-word term recognition methods on a legal corpus. *Corpora, 9*(1), 83-107.

Milne, D., & Witten, I. H. (2013). An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, *194*(1), 222-239.

O'Sullivan, I., & Chambers, A. (2006). Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing*, *15*(1), 49–68.

Scott, M. (2008). *WordSmith Tools version 5*. Liverpool, UK: Lexical Analysis Software.

Wiley, D. (n.d.). Defining the "open" in open content. Retrieved from http://opencontent.org/definition/

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, *13*(4), 257–238.

Varley, S. (2009) I'll just look that up in the concordancer: Integrating corpus consultation into the language learning environment. *Computer Assisted Language Learning*, *22*(2), 133–152.