Agnieszka Leńko-Szymańska and Alex Boulton (eds.),Multiple Affordances of Language Corpora for Data-driven Learning

Christopher Gledhill

Book review

Christopher Gledhill
Université Paris Diderot, CLILLAC Arp

Agnieszka Leńko-Szymańska and Alex Boulton (eds.). 2015. *Multiple Affordances of Language Corpora for Data-driven Learning* (Studies in Corpus Linguistics 69). Amsterdam/Philadelphia: John Benjamins Publishing Company. 312 pp. ISBN 978-9-0272-0377-9.

This collection of papers attempts to answer a very practical question: what are the best tools and techniques which language teachers and language learners can use to exploit corpus data, online text archives and other "language-rich" resources? In this volume the editors, Agnieszka Leńko-Szymańska (Institute of Applied Linguistics, Warsaw) and Alex Boulton (CRAPEL-ATILF, Nancy Université), present a selection of twelve papers from the 10th Teaching and Language Conference (TaLC, Warsaw, 11th-14th July 2012). In the early days of the "corpus revolution" (i.e. the 1990s, at least from the point of view of this reviewer), the TaLC conferences represented a pioneering attempt to apply the new methods and findings of corpus linguistics to the question of language learning. Since this volume is in many ways an anniversary edition for the TaLC community, it is pleasing to find many of its contributors are the people who have been instrumental in the field (Guy Aston, Maggie Charles, Lynne Flowerdew, Chris Tribble...). As well as presenting new data and observations, these authors also provide many key insights on how the field has moved on. In this respect, one name in particular crops up with regularity: Tim Johns, a highly original and sorely-missed member of the *Cobuild* dictionary project (Birmingham University). As pointed out in several places in the book, it was Tim Johns who framed many of the core concepts that are still used in corpus-based language analysis, and it was he who came up with the term "data-driven (language) learning" (henceforth abbreviated to DDL).

The book appears as number 69 in the John Benjamins collection *Studies in Corpus Linguistics*; this follows another volume edited by Alex Boulton, Shirley Carter-Thomas and Elizabeth Rowley-Jolivet which was reviewed in *ASp* (Bordet & Pic 2012). The focus of this previous collection was how to use DDL for teaching and research in English for Specific Purposes (ESP). In the present volume, the onus is on what tools and techniques can be used in order to exploit language corpora in a virtual or face-to-face environment, either for language learning or for related skills.

The book is made up of four sections. Each section comprises three chapters, each grouped by common themes: an introductory section presenting a theoretical

or historical perspective, and then three more specific sections on "corpora for language teaching", "corpora for skills development" and "corpora for translation training". The book is rounded off by a thirteenth chapter by one of the editors (Alex Boulton), who discusses the future of language corpora in the light of current internet usage. In the following paragraphs, I summarize the main points of each chapter, including – if appropriate – a brief comment at the end of each.

1. Introduction: Data-driven learning in language pedagogy

In their introduction, Agnieszka Leńko-Szymańska and Alex Boulton discuss the different underlying arguments that emerge throughout the book (e.g. the dearth of DDL at elementary and intermediate levels, the perceived technophobia that seems to put many language students off using a corpus, etc.). The authors also provide an explanation for the intriguing buzzword used in the title: 'affordances'. Apparently this term was inspired by perception psychology, and more particularly by James Gibson. As the editors put it: "[an affordance] is a person's perception of the environment that prompts some course of action. Affordances thus refer to the properties of an object in the environment enabling any kind of activity..." p. 1). The stage is thus set for a discussion not only of the corpus as a resource for making linguistic observations, but also as a means of changing the learner's perception of language and what he/she can do with corpus data.

2. Data-driven learning and language learning theories: Whither the twain shall meet

In this chapter, Lynne Flowerdew reviews a selection of studies on the effectiveness of DDL. She divides these into three main approaches. The first, the "noticing hypothesis", emphasizes the importance of conscious learning strategies. For example, a study of this type can involve students speculating about what potential keywords and patterns may occur in a given text; the students are then asked to consult a corpus to validate their predictions. It is notable that this approach tends to see corpus work as the main focus of the DDL activity ("how to use a corpus to better learn a language"). Flowerdew then looks at what she terms "constructivist approaches". This tends to involve the development of dedicated, task-dependent learning environments. The complexity of this kind of study means that it has not often been attempted, but Flowerdew cites one or two examples, such as the "Check my Words" toolkit: here the learner has a single task to complete, but is given several tools to use, all on the same platform (including such activities as: accessing a corpus, consulting a grammar guide or pulling down various menus with hints and examples of usage patterns, etc.). Finally, Flowerdew points out that very few studies have attempted what she calls a "sociocultural" (Vygotskyan) approach, i.e. a study that de-emphasises the corpus and linguistic observation, and instead relies on staged learning and cognitive scaffolding. One exception is Chau (2003), who asks students to construct a database and create a dictionary, while only using the corpus indirectly, as a means of reaching some other particular learning objective. In conclusion, Flowerdew suggests that although there have been plenty of studies which attempt to demonstrate empirically that DDL is effective, the evidence according to her is still rather

unconvincing. She suggests that this is often because of the very small scale and timespan involved in many studies. In addition, and perhaps most crucially, little is known currently about how different individual learning styles might impact the effectiveness of DDL.

3. Teaching and language corpora: Perspectives from a personal journey

In this chapter, Christopher Tribble recounts his experience of over thirty years of language teaching using "corpus-informed learning". In the first half of the paper, he argues that although corpus linguistics and DDL have undergone profound changes over this period, notably thanks to the democratization of technology (from tape recorders in the 1970s to portable computers in the 1990s), these developments have hardly had any impact at all on mainstream language teaching and learning, apart from some indirect applications (reference materials, syllabus design). In the second half, Tribble sets out the results of two surveys of language practitioners on various mailing lists (including Linguistlist and Corpora List) between 2008 and 2012. The results are sometimes surprising. For example, he notes a marked reduction in the types of institutions which use corpus-informed learning: this activity is increasingly restricted to universities and academic institutions (from around 40% to nearly 80% of respondents). There has also been a corresponding reduction in the range of applications of corpus work (LSP/Business students for example dropping from 10 to 5%) and a similar drop in the expected proficiency levels of target students (e.g. CEFR level C2 from over 10% to around 7%). Tribble also presents some intriguing findings on his respondents' favourite publications (coming in at number one is O'Keefe *et al*. 2007 *From Corpus to Classroom*) as well as electronic resources (surprisingly the BNC and COCA are still the best known). In an interesting and characteristic twist, Tribble has made a corpus out of the respondents' answers to his questions and analyses their 'n-grams' (recurrent phrases) in order to provide a brief but telling meta-analysis of their discourse. Thus, for example, among the most frequent reasons for *not* using a corpus is the productive, but also depressingly predictable lexico-grammatical pattern: "(they, many teachers, students, the participants…) (do not, did not) know how to use (the corpus, this technology, their knowledge, this resource)".

Part I. Corpora for language learning

4. Learning phraseology from speech corpora

This chapter revisits two key insights from corpus linguistics: 1) language data is largely made up of frequently recurring multi-word patterns (variously termed 'phraseological units', 'collocations', etc.), and 2) proficient language users depend on these phrases in order to routinely produce – and predict – meaningful discourse in a largely automatic, idiomatic manner (cf. John Sinclair's "idiom principle"). In this chapter, Guy Aston focuses on how trainee interpreters might benefit from the analysis of such phraseological units by using a corpus. It is currently not possible to consult a parallel corpus of texts which have been interpreted into a target language. Aston therefore does the next best thing, which is to use the well-known TED talks as a data source. Many TED talks have been subtitled into various

languages, and these can be downloaded as text (.txt) files with timecodings (i.e. links to points in the video). These texts can then be aligned and searched in order to hear the original speech and to see the translator's subtitles at the same time. Aston makes the interesting point that it is possible not only to look for regularities of expression in the corpus, but also to study pronunciation and prosody. For example, the phrase "in other words" with the stress on "words" is typically used as an independent tone group to reformulate or explain a neighbouring word or phrase, while the phrase "in other words" (with the stress on "other") is typically embedded in a longer tone group, and generally functions as a discourse connective. Aston concludes by giving some practical examples of how this type of analysis can be used in the classroom: a) listening to segments and repeating them aloud, b) attempting to utter a complete segment before it is finished and c) reading segments aloud before hearing them. These may seem to be rather prosaic activities, involving somewhat routine patterns of language use. However Aston's point here is that one of the key skills which interpreters need to learn is timing and fluency, as well as the ability to predict what a speaker is saying. It therefore seems fair enough to teach students to familiarise themselves with the most recurrent features of a relevant genre or text type. Indeed this is what Aston and many corpus linguists mean by 'phraseology': naturally occurring idiomatic language, with all of its predictable patterns of speech and recurrent turns of phrase.

5. Stealing a march on collocation: Deriving extended collocations from full text for student analysis and synthesis

In this chapter, James Thomas gives an overview of the latest tools and methods which corpus linguists have been using to exploit corpora. The starting point is a brief discussion of "deviations" (Thomas' term for errors/mistakes) made by Czech-speaking trainee teachers who are following advanced EAP courses, as well as specialists in various domains (notably Informatics). Thomas then gives a detailed introduction to the well-known online corpus toolkit Sketch Engine, demonstrating functions which some readers may already be familiar with (such as the Word Sketch, which sets out the main grammatical relationships between a given word and its collocates), as well as less familiar tools, such as logDice ("lists of collocates where high-ranking items tend to accord with intuition"). Thomas then explains how he exploits these tools in the classroom. For pedagogical purposes, Thomas finds it useful to make a distinction between what he calls a) two-lexeme collocations and b) extended collocations. He gives several examples of how these items can be identified in the language classroom. For example, the item "scholarship" can be associated with an extended collocation such as "a scholarship is awarded [by an institution] [to a student] to study [a subject/skill] [somewhere]". Thomas then looks at how students can examine the extended collocations of key words throughout a single text. This paper seems to be full of new and useful concepts ("C+", "topic trails"…), although many of these ideas have already been explored elsewhere, but using different metalanguage (e.g. "extended collocations" are variously known as discontinuous frames, lexicogrammatical patterns, extended phraseological units, etc.).

6. A corpus and grammatical browsing system for remedial EFL learners

Kiyomi Chujo, Kathryn Oghigian and Shiro Akasegawa start off this chapter by pointing out the very low proficiency scores obtained on TOEFL and TOEIC tests by Japanese learners, as compared with many other countries. Although it might be thought that the introduction of DDL methods would have improved this situation, the authors claim that there are still many obstacles to this, not least a lack of appropriate teaching materials and methods which have been designed for learners at lower levels of proficiency (beginners and remedial). Chujo *et al.* aim to improve this situation by introducing a corpus-based platform named "Grammatical Pattern Profiling System" (GPPS). However, as it is presented in this chapter, the platform does not seem to address the needs of beginners: here it is used to search for examples of rather specialised grammatical patterns (examples cited in this chapter include: possessive nouns, passive voice and "subjunctive past/subjunctive wish"). Intriguingly, the authors claim that GPPS makes use of a specially-adapted corpus, the "Sentence Corpus of Remedial English" (SCoRE). This involves thousands of "sentences" (segments rather than whole texts) which have been especially selected and simplified, so that they correspond to the lower end of (USA) reading grades and word familiarity levels. Each sentence has also been translated into Japanese (using machine translation, and then manually corrected) so that learners and teachers can analyse them. The system therefore depends on invented data such as "These are the people (whom) I call my family" (p. 122). I admit to being somewhat perplexed by this approach. I was under the impression that the whole point of using a corpus was to examine authentic, naturally-occurring data (cf. John Sinclair's "trust the text"). To their credit, the authors themselves admit that this approach to naturally-occurring data may be limited (p. 124). They go on to point out that it is important for low-proficiency learners to be able to access clear, uncluttered instances of grammatical constructions. Considering the terrible language one sometimes finds on the internet, perhaps Chujo, Oghigian and Akasegawa are right after all: what matters for some learners is simplicity, clarity as well as (maybe) some artificial correction.

Part II. Corpora for skills development

7. Same task, different corpus: The role of personal corpora in EAP classes

In this chapter Maggie Charles looks at how a corpus-driven syllabus can be designed so that students of different disciplines can be taught in the same EAP class. Charles gives a very clear and precise example of how such a course can be structured (p. 135). In the first place, the students need to construct their own "do-it-yourself" corpus. They are then given the same generic exercise (Charles sets out a model worksheet in the appendix, p. 154). The exercise is so constructed that students of different specialities can be asked to achieve the same learning objectives, even when they have different data and different results. Since the exercise addresses features which are typical of EAP genres, such as reporting verbs and modals, it is usually possible for the students to find these features in their corpora. In the following sections, Charles describes how the students use the different tools provided by the AntConc programme to analyse their data (i.e.

Concordance, Wordlist, Collocates, Concordance Plot). Charles ends the paper with a discussion of her students' feedback. At one point (p. 147) her students encountered the recurrent problem of what to do with non-native-speaker texts: do they include them in the corpus? This leads to a very interesting discussion on perceptions of error, but although Charles' suggested answer ("increasing the size of the corpus may help") seems an obvious solution, this does not strike me as very convincing (how can you guarantee that you will not get more of the same?). Anyway, this discussion does lead Charles to make some very percipient remarks in her conclusion, especially the observation (p. 148) that by coming up with different results, the students are confronted with an "information gap" which may serve to fuel future discussion and perhaps foster future learning opportunities.

8. Textual cohesion patterns for developing reading skills: A corpus-based multilingual learning environment

In this chapter, Svitlana Babych discusses how students can be taught to identify textual cohesion patterns using an online learning environment. Her students are L1 (native) speakers of English at Leeds University who are studying Russian as their L2 and Ukrainian as L3. In the first part of the paper, Babych discusses the problems involved in analysing text connectors using traditional lexical concordancers, and looks at some of the linguistic literature, which has argued that corpus analysis should as far as practically possible maintain the visual, multimodal features of the texts which are being analysed. In the second half of the paper, Babych sets out how connectors are to be analysed using an online learning environment. The connectors identified for this project were collected automatically from a corpus of Ukrainian, Russian and English, and then categorized manually (p. 161). Babych then discusses how her class of language students were taught to use the online environment. This is a program which allows for the browsing of a comparable trilingual corpus of journalistic texts (English, Russian and Ukrainian). The aim is to integrate various activities such as "highlighting connectors", "exploring a multilingual thesaurus of connectors" or "summarizing cohesive profiles of a text as a frequency list of types and sub-types of its cohesive ties" (p. 169). These sound like valuable exercises indeed, especially since they focus on very specific sub-sets of reading skills. And some of Babych's practical recommendations for teaching activities are perfectly sensible, for instance: "One such activity might be to read through a text, deciding which statements are facts, and which are opinions, then analyze which words from the texts influenced their decision. They can also discuss what type of text it is, [...]" (p. 172). But on reading this and other suggestions, I could not help thinking "Well, OK, but this sounds like an old-school 'paper-and-pencil' exercise. Why would I need a corpus to do this?".

9. Exploiting keywords in a DDL approach to the comprehension of news texts by lower-level students

In this chapter, Alejandro Curado Fuentes explores how the analysis of keywords[1] may be of interest in the language classroom. Fuentes reports on an

[1] Keywords are lexical and grammatical items which are more statistically salient (i.e. occur with greater than expected relative probability) in a particular text type or register than in a comparable reference

experiment using fifty Spanish-speaking students of Business English divided into two groups. Both groups were taught to examine the basic phraseology and semantics of various keywords (including simple items such as < *defense* >, or more complex sequences such as < *prior defense secretary* >, < *Obama* + *say* > etc.). One group of students was trained in a computer lab on how to read concordances and corpus data. The other group underwent a more traditional course involving text-based reading comprehension skills. Both groups were given the same tests at pre-, mid- and post-experiment stages, with questions on: a) cultural knowledge, b) grammatical competence (identifying the correct translation of a keyword in co-text), and c) discourse cohesion (identifying the rhetorical function of a sentence). Fuentes provides statistical evidence to suggest that the corpus-users were "significantly better in post-tests" than the control group (p. 189). The experimental group also provided some interesting feedback about their experience: unsurprisingly, they often had trouble with formal linguistic analysis, and at early stages of the experiment they resented the complexity and messiness of concordance data. Later on, however, these same students seemed more positive about the tool (p. 193). In his conclusion, Fuentes provides some interesting reasons for the apparent success of DDL among these students: significantly, the experimental group liked not only the fact that the texts were authentic, but also that the language task seemed authentic. Fuentes also suggests that although DDL posed a significant challenge for some students, the complexity of the task also served to promote "increased engagement and participation, and therefore led to good results" (p. 194).

Part III. Corpora for translation training

10. Webquests in translator training: Introducing corpus-based tasks

In this chapter Teresa Molés-Cases and Ulrike Osier report on the "Corpus Valencià de Literatura Traduïda" (COVALT) and on a series of experimental translation classes in which Spanish-speaking (Castilian) and Catalan-speaking students of literary translation are trained to translate from English and German (the students are intermediate B1-2). The authors also introduce the "Webquest" approach to language leaning: this is a method of breaking down exercises into simple, interactive, research-based activities. From this point of view, the corpus is not the main focus of students' activities (indeed the teacher will often provide pre-analysed corpus material, thus avoiding the "technophobia problem" sometimes associated with corpus work). Rather, the main function of a Webquest is to get students to use a variety of online sources in order to accomplish a (relatively simple) task. Generally speaking, Webquests make use of two different types of resource. First, there are language-oriented resources, such as the bilingual dictionary Pons (<http://www.lingue.es>), as well as translation websites such as Linguee (<http://www.lingue.es>) and Linguatools (<http://www.linguatools.de>). Most language students and trainee translators are familiar with these tools. The second, and perhaps most important resource is the forum or common space where

corpus.

students can store, organize and present their findings. The authors mention many examples of this, from the well-known Moodle learning platforms, to less formal tools such as the Pinboard or Notepad functions from the Learningapps website (<http://learningapps.org>) or the website Voki which, intriguingly, allows students to create speaking avatars (<http://www.voki.com>). As the authors point out in their conclusion, a translation exercise should not be seen as a simple homogenous activity: rather it is "a gradual progression consisting in an initial phase of linguistic comprehension and production, an intermediate stage of reformulation and translation and a final phase of analyzing and comparing [...]" (p. 219).

11. Enhancing translator trainees' awareness of source text interference through use of comparable corpora

This chapter explores two related and rather difficult notions in translation studies: 1) interference: the extent to which a source text (and its source culture/language system) may affect a translated text in various ways, and 2) Toury's law (named after Gideon Toury): the hypothesis that a target text produced by an inexperienced translator generally displays more features of the source text than a text produced by an experienced translator. In the first part of this chapter, Josep Marco and Heike van Lawick examine a variety of studies which have either demonstrated or downplayed the importance of interference. As might be expected, no clear answers emerge, but this discussion does at least throw up some of the key issues at stake (translation universals, issues associated with language contact, the problems involved in matching text types, using comparable or parallel corpora, etc.). Marco and van Lawick conclude that, in the main, non-translations tend to resemble reference corpora more than translations, especially in specialised and technical domains.

In the second part of their paper, the authors set out a brief experimental study on the relative benefits of using a corpus in order to raise awareness about interference among trainee translators. As with the previous contribution, the subjects are Spanish- and Catalan-speaking students of literary translation at the Universitat Jaume I, Spain. The experiment involves a very clear set of tasks: 1) ask students to translate a short extract from English or German into Catalan (the source text in English was *The Great Gatsby* by F. Scott Fitzgerald, similar texts were used for German to Catalan), 2) provide students with an existing translation of a similar text (an equivalent literary translation), 3) ask the students to collectively identify a list of (undesired) examples of interference, at the same time as identifying the relative frequency of occurrence of these forms in a comparable corpus of narrative texts in Catalan, non-translated texts and translations from the same source language, 4) ask the students to work again on the professionally translated text, and finally 5) ask the students to go back to their own translation and to produce a revised version. Overall the authors report positive outcomes, with the students producing more natural-sounding, idiomatic translations (the students were asked for feedback, but were not given tests). The chapter ends with a frank discussion of the limitations of this kind of study, most notably pointing out the very small number of students involved, as well as the lack of statistical analysis on certain key features (such as the potential interference phenomena that may not

have been identified by the students in phase 3 of the study).

12. Using a multimedia corpus of subtitles in translation training: Design and applications of the Veiga corpus

In this chapter, Patricia Sotelo points out the many complex skills that are now expected of trainee translators, and suggests that training in subtitling can be a useful way of teaching many of these. In the first part of the paper, Sotelo points out that knowledge of information technology is now just as important to professional translators as knowledge of language. To support this, she cites the framework document of the European Master's in Translation (EMT), which outlines three core skills for all professional translators in the future: a) thematic competence (using databases and data-mining tools to explore specific domains and perform other language engineering tasks), b) information competence (using translation memory and multilingual corpora to perform computer-assisted translation) and c) technological competence (using software to edit and correct, modify or design all forms of electronic media). Sotelo then argues that the practice of subtitling constitutes probably one of the most IT-dependent branches of the translation industry, and thus provides an outlet for many of the requisite skills mentioned in the EMT framework. Sotelo also points out that there are many different subtypes of audio-visual translation, all of which have been created in response to new modes of media communication. This includes traditional activities such as dubbing and subtitling (translating foreign films), but also surtitling (for stage performances), audio description (voice-over for the blind) and intralingual subtitling (subtitles for the deaf, etc.). As far as interlingual subtitling is concerned, the skills required involve not only language comprehension, but also creative skills such as editorial judgment, creative timing and aesthetics. Although there are a handful of parallel and comparative corpora of subtitled media, very few of these make use of the original audio-visual material (for obvious technical and copyright reasons). In response to this problem, Sotelo reports on the Veiga multimedia corpus of subtitles, a corpus of over thirty films and programmes subtitled in English and Galician.

Unlike other projects reported in this volume, the aim of Sotelo's study is to familiarize students with the mechanics of audiovisual translation. Thus one of the first tasks in the project is to ask students to look for unusual typography, omissions or additions in either the English or the Galician corpus. The students are then encouraged to formulate hypotheses about why such features do not occur with the same frequency in each corpus (omissions are used when a character repeats him/herself; additions occur when the source text includes songs, and so on). In her conclusion, Sotelo admits that there is no empirical evidence to suggest that her students benefited directly from the use of the Veiga corpus. However, it seems to me that she has made a very good case for teaching subtitling not only to trainee translators, but to language students in general.

13. Applying data-driven learning to the web

In this final chapter, Alex Boulton rounds off the volume by arguing that it is perhaps time we started to see the web as a valid kind of corpus and the search

engine Google as a very basic type of concordancer, at least from the point of view of the novice language learner. Boulton admits that, if the web is used as a language corpus, it is often chaotic, ever-changing, and deeply skewed. And when Google is used a concordancer, it tends to be highly unreliable: it gives statistics which are by and large unverifiable, and its "snippets" (they can hardly be called concordance lines) cannot be viewed or manipulated in the same way as traditional concordancers. For this and other reasons, it is understandable that many corpus linguists still see the web and Google as second-rate resources. But while Boulton agrees with this, he then goes on to argue that the web and Google also have many advantages, especially from the point of view of the language learner. Firstly, the web offers immediate, free access to the very latest kinds of data, and in more massive quantities than has ever been available before. Secondly, Google offers the same basic functionalities as traditional purpose-built concordancing tools, but in a much more user-friendly environment. Finally (and this is something that even corpus linguists have been keen to exploit), the kind of language use that can be found on the internet involves its own very particular forms of discourse. As such the language of the web has already become the object of serious linguistic research (cf. the increasing amount of research currently being done on hybrid text types such as email, blogs, twitter, and so on).

In the second part of this chapter, Boulton shows some simple and disarmingly inventive ways in which it is possible to exploit the web for language learning. For example GoogleFight (<http://www.googlefight.com>) gives a graphic visualization of two or more competing phrases. And then there is the predictive text function of Google search produces grammatically correct but hilarious phrases, for example searching for the sequence < Can g > throws up hundreds of Chomskyesque interrogatives (often, for some reason, involving guinea pigs!). Such examples are an effective way of showing how almost any word in the language is typically associated with a rich set of very productive, but also highly predictable phraseological patterns. And finally, there is the extremely useful asterisk function which can be used to search for discontinuous phrase-stems, such as *"it has often been *ed that"* (note that in order to force Google to look for this exact sequence, the speech marks have to be left in). In conclusion, Boulton hypothesizes that it is already the case that many language users, including language professionals, translators and even corpus linguists make use of the web and the linguistic data that it throws up. So perhaps the time has come to see the web as a new kind of "reference corpus"?

General comments

There now exists an ocean of academic literature on corpus linguistics and teaching. If the readers of this review are anything like me, they may get a lot out of this collection, because it gives a flavour of what is currently going on. It is useful, for instance, to know just how far DDL and corpus-informed learning have progressed since the early days (by this I mean the 1990s, when people like me attended the first TaLC conferences). But it is also surprising – and also frankly reassuring – to learn how much the technology has stayed the same: the humble

concordancer does not seem to have evolved much, and in some instances seems to have regressed (in a good way, cf. the final chapter on Google searches).

Overall, I would recommend this book. It is a quality production. In terms of content, there are no poor chapters, although one or two contributions are weaker in terms of methodology. For example, as far as empirical research is concerned, some authors seem unwilling to present their results in ways that can be replicated, tested or debated. But the alert reader will spot these, and take them with a pinch of salt.

Finally, this collection does not push a particular theoretical approach, which is no bad thing. Here and there, there are challenges to how we see language (especially in the early chapters), and there are many useful reminders in the book about the importance of taking an empirical/data-driven/corpus-based perspective on language. However, as far as theory is concerned, although I come away from this book feeling generally enlightened, I also came away with the suspicion that many specialists in DDL have a bit of a blind spot, especially when it comes to any discussion of what Michael Halliday and others call "register" (cf. Halliday & Matthiessen 2014). As far as I understand it, 'register' corresponds to a particular constellation of formal lexico-grammatical features that can be used to characterise a particular text type. Usually it is not difficult to spot differences of register, even between superficially similar text types (e.g. the "popular science article" does not have quite the same configuration of features as the "academic research article", and this is not the same as the "oral conference presentation", etc.). Most DDL experts are often very careful to discuss genres, text types and the particular contextual biases which affect how their corpora or text archives are made up. However, they are not so keen to point out the formal linguistic features of these text types. And I am surprised when DDL analysts advocate using certain very traditional types of texts (such as "news reports" or "literary classics") as learning materials or even as reference corpora, without considering the particular linguistic features of these texts. There is also a certain shyness about discussing the particular types of text that their language learners may encounter in their professional lives. I suspect that this is because many researchers are not really familiar with language for specific purposes (LSP), and so they do not perhaps appreciate the issues involved in comparing a corpus of specialised texts with the general language, or with a corpus belonging to a different domain. Of course, for understandable reasons, many researchers do not have access to the kinds of language that are used in professional contexts. Nevertheless, I would still argue (as I did in the early days, see Gledhill 1998) that teachers and students need to be keenly aware not only of the specific genres they are likely to encounter in the professional world, but also of the specific lexico-grammatical forms they are likely to encounter. In other words, they need to be just as engaged in "learning a genre" as "learning (the rest of the) language".

Bibliographical references

Bordet, Geneviève & Elsa Pic. 2012. "Boulton, Alex, Shirley Carter-Thomas and Elizabeth Rowley-Jolivet (eds.). *Corpus-informed Research and Teaching in ESP"*. *ASp* 62, 109–115.

CHAU, Meng Huat. 2003. "Contextualising language learning: The role of a topic- and genre-specific pedagogic corpus". *TESL Reporter* 36/2: 42-54.

GLEDHILL, Christopher. 1998. "Learning a genre as opposed to learning French. What can corpus linguistics tell us?". In GEERTZ, W. & L. CALVI (eds.), *CALL, Culture and the Language Curriculum*. Berlin: Springer Verlag, 124–137.

HALLIDAY, Michael A. K. & Christian C. M. MATTHIESSEN. 2014. *Halliday's Introduction to Functional Grammar* (4rd Edition). London: Edward Arnold.

O'KEEFE, Anne, Michael MCCARTHY & Ronald CARTER. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

SINCLAIR, John McH. 1994. "Trust the text". In COULTHARD, M. (ed.)*, Advances in Written Text Analysis*. London: Routledge, 12–25.