Language for Specific Purposes and Corpus-based Pedagogy

Elena Cotos

Iowa State University, ecotos@iastate.edu

Follow this and additional works at: https://lib.dr.iastate.edu/engl_pubs

The complete bibliographic information for this item can be found at https://lib.dr.iastate.edu/engl_pubs/224. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.
Language for Specific Purposes and Corpus-based Pedagogy

Abstract
This chapter describes how corpus-based pedagogies are used for teaching and learning language for specific purposes (LSP). Corpus linguistics (CL) refers to the study of large quantities of authentic language using computer-assisted methods, which form the basis for computer-assisted language learning (CALL) that uses corpora for reference, exploration, and interactive learning. The use of corpora as reference resources to create LSP materials is described. Direct student uses of corpora are illustrated by three approaches to data-driven learning (DDL) where students engage in hands-on explorations of texts. A combination of indirect and direct corpus applications is shown in an illustration of interactive CALL technologies, including an example of an inclusive corpus-based tool for genre-based writing pedagogy. The chapter concludes with potential prospects for future developments in LSP.

Keywords
language for specific purposes, English for academic purposes, corpora, data-driven learning, computer-assisted language learning, corpus-based pedagogy, corpus-based technologies, automated feedback

Comments
This is the peer reviewed version of the following article: Cotos, Elena. "17 Language for Specific Purposes and Corpus-based Pedagogy." The Handbook of Technology and Second Language Teaching and Learning (2017): 248, which has been published in final form at doi:10.1002/9781118914069.ch17. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.
Chapter 17

Language for Specific Purposes and Corpus-based Pedagogy

Elena Cotos

Abstract

This chapter describes how corpus-based pedagogies are used for teaching and learning language for specific purposes (LSP). Corpus linguistics (CL) refers to the study of large quantities of authentic language using computer-assisted methods, which form the basis for computer-assisted language learning (CALL) that uses corpora for reference, exploration, and interactive learning. The use of corpora as reference resources to create LSP materials is described. Direct student uses of corpora are illustrated by three approaches to data-driven learning (DDL) where students engage in hands-on explorations of texts. A combination of indirect and direct corpus applications is shown in an illustration of interactive CALL technologies, including an example of an inclusive corpus-based tool for genre-based writing pedagogy. The chapter concludes with potential prospects for future developments in LSP.

Keywords

language for specific purposes, English for academic purposes, corpora, data-driven learning, computer-assisted language learning, corpus-based pedagogy, corpus-based technologies, automated feedback
Language for specific purposes (LSP) is a well-established domain in applied linguistics. Its mission is two-fold: (1) to provide foundational knowledge about language users’ linguistic needs, which vary considerably depending on the context and purpose of their language use, and (2) to inform the teaching and learning of context-specific language that learners need to acquire in order to successfully engage in target social practices. LSP has bourgeoned into various branches. Branches that concentrate on learners in various academic subject areas are nested under Language for Academic Purposes (LAP). Language for Occupational Purposes (LOP) is associated with professional and vocational contexts. The classification of LSP has also expanded to include language for sociocultural purposes, catering to the needs of socially or physically disadvantaged language learners. While LSP encompasses any language, English for specific purposes (ESP) has been at the forefront of the research in the past 50 years. Therefore, this chapter highlights ESP practices and research, which are also illustrative of work in other languages (e.g., Parodi 2015; Vergaro 2004).

LSP as a “specific-learner-centered language instruction” (Belcher 2009, 2) has undergone great transformations over the past 30 years due to the research tools and methodologies offered by Corpus Linguistics (CL), the study of large quantities of authentic language using computer-assisted methods. This chapter maps out corpus-based pedagogies in LSP, reviewing uses of corpora for reference, exploration, and interactive computer assisted language learning (CALL). First, I describe how corpora can serve as reference resources to create LSP materials. Second, I show direct student uses of corpora in data-driven learning (DDL) approaches where students engage in hands-on explorations of texts. I categorize these approaches into three strands, each with its essential concepts and background in CL research. Next, I show how indirect and direct corpus applications have been translated to interactive
CALL technologies, including an example of an inclusive corpus-based tool for genre writing pedagogy. The chapter concludes with potential prospects for future developments in LSP.

LSP corpora for reference

Widespread access to corpora by materials developers and researchers has created a dynamic relationship between LSP and CL, providing insights for needs analysis (see Dudley-Evans and St. John 1998) with rich descriptions of language in specific contexts of use. LSP researchers and materials developers use corpora for decisions regarding the design of needs-responsive curricula, course syllabi, teaching materials, classroom tasks, and assessment. Such uses are referred to as indirect because corpora serve as language sources that inform the content of materials and tasks to be used in the classroom, rather than being directly accessed by teachers and students. To provide language descriptions that closely meet specific learning needs, indirect uses require a relevant corpus of texts and software tools.

Corpora

Corpora are large machine-readable compilations of authentic texts. They can be general or specialized, depending on what kinds of texts are included in the corpus. The former contain large volumes of text, up to hundreds of millions of words, intended for a range of researchers wanting to investigate particular linguistic phenomena or to provide grammatical and lexical descriptions of a language as a whole. The latter, which are typically compiled for a specific project, can be much smaller and are intended to describe language use in specific contexts.

Specialized corpora are of direct relevance to LSP. Practitioners often choose to compile small corpora based on specific instructional needs. Needs-based corpora are more clearly patterned and more feasible for constructing, managing, and interpreting (Aston 1997). For example, several corpora of academic writing were compiled in Hong Kong: the Hong Kong
University of Science and Technology (HKUST) Computer Science Corpus, the Hong Kong Financial Services Corpus, and the Hong Kong Engineering Corpus. Other examples include the Jiaotong Daxue English of Science and Technology (JDEST) Corpus and the Guangzhou Petroleum English Corpus (GPEC). Spoken language is represented in the Michigan Corpus of Academic Spoken English (MICASE) and the Cambridge and Nottingham Spoken Business English Corpus (CANBEC). Corpora containing learner language have also gained popularity. A list of learner corpora can be found on the website of the Centre for English Corpus Linguistics (2016).

The advantage of specialized LSP corpora is that they represent language characteristic of the registers and genres of the contexts of interest to the specific purposes of learners. Registers share patterns of lexical and grammatical features of language that are determined by situational factors (e.g., written or spoken, formal or casual, scientific or technical). Genres are grouped into text types according to distinct sociocultural purposes and discourse conventions (e.g., research articles, grant proposals, business reports). In terms of register, corpus texts are viewed as language used in recurring situations in a society; in terms of genre, corpus texts are viewed as types of regularly recurring messages in a community (Ferguson 1994). The understanding of registers and genres is critical for teaching learners the specific language they need in order to successfully engage in communication with their discourse communities. Therefore, LSP practitioners have to carefully consider these concepts in addition to the subject or practice area.

Corpus tools
The most common tools used in corpus analysis informing LSP pedagogy are concordancing programs, which are text search engines with sorting functions. The first concordancers became available for personal computers in the 1980s (e.g., MicroConcord and Mini-concordancer), and
later Internet access enabled broad use of web-based concordancers such as *WordSmith* (Scott 1996), *MonoConc Pro* (Barlow 2000), *ConcGram* (Greaves 2009), *AntConc* (Anthony 2014), *WordSearch* (Cortes 2007), and *TextSTAT* (Benini 2010). When queried, these concordancers extract a “collection of the occurrences of a word-form, each in its textual environment” (Sinclair 1991, 32), which is displayed as lists of key words in context (KWICs) called concordance lines. Figure 17.1 shows sample concordance lines from the Corpus of Contemporary American English (COCA), which can be re-sorted to see patterns in the use of the word “develop.” The concordancer also shows the word’s synonyms, definitions, and its relative frequency in different academic domains. Co-occurring words like “develop” and “understanding,” known as collocates, are also displayed. The search was done with a wildcard (*) that replaced the end of the word. In this way, all the different forms of “devel*”—“develop,” “developed,” “developing” were extracted. Concordancers can also help identify the grammatical and syntactic patterns of search words, their shades of meaning, meaningfully associated collocates, and positional and constituency variation. Some concordancers allow for both phraseological and paradigmatic searches for categories like subject, predicate, and direct object (Flowerdew 2015).
Considering that “materials should be based on analyses of representative samples of the target discourse” (Hyland 2002, 113), practitioners generally use concordancers to identify frequencies and patterns of certain linguistic features in relevant corpora and then apply the results of their queries to teaching. As corpora lend themselves to the detection of lexis, concordancers have been used to produce frequency and range-based lists of academic vocabulary and collocation lists (Ackermann and Chen, 2013; Durrant 2009; Gardner and Davies 2014). Field-specific vocabulary and collocations have been included in instructional materials for business (Popescu 2007; Walker 2011), engineering (Mudraya 2006), nursing (Yang 2015), and agriculture.
(Martínez, Beck, and Panza 2009). Additionally, teachers use concordancers to create materials for classroom exercises that prompt students to test linguistic hypotheses, notice contextual meanings, examine collocations, and so on. Corpus-based exercises can range from fill-in-the-blanks and matching of split concordance lines to demonstrating different forms of a particular lexical item.

Because learner corpora capture learners’ difficulties in producing specific target language, they have entered the LSP arena with the same pace of pursuit as the corpora produced by native speakers (Granger and Paquot, 2013). Rankin and Schiftner (2011) recommend running comparative searches in a corpus compiled “in-house” from current students and in a relevant native-speaker corpus, and then developing teaching materials taking into account the linguistic differences identified. Similarly, contrastive analyses of native-speaker, learner, and first-language texts as well as of parallel corpora that contain the same texts in different languages are great sources for materials design (Teubert 2004).

Corpora are excellent resources for grammarians and lexicographers who produce reference books, grammars, and dictionaries. One of the most representative pedagogical volumes is the Longman Grammar of Spoken and Written English (Biber et al. 1999). It records an exhaustive frequency-based description of lexico-grammatical features of written prose, conversation, fiction, and news in terms of structural characteristics and patterns of use. The textbook Exploring Academic English (Thurstun and Candlin 1997) integrates exercises based on concordances for lexical items indicative of sets of rhetorical functions. Recently, the Professional English Online and the Business and Professional English series by Cambridge University Press have entered the market. Basturkmen’s (2010) Developing Courses for English for Specific Purposes is an important resource for teachers showcasing how descriptions of specialist discourse can be used to determine the curriculum.
LSP corpora for data-driven exploration

Exploiting corpora and concordancing tools with students in the classroom is known as direct corpus application in teaching. A key paradigm in direct corpus application in teaching is data-driven learning (DDL) (Johns 1986, 2002), which is an approach that engages students as independent analysts—detectives who explore authentic language use captured in a corpus. Corpora are thus used as enhancers of students’ linguistic intuition (Gavioli 2005). The teachers are mediators; they carefully designate the linguistic aspects to be analyzed by their students and determine the corpora, tools, and progression of corpus consultation activities.

Lead advocates of DDL in LSP emphasize the pedagogical strengths of this approach, arguing that it is “fully compatible with communicative language teaching; discovery learning and learning by doing; autonomisation and learning to learn; learner-centeredness and individualization; collaborative learning and creativity; task-based as process as well as product orientations; form and meaning in constructivism; with an emphasis on the authentic language of discourse by register / genre” (Boulton 2011, 575). In DDL, students adopt tools and techniques used by corpus linguists; however, they do not carry out the same kinds of analyses as those conducted by linguists. In what follows, I focus on three approaches from CL that have influenced DDL used in the classroom—bottom-up, top-down, and paired.

Bottom-up approach

The bottom-up approach to corpus-based pedagogy builds upon register analysis in CL research, which focuses on pervasive lexico-grammatical features that “occur frequently in the target text variety because they are well-suited functionally to the [...] situational context of the variety” (Biber 2010, 242). The starting point for CL is bottom-up automatic segmentation of all the texts in a corpus to identify different types of lexico-grammatical features and examine their patterns.
Table 17.1 lists some examples of register analysis studies using this approach. Many researchers examined written and spoken LOP workplace registers, although corpus studies of linguistic features especially abound in LAP.

Table 17.1: Examples of bottom-up register analysis studies.

<table>
<thead>
<tr>
<th>Research focus</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salient linguistic features</td>
<td>Ferguson (2001), Kuo (1999), Thomas and Hawes (1994)</td>
</tr>
<tr>
<td>Co-occurrences of multi-word units</td>
<td>Cortes (2004), Grabowski (2013), Gledhill (2000)</td>
</tr>
<tr>
<td>Co-occurrences of words irrespective of constituent / positional variation</td>
<td></td>
</tr>
</tbody>
</table>

Teachers have particularly favored DDL activities designed using techniques from bottom-up research methods, largely because of the availability and practicality of concordancers. They often encourage students to interpret concordance output vertically by examining frequency lists and concordance lines to identify patterns of language use. Alternatively, students can interpret concordances horizontally by reading the surrounding
context and noting specific language choices that express functional or cultural meaning (Braun 2005, 54). Gavioli (2005) provides helpful guidelines for ESP teachers along with a series of activities that engage students with specialized and bilingual corpora. Concordance-based tasks may include students’ inductive study of technical vocabulary, searching the corpus and examining KWICs to better understand appropriate use of language in context, conducting different types of searches for comparative purposes, identifying language misuse in learner language, and so on. Similarly, Mudraya (2006) illustrates activities that prompt engineering students to:

1) distinguish between general meaning and meaning as sub-technical vocabulary (*solution of problems* vs *solution from the absorber*),
2) supply adjective collocates for general and sub-technical meaning (*possible solution* vs *concentrated solution*),
3) supply verb collocates for general and sub-technical meaning (*attempt a solution* vs *immerse in solution*),
4) exemplify specific syntactic patterns, e.g. *solve + solution method*

(*solve/solves/solving/solved + with* (a vector approach) / *by* (drawing) / *using* (work-energy principle)).

Such analysis of technical language is ideal for the study of advanced academic writing, which is placed high on the LSP/ESP agenda. A wealth of publications, including some studies reviewed above, present uses of corpora of research articles in graduate-level writing courses. A classic example of a course for doctoral students is described by Lee and Swales (2006). Their tasks involved queries of lexico-grammatical and discourse patterns in different types of specialized corpora. For example, the students used web- and PC-based concordancers to:
1) generate word lists (e.g., reporting verbs),
2) concordance for errors (e.g., *at the end* v. *in the end*),
3) examine grammatical structures (e.g., *suggest that*),
4) guess/scrutinize word meanings (e.g., *continually* vs *continuously*),
5) generate examples of puzzling pairs (e.g., *seek* vs *search*),
6) guess and verify frequencies,
7) rank word classes most likely to trigger certain structures (e.g., ADJ/N/V + V-ing, *appropriate for modeling*),
8) analyze rare patterns (e.g., V (NP) + V-ing, *led into admitting*).

Top-down approach

The top-down approach, which builds upon genre research, entails analyses of the internal structure of texts. It is referred to as top-down because the starting point is identifying patterns of text organization using analytic frameworks of possible discourse units (Biber, Connor, and Upton 2007). One of the most productive genre frameworks was developed by John Swales (1981, 1990). He describes discourse in terms of two kinds of discourse units—“moves” (or communicative goals) and constituent “steps” (or rhetorical strategies that accomplish communicative goals). In top-down CL research, the focus is on discursive patterns, constraints, variation, and language choices, which conventionalize the rhetorical identity of texts and establish conformity with the expectations of the target discourse community. Table 17.2 exemplifies works that describe academic and professional genres.

Table 17.2 Examples of top-down genre analysis studies.
<table>
<thead>
<tr>
<th>Genre group</th>
<th>Genre investigated</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Graduate program</td>
<td>Samraj and Monk (2008)</td>
</tr>
<tr>
<td></td>
<td>applications</td>
<td></td>
</tr>
<tr>
<td><strong>Professional genres</strong></td>
<td>Medical consensus</td>
<td>Mungra (2007)</td>
</tr>
<tr>
<td></td>
<td>statements</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Law, management,</td>
<td>Lung (2008, 2011)</td>
</tr>
<tr>
<td></td>
<td>economics cases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Corporate press releases,</td>
<td>Catenaccio (2008), Cho and Yoon (2013),</td>
</tr>
<tr>
<td></td>
<td>earnings calls</td>
<td>Crawford Camiciottoli (2006)</td>
</tr>
<tr>
<td></td>
<td>Negotiation letters</td>
<td>Pinto dos Santos (2002)</td>
</tr>
</tbody>
</table>

In top-down genre-based teaching, move frameworks developed by researchers have become essential guidelines for students’ corpus explorations of the macro-structure of texts. The directions listed below, which are drawn from Swales and Feak (2004, 2012), Robinson et al. (2008), and Cotos (2014), illustrate how teachers introduce the moves and steps before referring
their students to the texts in the corpus. Prior to that, they may also work through a text sample requiring students to do the following:

1) identify major purposes of paragraphs and propose a move structure,
2) divide a sample text into moves,
3) identify and discuss move boundaries,
4) identify the step components of moves,
5) re-construct sentences in teacher-modified texts into their original order.

When directed to corpora, students are asked to:

1) examine texts for their organization into sections and subsections, noting the naming conventions and noticeable transitions,
2) examine corpus texts for moves that are present or absent,
3) identify patterns in rhetorical organization,
4) discuss similarities and differences among texts (e.g., from the same and from different journals),
5) reflect and explain reasons for alternative move structures.

Students also focus on the language choices made by authors, but in a deductive rather than inductive way. Commonly, they are asked to identify words or expressions that they think convey the functional meaning of moves or steps (e.g., has been extensively studied used to claim topic centrality). Or, they may try to deduct some general trends about the use of verb tenses and their relation to a particular communicative intent (e.g., this study reports vs. the purpose of this study was when announcing present research).

Rhetorically annotated corpora are particularly helpful for successful completion of such tasks. Cotos, Link, and Huffman (2016) describe how students conduct a rhetorical composition
analysis task, accessing a corpus of research articles annotated for moves and steps via the Callisto workbench. Because this tool uses colors to differentiate between tagged moves, it visualizes the move anatomy of each text enabling students to observe and notice:

1) the distribution of moves,
2) the sequence of moves,
3) the occurrence of steps within each move,
4) the moves and steps that appear more often than others,
5) the moves and steps that are uncommon in their discipline,
6) rhetorical overlap (i.e., segments that represent more than one move or step),
7) similarities and differences in the function of a given step in different texts,
8) content used to realize steps.

Paired approach

While the bottom-up and top-down approaches have a value of their own, integrating the two can be “the best of both worlds” (Charles 2007, 299). Therefore, CL researchers have combined them into a paired approach, establishing links between the lexico-grammatical features and the generic moves of specialized discourse. Table 17.3 lists several studies that reveal both micro-textual and macro-textual distinctiveness.

Table 17.3 Examples of combined bottom-up and top-down corpus studies

<table>
<thead>
<tr>
<th>Research focus</th>
<th>Genre</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic characteristics of moves</td>
<td>Research articles</td>
<td>Kanoksilapatham (2007),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Durrant and Mathews-Aydınlı (2011)</td>
</tr>
<tr>
<td>Topic</td>
<td>Context</td>
<td>Authors</td>
</tr>
<tr>
<td>---------------------------------------------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>Vocabulary-based discourse units</td>
<td>Research articles</td>
<td>Biber, Connor, and Upton (2007)</td>
</tr>
<tr>
<td>Collocations in moves</td>
<td>Application</td>
<td>Ding (2007), Marco (1999)</td>
</tr>
<tr>
<td>Interpersonal features in moves</td>
<td>Editorial letters</td>
<td>Flowerdew and Dudley-Evans (2002)</td>
</tr>
<tr>
<td>Rhetorical structure and pragmatic use of mood, modality, reference system</td>
<td>Sales promotion</td>
<td>Vergaro (2004)</td>
</tr>
<tr>
<td>Lexical bundles and discourse functions</td>
<td>Student and published academic writing</td>
<td>Chen and Baker (2010)</td>
</tr>
<tr>
<td>Rhetorical unity defined by moves and linguistic features denoting temporal and special aspects</td>
<td>Cover letters</td>
<td>Crossley (2007)</td>
</tr>
</tbody>
</table>

Some DDL practitioners exhibit a similar tendency in the classroom, implementing both the analysis of concordance queries and the analysis of text structure. For example, Weber (2001) developed activities for the teaching of legal essays. The students first read entire texts to identify prototypical rhetorical principles inherent to legal cases and selected vocabulary used to convey respective communicative intent. They then consulted the corpus to verify whether the selected expressions displayed patterns of regularity. In their courses on research writing, Charles (2007), Cortes (2007), and Flowerdew (2015) began by raising students’ awareness of
The patterns of macro discourse with top-down text analysis tasks, which drew their attention to rhetorical functions, communicative purposes, and various linguistic realizations. Transitioning to bottom-up analysis, they then taught students to conduct concordance searches of specific lexico-grammatical items.

The symbiosis between top-down and bottom-up corpus-inquiry can further be augmented with ethnographic insights. For instance, in Hafner and Candlin’s (2007) study, students in a professional legal training course were offered online corpus tools as an optional DDL affordance to facilitate their writing of legal documents for their needs outside the classroom. Finding that “students who construct themselves as apprentice lawyers bring this identity to the corpus consultation task,” Hafner and Candlin urge LSP teachers to “draw on professional practices when perceiving affordances in corpus tools and resources” (316).

LSP corpora for interactive computer-assisted language learning
Interaction is an important process in language learning, but conditions for interaction may be limited or unavailable, especially for LSP students. Corpus-based technologies can create enhanced learning conditions if designed as semi-intelligent or intelligent CALL (computer-assisted language learning) systems. Such systems are powered by natural language processing (NLP), a domain of artificial intelligence, which relies on training the computer to recognize spoken or written language. NLP methods somewhat mirror the paired register-genre approach to corpus analysis. Specifically, NLP is concerned with automatically detecting lexico-grammatical features in native speaker and learner corpora. Those features are then used to train computer models to automatically analyze genres and genre components (Stamatatos, Fakotakis, and Kokkinakis 2000). In this section, I briefly review NLP-based tools for reference, DDL, and automated feedback. Then, I describe an example of a pedagogical system that integrates these
three types of CALL tools in three respective modules to help students learn how to write research articles in English.

Interactive reference tools
Several technology-savvy teacher-researchers created novel interactive reference tools for academic English, which illustrate different ways in which corpora can be exploited to create more engaging materials. For example, Williams (2012) developed DicSci, a pattern dictionary of science verbs helpful for vocabulary learning. DicSci was built bottom-up from a corpus using collocational networks (i.e., chains of collocations extracted from a corpus by means of analyzing proximity and statistical procedures). Williams (2012) refers to it as “organic” because it groups words and presents word uses as growing naturally from the corpus data, as opposed to being determined by someone’s intuition. The word “control,” for instance, can be the node of a collocation network that contains several words most frequently co-occurring in a given corpus. One of its most frequent collocates is “group” that, in turn, most frequently co-occurs with “subjects.” Subsequently, “subjects” collocates with certain verbs (e.g., “recruit,” “instruct”) and adjectives (e.g., “obese,” “sick-listed”). This growing chain of collocates renders the semantic environment of the node word and represents natural language use as captured in the corpus. Kuo (2008) describes a somewhat similar collocation builder system that draws on parallel corpora in English and Chinese to provide collocation examples in both languages.

A few other tools were developed for more specific needs. Pinkwart et al. (2006) designed Legal Argument Graph Observer (LARGO), a tool for teaching law students how to make arguments and formulate warrants for deciding a case. LARGO enables students to graphically represent examples of legal interpretations with the hypotheticals they observe while
reading texts. Nehm, Minsu, and Mayfield’s (2011) *Summarization Integrated Development Environment (SIDE)* system assesses the accuracy of students’ written explanations of science topics. *Internet Writing Resource for the Innovative Teaching of English (iWRITE)*, a database-driven online tool for grammar and academic writing, makes use of a learner corpus that was collected “in-house” from students enrolled in L2 writing courses at a North American university (Hegelheimer and Fisher 2006). Bloch (2010) applied research results to the development of a web-based database for academic writing, which contains sentences with reporting verbs functioning as rhetorical devices. In a like manner, Henry and Rosenberry (2001) created a hyperlinked website for teaching application letters relying both on frequencies and on linguistic, discourse, and syntactic features characteristic of the steps that accomplish the moves of this genre.

**Interactive DDL tools**

Interactive DDL tools use NLP to enhance students’ data-driven explorations of native-speaker and learner corpora, particularly fostering deduction from automatically analyzed texts. Using a science and engineering corpus, Anthony and Lashkia (2003) developed the *Mover*, software that automatically identifies the moves in the abstracts of research articles. Students can collect a small corpus of abstracts and upload it to the *Mover*, which then labels each sentence in each text with a move. This software thus enables the students to analyze move sequences and draw conclusions about the rhetorical organization of abstracts. Also employing a genre-analytic approach, Chang and Kuo (2011) created *Moves And Keywords Engine (MAKE)*. This tool can be queried for keywords and re-current word combinations such as “the degree to which” and “in the context of,” known as lexical bundles, within single and multiple moves identified in a corpus of research articles in computer science. The open-source *Type Your Own Script (TYOS)*
allows students to exploit learner corpora of initial drafts and revised drafts with mark-up and corrections by the teacher, which highlight rhetorical strategies and language choices (Birch-Bécaas and Cooke 2012).

Interactive feedback tools

NLP-powered tools that scaffold learners with just-in-time feedback as they are producing language create conditions for a higher degree of interaction. Various interactive feedback tools have been developed for specific domains such as journalism (*Journalism*), navy (*CRES*), biology (*SAGrader*), business (*EPISTLE*), sociology (*APEX Assessor*), education (*ETIPS*), and psychology (*RMT*). A number of such tools are available for academic writing (e.g., *Criterion*, *MyAccess!*, *WriteToLearn*, *Folio*, *Writer’s Workbench*, *SkillWriter*, *Writing Roadmap*, etc.). Some writing systems are specific to scientific writing. For example, the *Intelligent Academic Discourse Evaluator (IADE)* analyzes students’ Introduction section drafts and provides color-coded move-level feedback. *IADE* also presents percent ranges for each move, comparing it with a corpus of Introductions from articles published in students’ disciplines. *Scaffolded Writing and Rewriting in the Disciplines (SWoRD)* is a web-based tool that implements reciprocal peer review to simulate the journal publication process (Cho and Schunn 2007). To improve peer feedback, Xiong, Litman, and Schunn (2012) developed a tool that instantaneously processes students’ peer reviews, detects the presence or absence of features important for quality feedback, and prompts the students to revise their reviews. A tool for writing strategy training, *Writing Pal*, generates feedback using sophisticated algorithms that assess lexical, syntactic, cohesion, and rhetorical features (McNamara, Crossley, and Roscoe 2013).
An example of “all-in-one” corpus-based tool for genre writing pedagogy

Pedagogical needs for reference materials, DDL practices for exposure to authentic language, and intelligent CALL for feedback-scaffolded language practice—all converge in a recently developed system for scientific writing called Research Writing Tutor (RWT) (Cotos, 2016). It was created using move analysis results derived from a corpus of 900 journal articles published in the top journals of 30 disciplines (Cotos, et al. 2015). The researchers manually annotated this multidisciplinary corpus for moves and steps, and RWT uses the annotations to retrieve and present the corpus data in three interrelated modules.

RWT for reference. A learning module, called “Understand Writing Goals,” contains web-based pedagogic enrichment materials (Braun 2005). The materials include definitions, explanations, and examples of all the moves and steps pertaining to Introduction, Methods, Results, and Discussion/Conclusion (IMRD/C) sections to be used by teachers when introducing the key rhetorical concepts of the research article genre. The students, in turn, can use these materials for knowledge consolidation. Or, in the case of autonomous learning, this module can serve as a study guide because it also includes a series of video mini-lectures explaining genre-specific content.

RWT for DDL. A demonstration module, called “Explore Published Writing,” gives students direct access to the corpus (both in its original and annotated form) in a way that allows for paired DDL explorations. It contains a function-based concordancer, which displays excerpts representing a particular step of a particular move, at the same time exhibiting a variety of language choices used to express different shades of functional meaning. From there, students’ corpus exploration can shift to top-down macro-level analysis by clicking on any of the concordance lines. That brings up the entire text color-coded for moves and glossed with steps. Such enhancement of corpus data fosters the exploration of co-texts beyond individual
concordance lines and the identification of structural patterns in the discipline. Alternatively, teachers may first direct their students to the annotated corpus with tasks that encourage horizontal reading of a given IMRD/C section, and then assign DDL tasks focused on linguistic means of expression in relation to their communicative purposes.

*RWT for interactive feedback.* The third, “Analyze My Writing” module, offers students the opportunity to apply their corpus observations to their own writing and creates conditions for iterative revisions of their drafts. This module operates with the help of an engine that automatically analyzes students’ IMRD/C drafts and returns multi-level rhetorical feedback. The feedback provided at macro level is operationalized through color codes, each indicative of a particular move (similar to *IADE*). *RWT* affords navigation flexibility, so the students can open the “Explore Published Writing” module in another window or tab and compare the distribution of colors in their draft with that of the texts in the annotated corpus. Another type of macro-level feedback displays an actual comparison of each move in the student’s draft with published disciplinary texts based on median numerical values. This numerical feedback shows range bars with percentages indicating that the draft may have “too much” or “not enough” of a given move, or that the percentage of that move falls within the “goal” range of the discipline. In addition to ranges, the comparison is visualized as summary pie charts. As the purpose of each move is expected to be accomplished with certain steps, the move feedback is accompanied by hints specifying which steps are used to an extent similar to the target discipline, and which steps are lacking or may need to be improved.

The feedback provided at micro level is generated for individual sentences in students’ drafts. This type of feedback takes the form of interactive comments or clarifying questions about the functional meaning of the sentence. For example, *RWT* suggests, “You are likely *introducing present research descriptively*” for the sentence “In this paper, we address
individual, maternal, and family stressors experienced by low-income adolescents.” Or, it may elicit a confirmation of the function of “Only one study to date has addressed how stress may moderate the relationship between food insecurity and adolescents’ probability of being overweight or obese” by asking “Are you indicating a gap?” Such comments are aimed at encouraging the students to think about intended meaning and to revise their writing where necessary. When revising with RWT, the students may need help with interpreting the feedback. To support them in these instances, the analysis module embeds on-demand glosses with brief explanations of the moves and steps next to the macro-level feedback. Students are also encouraged to “Learn More” or “See Examples” through hyperlinks to relevant content in the learning and demonstration modules, respectively. A note-taking feature added to the draft analysis interface also supports the revision process.

In sum, RWT amalgamates core requisites of corpus-based LAP/LSP instruction. Drawing on corpus analysis, the tool provides teachers with attested examples of specialized language use and with a platform for developing tasks responsive to student needs. RWT also accounts for the fact that teachers may lack familiarity with the writing conventions of specific disciplines. The tool’s corpus-based components and output are learner-centered and easy to interpret, at the same time presenting complex linguistic evidence. Moreover, RWT’s automated analysis and feedback on student writing adds an important learner-driven data dimension, which is beneficial for it draws learners’ attention to problematic areas in their own production (Nesselhauf 2004) and can enhance motivation (Seidlhofer 2002).

Future expectations
Saying that CL has a role to play in the advancement of LSP would be an understatement. Corpus methodologies provide versatile approaches and tools for the study of naturally occurring
texts at linguistic, functional, rhetorical, and pragmatic levels. To date, LSP has greatly benefited from corpus-based frameworks of analysis, descriptions of authentic language use, and interactive technologies. As the field moves forward, new directions will be charted for corpus-based and computer-assisted LSP.

Given that novel web-applications and computational models for text analysis offer exceptional techniques for utilizing CL findings, LSP is positioned to embrace innovative technologies for the design of instructional materials and principled curricula. Williams (2012), for instance, projects that mind mapping technology, which can display information in a relational way (as in the *DicSci* organic dictionary), will give users access not only to individual entries and phraseological descriptions, but also to super entries linking quasi-synonyms and writing assistance. Coding each word in a corpus with part-of-speech, case, number, and gender tags (known as morpho-syntactic tagging) will be overtaken by pedagogically motivated corpus annotations of both native-speaker and learner corpora (see Pérez-Paredes and Álcaraz-Calero, 2009). This will, in turn, enable new multi-modally human and computer-annotated corpora (Knight et al. 2009) as well as corpora annotated by learners to interfuse with applied NLP (McCarthy and Boonthum-Denecke, 2012) in systems capacitated for powerful language data export features tailored for specific learning goals.

Much like CALL, LSP will continue to evolve in terms of applications that will be both authentic and research-supported. This requires expansive CL study of specialized discourses as well as copious research on language teaching and learning supported by corpus-based materials and tasks. I anticipate that corpus data will be well aligned with pedagogic constructs. Arguably, future research will need to fully exploit the potential of corpus and computer-networking technologies to construct language-learning environments in the light of learning theories and to guide effective implementation of technology-enhanced LSP.
References


Ding, Huiling. 2007. “Genre Analysis of Personal Statements: Analysis of Moves in Application Essays to Medical and Dental Schools.” English for Specific Purposes, 26, no. 3: 368–392.


Pérez-Paredes, Pascual, and José Álcaraz-Calero. 2009. “Developing Annotation Solutions for Online Data Driven Learning.” *ReCALL*, 21, no. 1: 55–75. DOI: 10.1017/S0958344009000093


Yang, Ming-Nuan. 2015. “A Nursing Academic Word List.” English for Specific Purposes, 37: 27–38. DOI: 10.1016/j.esp.2014.05.003

Biographical Note
Elena Cotos is Assistant Professor in the English Department and the Director of the Center for Communication Excellence of the Graduate College, Iowa State University. Her research interests include EAP/ESP, genre analysis, automated writing evaluation, CALL, learner corpora, and language assessment. Her work was published in a number of professional journals, edited volumes, and a book-length monograph.