

An experimental design perspective on the affordances of concordances:  
Exploring the affordances of concordances from a language learning perspective

by

Oliver James Ballance

A thesis  
submitted to the Victoria University of Wellington  
in fulfilment of the requirement for the degree of  
Doctor of Philosophy

Victoria University of Wellington

2017



## Abstract

Language learner use of concordances attracts considerable research interest. Concordances are now being conceptualized as presenting language learners with multiple affordances (Leńko-Szymańska & Boulton, 2015), affordances being the latent value of an object determined by the use made of it and not necessarily by its design. However, empirical research has typically operationalized concordances in simple, monolithic terms, and thus, from an experimental design perspective, they are implicitly seen as providing a single set of language learning affordances. That is, the majority of research conducted on concordancing treatments is discussed in terms of concordancing in fairly simple undifferentiated terms (see overviews of research in Boulton, 2010a; Cobb & Boulton, 2015; C. Yoon, 2011). Thus, previous research has contributed to an understanding of whether concordances can facilitate language learning, but it has rarely addressed the issue of how the operationalisation of concordances in pedagogical contexts interacts with the language learning affordances provided. Insufficient engagement with this issue is important because, alongside studies that have shown strong facilitative effects, there are many studies of learner use of concordances that show little to no facilitative effect of the concordancing condition. To address this issue, this thesis addresses the construct of learner use of concordances from an experiment design perspective, helping to define the construct of concordancing and examining potential variables in learner use of concordances.

The thesis reports a series of studies that examine the construct of concordancing from both the perspective of concordance users' concordancing preferences and the perspective of the lexical qualities of concordances as texts. First, a quasi-experimental, quantitative survey of concordance users' concordancing preferences showed the construct of concordancing to comprise four distinguishable operational parameters: citation format, type of corpus concordanced, citation order and reading style. It then found correlations between these

parameters and three of four user groupings: grouping by frequency of concordance use, by linguistic relationship to concordance language, and by field of concordance use. It revealed that different types of concordance user can be characterized by preferences for different types of concordance use, and vice versa. In two further studies, quantitative analysis of concordances showed that manipulation of factors in concordance generation resulted in concordances with significant differences in measures of word frequency and type token ratio. These analyses showed the extent to which the affordances of concordances vary in relation to two key factors in concordance generation: which corpus is concordanced and level of corpus generality.

The findings of these four studies are discussed in relation to the definition of concordancing presented in Chapter 2 of the thesis and the affordances of concordances that have been discussed in the literature. Together the findings indicate that effective learner use of concordances is likely to be dependent on matching the operationalization of concordancing in pedagogical contexts to learner profiles. For this reason, research on learner use of concordances needs to adopt experimental designs that can account for variation in concordancing as a treatment condition. The findings are also discussed in relation to their practical implications for effectively operationalising concordancing in pedagogical practice and the development of pedagogical concordancers and concordance-based language learning materials.

## Contents page

Abstract.....	i
Contents page.....	iii
List of tables.....	vii
List of figures.....	ix
Chapter 1. Introduction .....	1
Background: Concordances in professional and pedagogical contexts .....	3
Learner use of concordances, learner use of corpora or data-driven learning? .....	6
Importance: Issues in research on the effectiveness and efficiency of learner use of concordances.....	10
Aims: Developing an experiment design perspective on learner use of concordances .....	13
Research Questions .....	19
Chapter 3, Study 1: .....	19
Chapter 4, Study 2: .....	19
Chapter 5, Study 3: .....	19
Chapter 6, Study 4: .....	20
Organisation of the thesis.....	20
Chapter 2. Literature review .....	25
Introduction.....	25
What are concordances and how have they developed? .....	27
Changes in the ontological status of concordances.....	28
Changes in the epistemological status of concordances .....	30
A definition of concordance.....	33
The language learning affordances of concordances .....	36
Concordances as a research tool .....	38
Concordances as a source of concentrated input text .....	48
Concordances as an index to ad hoc language learning resources.....	58
Variables in learner use of concordances .....	64
The affordances sought .....	65
The task the learner engages in.....	65
The texts concordanced.....	68
Which citations are presented .....	69
The format citations are presented in.....	70

Dynamic versus static access .....	71
Overview .....	72
Chapter 3. Analysing concordancing: A simple or multifaceted construct? .....	75
Introduction .....	75
Variables in concordance use with theoretical significance .....	76
Variable: The corpus used .....	77
Variable: The format of the citation .....	78
Variable: The order of citations .....	80
Variable: The concordance-based language learning task .....	81
Research questions .....	82
Methodology .....	83
Conceptual framework of the survey .....	83
The survey instrument .....	83
Data collection and participants .....	86
Data analysis .....	91
Results .....	93
Principal component analysis of Data Set A .....	93
Principal axis factoring of Data Set B .....	94
Clarifying the factor solution .....	95
Discussion .....	97
Limitations .....	100
Summary .....	101
Chapter 4. Correlations between concordance users and concordancing preferences .....	103
Introduction .....	103
Four ways of categorising concordance users .....	104
Factor scores .....	116
Research questions .....	116
Methodology .....	117
Data analysis .....	117
Results .....	118
MANOVA for <i>Length of use</i> .....	118
MANOVA and discriminant function analysis for <i>Frequency of use</i> .....	119
MANOVA and discriminant function analysis for <i>Language of use</i> .....	121
MANOVA and discriminant function analysis for <i>Field of use</i> .....	123

Discussion .....	125
Length of use.....	126
Frequency of use .....	126
Language of use .....	127
Field of use.....	129
Limitations .....	130
Summary .....	131
Chapter 5. Exploring the authentication burden of concordances through word frequency measures.....	133
Introduction.....	133
Literature review .....	134
Existing approaches to concordances as difficult texts: grading outputs and inputs .....	134
Word frequency as a gauge of readability.....	136
Research on texts, word frequencies and vocabulary load .....	137
Research questions.....	140
Methodology .....	140
Developing word frequency lists .....	141
Developing corpora.....	145
Sampling frame .....	146
Generating data .....	150
Scoring .....	150
Results.....	153
Experiment 1: Testing corpus concordanced as a variable in the vocabulary load of citations .....	154
Experiment 2: Retesting corpus concordanced as a variable in the vocabulary load of citations .....	156
Experiment 3: Comparing citations from genre corpora with citations from a graded reader corpus, estimating practical effect sizes.....	159
Discussion .....	162
Limitations .....	164
Chapter 6. Exploring lexical repetition in concordances .....	167
Introduction.....	167
Literature Review.....	168
Type token ratio and more advanced measures .....	168
Type token ratio as a measure of lexical variation and repetition.....	169

Type token ratio as a measure of authentication burden in LSP settings.....	170
Type token ratio and lexical patterning affordances .....	174
Type token ratio and corpus composition .....	175
Research questions.....	176
Methodology .....	176
Corpora used .....	176
Data generation .....	178
Scoring .....	179
Results.....	180
Analysis of variance for items 1-50.....	180
Analysis of variance for items 51-100.....	181
Discussion .....	182
Limitations .....	185
Chapter 7. Discussion and conclusion .....	189
Summary of findings.....	189
Study 1. Analysing concordancing: a simple or multifaceted construct? .....	189
Study 2. Correlations between concordance users and concordancing preferences .....	190
Study 3. Exploring the authentication burden of concordances through word frequency measures.....	191
Study 4. Exploring lexical repetition in concordances .....	193
Implications for theory.....	194
Prototype 1. Analysis .....	196
Prototype 2. Reference.....	199
Implications for research.....	201
Methodological implications, limitations and reflections.....	203
Survey research .....	204
Lexical analysis research .....	207
Future directions for research .....	209
Recommendations for pedagogical practice .....	210
Appendices.....	215
Appendix 1: Screenshots of the questionnaire.....	215
List of references.....	231



## List of tables

Table 1 Affordances sought.....	66
Table 2 Task engaged in.....	66
Table 3 Texts concordanced.....	69
Table 4 Citations shown.....	70
Table 5 Format citations are presented in .....	71
Table 6 Dynamic or static access .....	72
Table 7 Hypothetical dimensions of concordancing conceptualised as dichotomies .....	84
Table 8 Survey outline .....	85
Table 9 Respondents' backgrounds.....	87
Table 10 Respondents' 1st language/s .....	88
Table 11 Content words occurring 10 or more times and their collocates.....	89
Table 12 Principal component analysis of Data Set A (n = 91) .....	93
Table 13 Structure matrix of principal axis factoring Data Set B (n = 90).....	95
Table 14 Principal axis factoring of the complete sample (n = 181) .....	96
Table 15 Factor loadings of items for complete sample (n = 181) .....	97
Table 16 Frequency count for scores on Length of use .....	105
Table 17 Frequency count for scores on Frequency of use.....	107
Table 18 Frequency count for Language of use.....	110
Table 19 Frequency count of other domains of concordance use in teaching (n = 35) .....	113
Table 20 Field of use by percentage of time spent.....	115
Table 21 Frequency count for Field of use as a variable in this study.....	115
Table 22 Descriptive statistics for factor scores .....	116
Table 23 Descriptive statistics for Length of use in relation to Format, Corpus, Order and Task .....	118
Table 24 Descriptive statistics for Frequency of use in relation to Format, Corpus, Order and Task .....	119
Table 25 Discriminant function analysis eigenvalues: Frequency of use .....	120
Table 26 Discriminant function analysis Wilks' Lamda: Frequency of use.....	120
Table 27 Discriminant function analysis functions at group centroids: Frequency of use .....	120

Table 28 Discriminant function analysis structure matrix: Frequency of use .....	120
Table 29 Descriptive statistics for Language of use in relation to Format, Display, Corpus, Order and Task .....	121
Table 30 Discriminant function analysis eigenvalues: Language of use .....	122
Table 31 Discriminant function analysis Wilks' Lamda: Language of use .....	122
Table 32 Discriminant function analysis functions at group centroids: Language of use .....	122
Table 33 Discriminant function analysis structure matrix: Language of use.....	122
Table 34 Descriptive statistics for Field of use in relation to Format, Corpus, Order and Task .	123
Table 35 Discriminant function analysis eigenvalues: Field of use.....	124
Table 36 Discriminant function analysis Wilks' Lamda: Field of use .....	124
Table 37 Discriminant function analysis functions at group centroids: Field of use.....	125
Table 38 Discriminant function analysis structure matrix: Field of use .....	125
Table 39 Approximate proportions of different word types at different frequency levels in texts (G. Kennedy, 1998, p. 96).....	138
Table 40 Search terms selected for Citation Set 1: Academic prose, Fiction, News and Graded readers .....	149
Table 41 Search terms selected for Citation Set 2: Academic prose, Fiction, News .....	149
Table 42 Citation Set 1 scores on VCL and MWF.....	154
Table 43 Summary of effect sizes observed in Mann-Whitney U tests in Experiment 1 .....	156
Table 44 Citation Set 2 scores on VCL and MWF.....	157
Table 45 Summary of effect sizes observed in Mann-Whitney U tests in Experiment 1 & 2 ....	158
Table 46 Citation Set 1 scores on VCL and MWF compared with Graded Reader scores.....	160
Table 47 Summary of effect sizes for Mann-Whitney U tests in Experiment 3 .....	161
Table 48 Summary of corpora used .....	177
Table 49 Search terms used in STTR studies.....	179
Table 50 Summary of STTR comparisons for Level of corpus generality .....	181
Table 51 Analysis and reference as composites of concordancing preferences in relation to aspects of concordancing hypothesised as harder or easier .....	195

## List of figures

Figure 1 Percentages for Language of use: Yes, Occasionally, No [n = 181].....	109
Figure 2 Percentages for Field of use: Yes, Occasionally, No [n = 181] .....	111
Figure 3 Percentages of concordance use in teaching by domain [n = 124].....	112
Figure 4 Percentages for using concordances in teaching: Directly, Indirectly or Both [n = 124] .....	114
Figure 5 Two models of concordance use with prototypical aspects of concordancing in relation to selected categories of concordance user with means of group centroids .....	196



## Chapter 1. Introduction

Despite the profound influence of concordancing software on how we think about language and intense research interest in the possible benefits of using concordances with language learners (see overviews of research in Boulton, 2010b; Cobb & Boulton, 2015; C. Yoon, 2011), it is commonly held that concordances are rarely used by learners in mainstream language teaching contexts (Ädel, 2010; Boulton, 2008; Braun, 2005, 2006, 2007; Cotos, 2014; Flowerdew, 2012a; Götz & Mukherjee, 2006; Mukherjee, 2006; Pérez-Paredes, 2010; Römer, 2010, 2011; Tono, Satake, & Miura, 2014; Tribble, 2013). Indeed, when teachers have been surveyed, there appears to be little enthusiasm for learner use of concordances in the majority of pedagogical contexts, such as English as a Foreign Language (EFL) contexts (Leńko-Szymańska, 2014; Mukherjee, 2004; Tribble, 2015). This mismatch between researcher interest and teacher uptake represents something of a paradox, and it is itself an issue that has begun to receive an increasing amount of research interest. One possible explanation is that there is insufficient teacher training in relation to learner use of concordances (Braun, 2007; Flowerdew, 2012a; Mukherjee, 2004, 2006; Römer, 2010, 2011; Tono et al., 2014; Tribble, 2013). Limited access to resources such as computers, software and suitable electronic texts is another possible explanation (Ädel, 2010; Boulton, 2008, 2010a; Breyer, 2011; Gavioli & Aston, 2001; Römer, 2010). In addition, thinking about this issue more broadly, it may be that the cognitive complexity of concordance use conflicts with the learning styles and language proficiency levels of many language learners (Ädel, 2010; Boulton, 2010a; Braun, 2007; Coxhead, 2014; Flowerdew, 2012a, 2012b; Gavioli, 2001; Hunston, 2002; C. Kennedy & Miceli, 2001; Levy, 1990; O'Sullivan & Chambers, 2006; Partington, 2001; Pérez-Paredes, 2010; Pérez-Paredes, Sánchez-Tornel, & Alcaraz Calero, 2012; Stevens, 1991a; Sun, 2003; Varley, 2009; H. Yoon & Hirvela, 2004). This thesis takes a slightly different approach to the apparent paradox and suggests that some of the problem may be insufficient engagement with the construct of concordancing from the language learning perspective.

The thesis takes a step back from research that involves direct observation of learners using concordances and instead takes the construct of concordancing as its object of study. The aspects of this construct that are explored are selected on the basis of a review of the literature concerning learner use of concordances (and hence the subtitle of the thesis *Exploring*

*concordances from a language learning perspective*), but the studies reported do not involve observation of learner use; rather, the aim of the studies reported is to identify and explore variables within the construct of concordance use that have relevance to language learning (and hence the thesis' main title, *An experimental design perspective on the affordances of concordances*). In other words, to address the apparent paradox of exponential interest in learner use but minimal uptake, the thesis reports four studies, each of which contributes to our understanding of variables that are relevant to learner use of concordances, variables that may go some way towards explaining the apparent mismatch between interest and uptake.

The results of the studies reported in this thesis have implications for future research on learner use of concordances and the operationalisation of learner use of concordances in pedagogical contexts. Specifically, the studies reported here complement existing research with evidence-based accounts of the construct of concordancing, evidence of correlations between types of concordance user and types of concordance, measures of concordance vocabulary load as a correlate of which corpus is concordanced, and measures of lexical variation/repetition in concordances as a correlate of the level of relative homogeneity/heterogeneity of the corpus concordanced. The findings help to clarify the relationship between how the construct of learner use of concordances is operationalised and the potential language learning affordances of concordances. The thesis argues that a key to increased uptake of learner use of concordances is recognition of the fact that concordance use is a complex, multifaceted construct, different operationalisations of which may be offering different, sometimes competing affordances. It also underlines the scope for pedagogical mediation of concordances that is available through controlling the characteristics of the texts concordanced, providing preliminary estimates of the extent to which these factors can predict the linguistic difficulty of concordances on two key types of measure: word frequency measures and type-token ratio. This is important for facilitating the transfer of research on learner use of concordances to broader pedagogical contexts. In terms of research itself, the findings underline the need for experiment design to control for or manipulate more variables in how concordancing is operationalised, to report the details of how learner use of concordances has been operationalised in greater detail and for the results of many studies to be interpreted far more narrowly. The thesis also provides some preliminary indications of how decisions made in regards to the details of learner use of concordances are likely to interact with their affordances.

This chapter presents an overview of how concordances have been used in professional and pedagogical contexts and a critique of a recent meta-analysis of the effectiveness and efficiency of learner use of concordances. It then problematizes the research paradigm from an experiment design perspective, arguing that research to date has limited generalisability because of the interaction of insufficient construct definition and the assumptions of the various models of experiment design available. It calls for research to move beyond the question of whether learners can use concordances successfully (it seems pretty clear by now that they can) and to start charting salient variables within learner use of concordances.

### **Background: Concordances in professional and pedagogical contexts**

A concordancer allows a user to quickly and effortlessly search through an electronic text or text-database indexing every occurrence of a user-specified, computationally definable form: a word, a group of words, a morpheme, a combination of orthographic characters and wild cards or regular expressions, even a particular user-defined tag (assuming that is that the text itself has first been annotated with a corresponding set of tags). However the form has been defined, each instance of its occurrence within the text searched can be termed *a citation*, or more informally a *hit* or a *line*. The index of these citations that has been generated by the concordancer can then be used to compile a new text: first, one or more citations are extracted along with a prespecified span of its co-text (the words that co-occurred with each citation of the indexed form); second, the extracted citations are then compiled and returned to the user as a new text. This new text is composed entirely of these discontinuous citations, and this type of discontinuous text has revolutionised scholarly thought about language and language description (cf. Tognini-Bonelli, 2010).

The effect of concordances on language theory and language description cannot be overstated. It is not uncommon to see corpus linguistics described in terms of a revolution (Braun, 2005; Flowerdew, 2012a; Mukherjee, 2006; Römer, 2011), and concordancing has played a key part in this revolution. In *Corpus, Concordance, Collocation*, Sinclair (1991, p. 100) remarks that, “language looks rather different when you look at a lot of it at once”. Of course, it is not that case that prior to concordancing software there was in some sense a shortage of language data that could be viewed simultaneously; rather, it is that the process of extracting

citations and viewing them simultaneously has revealed underlying regularities to language use that had previously escaped description. Many regularities in how a linguistic form is used are ordinarily masked by the individuality of a form's use in a particular context and the interspersed of large quantities of weakly related forms between each occurrence of the repeated form; by removing the majority of intervening forms, concordances highlight the regularities of usage in a form's immediate context, regularities that belie the individuation of usage in each particular context of use. Thus, it is not so much that language looks different *when you look at a lot of it at once* (great libraries have existed since antiquity); rather, it is that language looks rather different *when concentrated*. This concentrated observation of language was brought about by the use of concordances, and examining bodies of text through concordances has been fundamental to the development of important new theories of language: for instance, John Sinclair's *Idiom Principle* (Sinclair, 1991), Michael Hoey's theory of *Lexical Priming* (Hoey, 2005), and Patrick Hanks' *Theory of Norms and Exploitations* (Hanks, 2013). The point to be made here is that the view of language afforded by concordances was instrumental in the development of these theories.

It is also worth noting that concordances have not only had a profound effect on how language is theorised, but also on more practically orientated language work. For instance, modern grammars and lexicographical projects make extensive use of concordances (Frankenberg-Garcia, 2014a; Hanks, 2008, 2013; Mukherjee, 2006). Concordances are also increasingly being seen as an essential tool for professional translators (Kübler, 2011; Laviosa, 2012). Indeed, it seems that concordances have become a fundamental part of the language worker's tool kit. Thus, language learning materials are now increasingly commonly informed by corpus evidence (Boulton, 2010a; Braun, 2006; Cotos, 2014; Frankenberg-Garcia, 2014a; M. McCarthy & McCarten, 2012; Mukherjee, 2006; Römer, 2010), and so authors are very likely to have drawn upon concordancing for one purpose or another during the production of language learning materials.

Alongside the prodigious influence of concordancing on language description and other professional interest in concordancing, there is also a long history of interest in the direct exploitation of concordances with language learners. That is, not abstracting from concordance data useful descriptions of language which then inform what learners are taught or the materials that are written for them (indirect use), but presenting learners with a concordance itself (direct



use). This distinction, often attributed to Geoffrey Leech, was popularised by Ute Römer, although the definition provided here differs slightly from Römer's treatment as I would wish to explicitly exclude teacher use of concordances in materials production from *direct* use (cf. Römer, 2011). This thesis is interested in learner use of concordances, not learner use of concordance-informed materials. Examples of learner use of concordances as defined here abound in the literature: some indicative examples are given below.

The 1980s saw a number of publications that began to explore the potential of using concordances with language learners (Honeyfield, 1989; Johns, 1986; Leech & Candlin, 1986; McKay, 1980). Since the 1980s, the publication rate has been steadily increasing and there are now hundreds of publications that deal with learner use of concordances (Boulton, 2010b; Cobb & Boulton, 2015; Daskalovska, 2013; Römer, 2011). Some particularly noteworthy early publications would include papers from Tim Johns, who coined the term *data-driven learning* (DDL) (Johns, 1991a, 1991b, 2002), Chris Tribble and Glyn Jones' *Concordances in the classroom: A resource guide for teachers* (Tribble & Jones, 1990), pioneering studies of the effectiveness of learning vocabulary with concordances from Tom Cobb (1997, 1999), Jennifer Thurston and Christopher Candlin's concordance-based vocabulary text book, *Exploring academic English: A workbook for student essay writing* (Thurston & Candlin, 1997, 1998), and the proceedings of the biannual Teaching and Language Corpora conference (TaLC) (for bibliographical details as far as 2014, see Boulton & Pérez-Paredes, 2014, p. 122). It is also worth mentioning the increasing availability of pedagogically orientated concordancers. Tom Cobb's website, [www.lextutor.ca](http://www.lextutor.ca), has always had a pedagogical orientation, integrating its concordancing functionality with a number of language learning exercises and other pedagogical exploitations of concordancing. More recently, new pedagogically orientated concordancing websites have emerged, such as <https://www.sketchengine.co.uk/skell/>, [www.just-the-word.com](http://www.just-the-word.com) and <http://flax.nzdl.org/greenstone3/flax>.

There are of course numerous other publications and resources that could be mentioned, especially once we recognise that much research on *learner use of concordances* is framed in terms of *learner use of corpora* or *DDL*. For instance, a recent meta-analysis of studies of *corpus-based learning* identified "132 papers which seek to evaluate some aspect of corpus use in foreign or second language (L2) learning and teaching" (Cobb & Boulton, 2015, p. 483). The vast majority of these papers report on the effectiveness of concordance use, as opposed to

studies of pedagogical exploitation of other corpus tools such as word lists or distribution plots. Indeed, almost all research on learner use of corpora or DDL is centred upon use of concordances. Consequently, because these terms are so often used interchangeably in the literature, this thesis treats papers that report exclusively on learner use of concordances (that is, papers that do not mention learner use of other corpus tools) as papers about *learner use of concordances*, regardless of how they have been framed. Similarly, in drawing upon the vast theoretically orientated literature, the framing of a discussion in terms of *corpora* or *DDL* does not disqualify it as relevant to learner use of concordances as long as it is clear in the text that the authors are indeed discussing learner use of concordances. However, before explicating the approach taken in this thesis in greater detail, it is perhaps necessary to pause here and untangle the subtle relationships that exist between these various terms.

### **Learner use of concordances, learner use of corpora or data-driven learning?**

A considerable proportion of this thesis concerns the definition of *concordancing* as a theoretical construct (see Chapters 2 and 3); however, in the way of some preliminary remarks, we can observe that, in the information age, a concordance is typically thought of as a text generated by a concordancer. As described above, a concordancer is a piece of software that searches through an electronic source-text, indexes occurrences of a user-specified form, extracts citations of said form and then compiles a new text composed of the citations extracted. I use the term *source-text* here, because it is not necessary for the text concordanced to be a corpus; a concordancer can produce a concordance of any piece of suitably formatted electronic text, even one single text. Any use of a text that has been generated via a concordancer by a learner can be termed *learner use of concordances*; the particular characteristics of the source-text are not relevant to this term so defined.

*Learner use of concordances* is often conflated with *learner use of corpora*, though the difference in reference is of course apparent on reflection. Learner use of concordances refers to use of a text that has been generated through concordancing; learner use of corpora refers to use of whatever is designated by the term *corpus*. Thus, the distinction between learner use of concordances and learner use of corpora is complicated to some extent by inconsistent use of the term *corpus*. Many researchers maintain a distinction between a corpus and other collections of texts (Adolphs, 2006; Biber, Conrad, & Reppen, 1998; Biber & Reppen, 2015; Cheng, 2012;

Davies, 2015; G. Kennedy, 1998; Reppen, 2010a; Sinclair, 1991; Teubert & Čermáková, 2007). The point of distinction is that, in a strict sense, a corpus is a collection of texts that has been compiled so as to provide a representative sample of a variety of language. From their perspective, to the extent that the composition of a collection of texts successfully provides a representative sample of a variety of language, it can be considered a corpus. If a collection of texts does not warrant claims of representativeness, it is not a corpus, just a collection of texts: a text archive or a text database. However, not all researchers adhere to this terminological distinction, and much discussion of learner use of concordances involves concordancing collections of text that are referred to as corpora but would not meet a strict definition of a corpus as a collection of texts that warrant a claim to represent a variety of language (Aston, 2001; Bernardini, 2004; Braun, 2005, 2006; Chan & Liou, 2005; Cobb, 1997, 1999; Gavioli, 2005; Lee & Swales, 2006; Partington, 2001; Stevens, 1991a; H. Yoon & Hirvela, 2004). Of course, this is irrelevant to whether such research constitutes research on learner use of concordances, although the corpus-database distinction does have important implications for the affordances of the concordances used in such studies (these implications are discussed in Chapter 2). At present, it will suffice to say that both the strict and loose definitions of corpora refer to collections of texts, with the implication being that the source-text concordanced must be more than a single text; however, as the definition of learner use of concordances above makes clear, it is possible for a learner to make use of a concordance of even a single text. In fact, a number of key figures have indeed discussed learner use of concordances of single texts (Aston, 2001; Götz & Mukherjee, 2006; Johns, Hsingchin, & Lixun, 2008). Thus, to reiterate, learner use of concordances refers to learner use of a text that has been produced through concordancing, while learner use of corpora does not refer to how a text has been produced, but rather, to learner use of a particular type of text, however that text type is defined. Although a concordancer is the primary qualitative means by which a corpus can be interrogated (Hyland, 2015; Rayson, 2015), learner use of concordances does not necessarily entail use of a corpus, and learner use of a corpus does not necessarily entail use of a concordance.

Both learner use of concordances and learner use of corpora are sometimes glossed as *data-driven learning* (DDL). However, again, reflection on these terms does not appear to support identity between any of these terms. To conflate DDL with either of these terms would be to mistake learner use of various types of texts with a particular way of using a text. Boulton

(2011) provides a thorough examination of the term *data-driven learning*, but concludes that there appear to be neither necessary nor sufficient conditions for identifying a particular pedagogical activity as an instance of data-driven learning. Instead, he recommends a proto-type account of DDL in which DDL is constituted by more or less adherence to certain core features of DDL in general:

In the case of data-driven learning, we might propose a prototype such as the following: The hands-on use of authentic corpus data (concordances) by advanced, sophisticated foreign or second language learners in higher education for inductive, self-directed language learning or advanced usage. Yet virtually any part of this description might be absent in a particular case [p. 572]

This rather amorphous definition, while capturing the range of research described as data-driven learning effectively, can perhaps be narrowed down considerably if our concern is not to accommodate the full range of research that has been presented under this guise. Tim Johns, the originator of the term, seems to have explicitly used the term to cover an eclectic range of language-learning activities connected by a common approach that aimed “to confront the learner as directly as possible with the data, to make the learner a linguistic researcher” (Johns, 2002, p. 108). Defined this way, it seems clear that DDL is not constituted by use of particular resources (concordances or corpora), but by the learners’ orientation to a language input. From this perspective, both concordancers and corpora are a means to an ends in data-driven learning: corpora provide a desirable quantity and quality of data, concordances provide an effective means of martialling the data, but DDL primarily concerns the learner’s orientation to the data, not the use of concordances per se. Reference to DDL in this thesis will follow John’s relatively restricted definition as, in combination with the definitions of concordance use and corpus use presented above, it promotes conceptual clarity in distinguishing the wide range of pedagogical activities that can be referred to by such terms.

This thesis is written from the perspective that concordancing may afford language learning opportunities irrespective of whether it is a corpus that is concordanced, or whether the learner approaches the concordance as a linguistic researcher or something else. Hence, returning to the paradox of learner use of concordances, one possible explanation for the relatively slight uptake of concordances in pedagogical contexts may be its blurring with other related but clearly

distinguishable pedagogical movements. Discussion of the relative merits of true corpora (in the sense of representative corpora) in comparison with other collections of text selected on pedagogical grounds are common place in the literature (Braun, 2005, 2006, 2007; Gavioli, 2001; Lee & Swales, 2006; Levy, 1990; Mishan, 2004; Partington, 2001; Römer, 2010, 2011). John's metaphor of *learner as researcher* has been expanded to cover *hard* and *soft* versions of DDL (Gabreilatos, 2005), the former requiring far more researcher like behaviour than the latter, and other scholars have suggested alternative metaphors such as *learner as traveller* (Bernardini, 2004), *learner as chef* (C. Kennedy & Miceli, 2010) or *learner as spy* (Gavioli, 2005). Different ways of conceptualising learner use of concordances are discussed in detail in Chapter 2, but without going into the details of each suggestion here, the point to note is that each of these alternative conceptualisations focuses on learner use of concordances, but each repositions the learner's relationship to the data presented in the concordance vis-à-vis Tim John's original suggestion. In other words, it is possible that a proportion of resistance to learner use of concordances in many pedagogical contexts may be related to assumptions about how concordances may be exploited by learners: for instance, the assumptions that concordances can only be used with corpora or that the best way for a learner to approach a concordance is from the perspective of a linguistic researcher. While each of these operationalisations of concordancing may represent a facet of optimal operationalisation of concordancing in one particular pedagogical contexts, each is only one of many possible operationalisations of concordancing in pedagogical contexts. Which operationalisations of concordancing are most suited to which learners in which contexts is of course an empirical question, and very much the driving force behind this thesis.

The fundamental question that this thesis engages with is whether learner use of concordances is best conceptualised as use of a single, simple language learning resource or, alternatively, as a complex, multi-dimensional construct. In the former view, apparent differences could perhaps be accommodated under a prototype approach comparable to Boulton's (2011) suggestion that data-driven learning is best characterised by a prototypical realisation, with particular operationalisations of concordancing conforming to the prototypical realisation to various degrees. In the latter, concordancing is more appropriately conceptualised as a multidimensional construct within which there are competing realisations of concordancing that are sometimes incompatible and frequently offering a range of different language learning

affordances to varying extents. The thesis provides evidence in favour of the latter account, and argues that progress in the field of learner use of concordances depends on researchers engaging more deeply with concordancing as a complex theoretical construct. Indeed, the first year of my PhD was spent trying to design a fairly typical study of the effectiveness of learner use of concordances, but the process of reviewing the relevant literature, attempting to construct an appropriate experimental design and attempting to produce and specify appropriate experimental treatments and conditions led me to question whether the construct of concordancing has been defined effectively enough for new research on the effectiveness of learner use of concordances to provide significant new insights. The next section will discuss existing research on learner use of concordances from this perspective.

### **Importance: Issues in research on the effectiveness and efficiency of learner use of concordances**

There is an extensive body of both qualitative and quantitative research on learner use of corpora and data-driven learning (see Chapter 2), the vast majority of which concerns learner use of concordances (see overviews in Boulton, 2010b; Cobb & Boulton, 2015; C. Yoon, 2011). Due to the large number of studies that have been published, meta-analysis is a useful way of synthesising quantitative results in this field. Cobb and Boulton (2015) identify 116 individual studies published between 1989 and 2012, and then present a meta-analysis of the 21 studies that met their criterion for inclusion within the analysis: that is, studies that reported a quantitative learning outcome or increased capacity to use language successfully, that included either a pre-test/post-test design (studies of *effectiveness*) or else a treatment-group/control-group design (studies of comparative *efficiency*), and that reported essential data such as groups sizes, means, and standard deviations. The meta-analysis of the eight studies addressing *effectiveness* reported a mean effect size of 1.68 (SD = .84, 95% [1.36, 2.00]); the meta-analysis of the thirteen studies addressing *efficiency* reported a mean effect size of 1.04 (SD = .73, 95% [0.83, 1.25]). These results indicate that the learner use of concordances reported in these studies was generally effective, and was generally more efficient than the learning conditions used as control conditions in these studies. Thus, in many respects, these results just underline the mysterious

mismatch between research evidence in favour of learner use of concordances and limited uptake in mainstream pedagogy, an issue that has frequently been discussed, as indicated above. However, it may also be that this existing body of research on learner use of concordances has problems with interpretability and transfer to practice.

As useful a technique as meta-analysis is for synthesising the results of separate studies, it is only as useful as it is interpretable. Ross and Mackey (2015, p. 220) make the following observations:

A key advantage of meta-analysis is that it provides the basis for what researchers can expect to find if studies similar to those in the meta-analysis are replicated. When meta-analyses result in a nonzero effect size, further experimentation anticipating a zero-effect outcome, that is, one that establishes significance in relation to a null hypothesis, eventually becomes superfluous.

Thus, one might be tempted to conclude from the Cobb and Boulton meta-analysis that there is conclusive evidence of the effectiveness and efficiency of the types of studies included in the analysis. However, two important issues in meta-analysis complicate this interpretation. First, as Cobb and Boulton, acknowledge, their analysis was only able to include 18% of the studies that they identified. This is unfortunate, as it is not difficult to identify studies that were not included in the meta-analysis and report non-significant results (Allan, 2006, 2010; Boulton, 2009b, 2010a; Kaur & Hegelheimer, 2005). Indeed, this issue is especially problematic considering the *file drawer problem* that all meta-analyses are vulnerable to.

In the *Handbook of parametric and nonparametric statistical procedures*, Sheskin (2004), defines the file drawer problem as the extent to which a meta-analysis is biased by systematic differences in which studies are available to the meta-analyst. He discusses this problem from the perspective of the bias introduced by differences in the results obtained in studies that are submitted and accepted for publication in comparison to the results obtained in studies that were conducted but never submitted for publication or were submitted but rejected. This concept of bias introduced by unavailable studies can also be extended to studies that were available to researchers but had to be excluded due to problems with their design or reporting: around 82% of the studies identified by Cobb and Boulton as potential candidates for inclusion. The equations for calculating exactly how many such excluded studies would be needed to

significantly alter the result obtained from a meta-analysis are quite complicated (for a brief discussion see Sheskin, 2004, p. 1038); however, only being able to retain a small proportion of the studies that were available, together with a significant number of null-result studies that were available but could not be included in the analysis, suggests that the results reported by Cobb and Boulton may be particularly vulnerable to bias. Indeed, the authors make it clear that they are aware of this issue and that they consider their results to be those of a preliminary study and not conclusive proofs of effectiveness or efficiency.

The second problem with casual interpretation of this meta-analysis concerns the broad range of studies included (Cobb & Boulton, 2015, p. 494). As the quote from Ross and Mackey indicates, the transferability of meta-analysis results is contingent on similarity between the studies in the meta-analysis and the situation the results are to be transferred to. In fact, meta-analysis assumes homogeneity amongst the studies included in the analysis, and differences between the studies in a meta-analysis constitute a significant methodological challenge:

Pooling the results of multiple studies that evaluate the same hypothesis is certainly not a simple and straightforward matter. More often than not there are differences between two or more studies which address the same general hypothesis. Rarely if ever are two studies identical with respect to the details of their methodology, the quality of their design, the soundness of execution, and the target populations that are evaluated. (Sheskin, 2004, p. 1025)

The implication is that a meta-analysis conducted on a diverse group of studies is correspondingly difficult to interpret: the more diverse the studies pooled in the meta-analysis are, the less clear it is which aspects of the studies constitute the common core that the pooled results apply to. Hence, Cobb and Boulton (2015) suggest that their readers not only attend to the overall results of their analysis, but to differences in the effect sizes of the studies included; they suggest such a reading of their meta-analysis as a useful step in identifying which sub-variables may be operative within studies.

In summary, Cobb and Boulton (2015) is probably best considered as providing a clear and convincing indication that learner use of concordances *can* be very effective and efficient from both a learning outcomes and a learner use of language perspective, but it cannot be seen as providing a reliable indication of the outcomes of learner use of concordances *in general* because



of the likelihood of bias introduced by the number of studies that had to be excluded from the analysis and the heterogeneity of the studies that were included. Consequently, there remains a need for continued research on learner use of concordances, particularly research that moves beyond the question of whether learner use of concordances can be effective or efficient and towards research designs that will support more robust generalisation. That is, for research on learner use of concordances to engage more successfully with principles of experiment design (Cobb & Boulton, 2015).

### **Aims: Developing an experiment design perspective on learner use of concordances**

This thesis explores learner use of concordances from the perspective of experiment design. Specifically, it explores the construct of concordancing from both theoretical and empirical perspectives, and then examines the lexical qualities of concordance-texts from the perspective of treatment levels within experiment design. The rationale for these studies is quite straightforward. A core principle of research design is that any potential source of variability that may have an effect on the outcome of a study should be accounted for (Hudson & Llosa, 2015, p. 94). Potential sources of variability should first be identified through thorough construct definition, and then controlled for or manipulated via research design. This is important because every source of variability that is left unaccounted for represents a limit to the generalisability of the findings reported (Purpura, Brown, & Schoonen, 2015; Winer, Brown, & Michels, 1991). Thus, the issues explored here are akin to the issues discussed in the *language as fixed effect fallacy* (Clark, 1973), Clark's provocative critique of how language prompts are operationalised in experiment design. Indeed, the full implications of the language as fixed effect fallacy continue to attract much interest in statistically minded discussion of applied linguistics (Johnson, 2008; Larson-Hall, 2010; Norris, Ross, & Schoonen, 2015; Woods, Fletcher, & Hughes, 1986). In brief, Clark observed that the particular words or sentences used in an experiment are rarely exhaustive of the potential candidates for inclusion, and so the particular words or sentences used in an experiment may actually constitute a significant variable that needs to be considered by the researcher. Clark's suggestion is that the research design treats language as a random effect and not a fixed effect. Modelling language as a random effect controls for the variation that may be attributed to the particular sample of language used in the

experiment, thereby permitting generalisation to the population of possible language that could have been sampled. While Clark was concerned with *the language as fixed effect fallacy*, this thesis draws attention to the potential *concordance as fixed effect fallacy*: the idea that the particular concordance used in an experiment constitutes a significant variable that needs to be addressed by the researcher.

The studies reported in Chapters 3, 4, 5 and 6 of this thesis depart from the hypothesis that concordancing is a complex, multidimensional construct and that different operationalisations of concordancing are a source of variability in studies of learner use of concordances. As we will see below, the research designs adopted in the majority of research on learner use of concordances implicitly conceptualise learner use of concordances as a simple, monolithic treatment condition; from Clark's perspective, as a fixed-factor with a single level. If this is not an appropriate characterisation of learner use of concordances, the generalisability of such studies is severely compromised and so their findings need to be reinterpreted: that is, they do not report on learner use of concordances in a general, unqualified sense, but on particular operationalisations of learner use of concordances. By identifying potential variables within the construct of concordancing, the studies reported here speak to the need for future research on learner use of concordances to engage with the operationalisation of concordancing in experiment design from the perspective of generalisability. The studies reported in this thesis also have interesting implications for teaching, learning and materials design, but the driving question behind the research reported here concerns the identification of factors that need to be manipulated, controlled for or otherwise reported on in research on learner use of concordances. Hence, fundamental to understanding the aims of this project are a clear understanding of the implications of construct description and treatment sampling in experimental design.

The inferences that can be made on the basis of any particular set of experimental results depend on the experiment design employed, and one key factor in experiment design is the operationalisation of treatment conditions. The relationship between the types of inference that can be made and the operationalisation of treatment conditions depends on the adequacy of the construct definition employed. Hence, the first step in experimental design is usually identified as construct definition (Punch, 2003; Scholfield, 1995). In *Statistical Principles in Experimental Design*, Winer et al. (1991, p. 83) provide a very clear discussion of these issues before presenting a detailed description of the two main models of experimental design, designs which

operationalise treatments as fixed constants, and designs which operationalise treatments as variance components:

In general, corresponding to each treatment about which the experimenter seeks to make inferences, there is assumed to be a population of normally distributed criterion measures. The number of such populations is equal to the number of treatments. Assume that the number of treatments is  $K$ . In the experiment, data are obtained on  $k$  of the possible  $K$  treatments. If  $k = K$ , then observed data are available on all treatments in the domain of treatments to which inferences are to be made. If  $k$  is less than  $K$ , observed data are available on only some of the treatments about which inferences are to be drawn. The importance of this distinction is that it determines the nature of the generalisations which can be made by the researcher. In this section the case in which  $k = K$  will be considered. This case is called Model I or the *fixed constants* model and limits the experimenter to making statements only about levels of the treatment actually included in the experiment. Model II, the *variance component* model in which  $k < K$ , will be considered in the next section. (Winer et al., 1991, p. 83)

Of course, the true number of possible treatment conditions ( $K$ ) cannot be determined arbitrarily; the number of possible treatment conditions is identified through thorough construct definition. If research on learner use of concordances operationalises learner use of concordances as a single treatment condition without levels, the experimental design implicitly adheres to Model I:  $k = 1$  and  $K = 1$ . As long as learner use of concordances is a simple, monolithic construct, with apparent differences between the conditions of learner concordance use representing ephemeral variation upon a single prototypical activity, this is not a problem: the results of the experiment can be generalised to learner use of concordances in general. If, however, there are potentially significant differences between variations in learner concordance use, the generalisability of the results is severely compromised because  $K > 1$ , but  $k = 1$ . Very few studies of learner use of concordances have begun to address this issue.

Two clear examples of research in which experimenters have engaged with the issue of variation in the operationalisation of learner use of concordances are Sun & Wang (2003) and Boulton (2009b). Sun and Wang (2003) operationalise concordancing in two distinct treatments: an inductive and a deductive sentence correction task ( $k = 2$ ). Boulton (2009b) operationalises concordancing in a *key word in context* (KWIC) treatment and a *sentence format* treatment ( $k = 2$ ). However, by the same token, it has to be noted that the concordancing task that Boulton's learners engaged with was neither the inductive nor deductive task that Sun & Wang's learner's

engaged with, and Sun & Wang only presented their learner's with one of the citation format conditions employed in Boulton (2009). Boulton's learners used concordances as a reference resource during a gap fill test, and Sun and Wang's learners only used KWIC format concordances. Hence, examining the operationalisation of concordancing in even just these two studies,  $K$  equals at least four: sentence format citations used inductively, sentence format citations used deductively, KWIC format citations used inductively and KWIC format citations used deductively. However, these are not the only two aspects of how concordancing was operationalised in these studies that may be thought to be potentially significant: for instance, the two studies also utilised different source texts to generate the concordances that learners were presented with. For this reason, the results of these studies cannot be discussed in simple terms of learner use of concordances under Model I:  $k$  appears to be significantly smaller than  $K$  in each design.

Looking at published studies of learner use of concordances more widely, it soon becomes clear that there are many potentially significant variables in how learner use of concordances has been operationalised. Perhaps most obviously, there is great variety in how language learners have interacted with concordances. Allan (2010), Boulton (2010b) Daskalovska (2013) and Gordani (2013) use concordance treatments based on the classic DDL model in which learners are presented with a concordance accompanied by a set of questions that aim to guide the learner through a series of research-like tasks: identifying some feature of the language in the concordance, classifying the language identified, and then forming a generalisation on this basis. In contrast, Boulton (2009), Cobb (1997), Chan and Liou (2005) and Kaur and Hegelheimer (2005), present concordances alongside gap-fill tasks; Chan and Liou (2005) report that their treatments also included sentence translation exercises, while Kaur and Hegelheimer (2005) also asked learners to use concordances in what they term sentence building exercises. These operationalisations of learner use of concordances in the experimental literature by no means exhaust the wide range of tasks that have been suggested in the wider literature (see for instance Aston, 2001; Honeyfield, 1989; Mukherjee, 2006; Römer, 2011).

To illustrate one more potential source of difference, there is also considerable diversity in terms of the source texts that have been used to generate concordances in studies of learner use of concordance. For example, some studies use large general corpora, such as the BNC (Boulton, 2010a; Daskalovska, 2013; Gordani, 2013; Kaur & Hegelheimer, 2005; Pérez-Paredes

et al., 2012); other studies have used corpora composed of materials explicitly targeted at language learners (Allan, 2010; Cobb, 1997, 1999); and yet others have drawn upon a parallel English-Chinese newspaper corpus, *Sinorama* (Chan & Liou, 2005; Lai & Chen, 2013). Given that the content of a concordance is determined by the interaction between search term and source text (the search term determines the node of each citation and the source text determines the cotext extracted along with it), it stands to reason that concordances extracted from different source texts are likely to be an important variable in learner use of concordances.

Exploring published research in this way, it seems clear that learner use of concordances has been operationalised in many different ways. It implies that  $K$  for learner use of concordances is actually  $>1$ , and so experiments that operationalise learner use of concordances as  $k = 1$  cannot be confidently generalised to learner use of concordances in general. Consequently, it is worth considering the extent to which experiment design Model II offers a solution to this issue.

Experiment design Model II is the variance component model. It is appropriate when it is not possible to gather observations on all possible treatments within an experiment. Winer et al. (1991) provide a concise exposition of the difference between the assumptions of Model I (the fixed constants model) and Model II (the variance components model) and their implications:

One of the basic assumptions underlying Model I is that all treatments about which inferences are to be made are included in the experiment. Thus, if the experiment were to be replicated, the same set of treatments would be included for each of the replications. Model I is usually the most appropriate for a single-factor experiment. If, however, the  $k$  treatments that are included in a given experiment constitute a random sample from a collection of  $K$  treatments, where  $k$  is small relative to  $K$ , then upon replication a different random sample of  $k$  treatments will be included in the experiment. It is quite rare for experiments to be designed and executed with the treatments selected randomly from a population of treatments in the social, behavioural and biological sciences. Accordingly, Model I is usually the ... model [employed]. One result is that the outcome of those studies cannot be generalized to a population of treatment levels. In principle, experiments can be carried out under the conditions specified by Model II and results can be generalized to the population of treatment levels when that occurs. Model II covers this case. (Winer et al., 1991, p. 92)

In Model II, instead of including all possible treatments within an experiment, the experimenter takes a random sample of all possible treatments, and random sampling of treatments provides

the logical justification for generalising the experimental results obtained to treatment conditions that were not operationalised in the study. Consequently, Model II does provide a potential solution to the generalisability of research on learner use of concordances, but in order for it to be used, we would need a random sample of  $K$ , possible learner use of concordance treatments. Clearly then, we would first need an inventory of variables within learner use of concordances from which to sample as such an inventory would be a prerequisite of specifying the sampling frame for a random sample. Given such an inventory, it would then be possible in principle to randomly sample treatments, allowing the use of Model II ( $k < K$ ), but as with Model I, there appears to be considerable work to be done in construct definition before such an approach would be possible.

Whichever experimental model is adopted, if concordancing is a multidimensional construct, and analysis of the data reported in Chapter 3 suggests that it is, then construct definition and operationalisation of treatments are central to making generalisations from experimental results. This thesis aims to push this research agenda forward. First, if the construct of learner use of concordances implies that  $K$  is actually quite large for learner use of concordances in general, then research that has operationalised a single treatment condition in a study that conforms to Model I must redefine  $K$  in that study. Without random sampling of treatment conditions, Model I must be used, and so inferences can only be made on the basis of  $k = K$ . Much existing research that purports to examine learner use of concordances in general may need to be reinterpreted as research on a particular type of learner use of concordances because the results of studies under Model I are only transferable to operationalisations of learner use of concordances that closely parallel those reported in that study. Thus, second, construct definition and identification of variables has an important role to play in improving reporting standards in research on learner use of concordances, something which has been criticised by Boulton (2010b). When Model I has been used, deciding whether the operationalisation of learner concordance use reported in a study is comparable to any other relies on the reporting of salient variables within operationalisation; the more detailed the account of operationalisation provided, the more confidence a reader can have in the transferability of the results. Construct definition and identification of variables has an important role to play in specifying what those variables are. Third, as mentioned above, to realise the ambitious agenda of conducting generalisable research under Model II, researchers would need a sufficiently thorough construct definition for

concordancing treatments to be randomly sampled from the full range of concordancing treatments possible. Such a project may perhaps prove unrealistic in terms of learner use of concordances in general, but it may have heuristic value in facilitating the identification of more narrowly defined domains of learner use of concordances. In brief, identification of variables within the construct of learner use of concordances has an important role to play in moving the field forwards (cf. Cobb & Boulton, 2015), and this thesis aims to fill this gap in the research.

## **Research Questions**

The agenda for the research reported here has been set out above. Here, I set out the specific research questions addressed by each of the four studies reported.

### **Chapter 3, Study 1:**

#### **Is concordancing a simple or multidimensional construct?**

1. *Does concordancing appear to be a unidimensional or multidimensional construct?*
2. *What are the factors that distinguish different realisations of concordancing?*

### **Chapter 4, Study 2:**

#### **Can concordancing preferences be predicted from user characteristics?**

3. *To what extent can preferences for different realisations of concordancing be predicted by concordance user characteristics?*

### **Chapter 5, Study 3:**

#### **Do concordances present systematically different authentication burdens?**

4. *Do concordances extracted from three different authentic corpora have statistically significant differences on measures of word frequency?*
5. *What is the magnitude of differences in word frequency measures between concordances extracted from different authentic corpora?*
6. *What is the magnitude of differences in word frequency measures between concordances extracted from different authentic corpora as compared with word frequency measures of concordances extracted from a corpus of graded readers?*

### **Chapter 6, Study 4:**

#### **Do concordances contain systematically different ratios of lexical variation and repetition?**

7. *Do concordances extracted from corpora with differing levels of generality have significantly different standardised type token ratios?*
8. *What is the magnitude of difference between the type token ratio scores of concordances extracted from corpora of differing levels of generality?*

### **Organisation of the thesis**

An important research goal for the field is the identification of variables within the construct of learner use of concordances, and this thesis makes a contribution to this project through a series of inter-related studies. In Chapter 2, the thesis reviews the literature on concordances so as to develop an appropriate framework for the inclusion or exclusion of a given realisation of a technology as a concordance or not. It then reviews the literature on learner use of concordances from an affordances perspective. From such a perspective, individual features of a technology may be seen as either facilitating or impeding a particular pedagogical goal (Levy & Stockwell, 2006). Particular features of concordances are mapped onto the various purported affordances of concordances, and this analysis suggests that concordancing is not a simple, monolithic construct but a complex, multidimensional construct and, furthermore, that not all of the affordances of a concordance can be available to a user simultaneously. The analysis implies that the affordances of concordances may be expected to differ between distinguishable paradigms of learner use of concordances, and that variables relevant to these affordances should be identified and explored. The studies reported in the next four chapters engage with this issue. However, because each study employs a distinct research design, there is no separate methodology chapter in this thesis; instead, each chapter reporting a study contains a method section detailing the method employed in the study reported in that chapter.

Chapter 3 seeks to complement the theoretical discussion of the construct of learner concordance use in Chapter 2 with empirical evidence. It explores the construct of concordancing through a survey of concordance users. The survey gathers data on concordance users' preferences in regards to five of the potential variables in learner concordance use that



were identified in Chapter 2: the formatting of citations in the concordance, the alignment of citations in a concordance, the ordering of citations in a concordance, the relative generality of the texts composing the corpus that the concordance is extracted from, and the users' manner of engagement with the concordance. Using a series of factor analyses, the underlying structure of the data is identified. Concordance users' preferences are identified with four distinguishable dimensions of variance, indicating that concordance use is not a monolithic construct and should not be treated as such.

Chapter 4 identifies correlations between the concordance users' scores on the four dimensions of variance identified in Chapter 3 and four ways of categorising users: how long they have been using concordances for, how frequently they use concordances, the field of language use that they use concordances in, and their linguistic relationship to the concordances they use. The analysis finds predictors for citation formatting preferences, citation ordering preferences and concordance engagement preferences. The analyses indicate that preferences for particular configurations of concordance use correlate with particular user groups, and the relevance of the findings to the language learning affordances of concordances is discussed. In particular, two significant modes of concordance use are identified and these are then associated with, on one hand, affordances of concordances that utilise traditional reading skills, and on the other, with affordances focused on the identification of lexical patterns. These results provide strong evidence that concordance user's concordancing preferences are not arbitrary but are likely to reflect the different affordances of different operationalisations of concordancing. In particular, they highlight the importance of task and format as variables in learner concordance use. However, no predictors are identified with preferences for concordances extracted from more or less general corpora, although this was the second largest dimension of variance identified in Chapter 3.

Chapter 5 engages with a key variable in learner concordance use that has implications for the readability of concordances: word frequency measures of authentication burden. This is important because Chapter 4 indicates that the readability of a concordance may be an important determinant of a concordance's affordances for some user groups, and word frequency is a useful indicator of a text's readability (Grabe & Stoller, 2001; Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2013; Schmitt, Jiang, & Grabe, 2011). This chapter reports a study in which the averages of citation level word frequency measures are compared in four

different conditions: concordances extracted from a corpus of academic prose, a corpus of news, a corpus of fiction, and a corpus of graded reading. It finds significant differences between conditions and estimates the effect sizes of the differences identified. It shows that concordances extracted from these different corpora have systematic differences on measures of word frequency, highlighting the importance of this variable in learner concordance use. Importantly, it shows that the differences between concordances extracted from authentic corpora are comparable in magnitude to the difference between those extracted from the graded corpus and those from lowest scoring authentic corpus. Interestingly, the data can also be interpreted as indicating that the fragmentary nature of concordances presents a lower vocabulary load than that reported elsewhere for whole texts. In sum, this chapter indicates that concordances may be able to provide traditional-reading-based affordances to learners at unexpectedly low proficiency levels, demonstrates the importance of corpus composition as a variable in learner use of concordances, and presents a strong case for reporting the vocabulary load of concordances that are used in studies of learner concordance use.

Chapter 6 examines the vocabulary presented in concordances from a different perspective. In contrast to Chapter 5 which employed measures of word frequency, Chapter 6 uses a measure of type-token ratio (TTR) (Johnson, 1939, 1944 cited in Jarvis, 2013, p. 16). Measures of TTR have theoretical implications for the affordances of concordances in two respects. First, TTR has particular relevance to the affordances of concordances in language for specific purposes (LSP) settings from an authentication burden perspective. In LSP contexts, TTR may be a better measure of a text's authentication burden than word frequency. Perhaps more importantly though, it is important to realise that TTR scores provide an indication of the extent to which a concordance presents lexical patterning. Low TTR scores underlie the presence of repeated collocates, word clusters and ngrams. For this reason, differences in TTR scores indicate an important variable if concordances are envisioned as providing an opportunity to observe lexical patterning (see Chapter 2 and Chapter 4): TTR scores provide an indication of the extent of a concordance's pattern observation affordances. The study reported in Chapter 6 analyses concordances extracted from corpora of three different levels of generality (cf. Chapter 2 and Chapter 3): that is, corpora of written English, written academic English, and written English from one particular academic journal. It reports significant differences and large effect sizes, and discusses the findings in relation to the affordances of concordances for language

learning. The findings indicate that it is possible to find systematic differences between the TTR scores of concordances and that the level of generality of the corpus used is a factor that can predict these differences. Consequently, both TTR score and level of corpus generality should be considered potentially significant variables in learner use of concordances.

Chapter 7 draws together the findings from Chapters 3, 4, 5 and 6, and triangulates them so as to present a richer description of the construct of concordancing in language learning. Synthesising the findings, it discusses the implications of the findings for both learner use, materials design and research on learner use of concordances. The chapter concludes with a discussion of the limitations of the research reported here and an agenda for future research in this area.



## Chapter 2. Literature review

### Introduction

This chapter is organised in two sections, the first presenting important background and contextual information, the second engaging with theoretical perspectives on learner use of concordances. In Chapter 1, it was noted that learner use of concordances is often equated with learner use of language corpora or with data-driven learning, and a distinction between these terms was made; perhaps unsurprisingly then, the term *concordance* is itself somewhat conceptually ambiguous. Consequently, the first goal of this chapter is to review the history and development of concordances so as to arrive at a definition of concordance that can be used to discriminate between the use of concordances and the use of other corpus tools, input texts or language learning resources. In the second part of this chapter, the technical definition of concordances developed in the first part of this chapter is used to engage with theoretical perspectives on the affordances of concordances. The literature on affordances is synthesised so as to provide a coherent framework for research on learner use of concordances and highlight the range of theoretical perspectives on learner use of concordances that this thesis engages with. However, a consequence of this framing of the literature review is that it is primarily concerned with theoretical perspectives, and not empirical studies.

Bitchner (2012) suggests that there are three types of literature appropriate to an applied linguistics thesis reporting empirical research: non-research literature, theoretical perspectives and research literature. Wallace and Wray (2011) make a comparable distinction between theoretical literature and research literature, also referring to the ideas of practice literature and policy literature. Such distinctions are however slightly problematic from the perspective of this literature review. It often appears that research literature is held in higher esteem than other types of literature and, as such, it may be thought that a literature review should engage most thoroughly with the available research literature. The problem is that there is not really any research literature that addresses comparable research questions to the research questions addressed in Chapters 3, 4, 5 and 6 (though see each Chapter for what there is). The experimental and quasi-experimental studies reported in Chapters 3, 4, 5 and 6 actually address gaps that

emerge from synthesis of the theoretical perspectives in the literature, so it is the theoretical perspectives that this thesis engages with that are discussed in this chapter.

As useful as the literature classification schemes mentioned above may be, they represent a simplification: the discussion and conclusion sections of the research literature present theoretical perspectives on research knowledge. This thesis addresses gaps in the research that emerge from such theoretical perspectives, and not research findings, whichever type of literature such perspectives emerge from. If we separate the type of literature from the type of knowledge, we find that there is no research knowledge pertaining to the construct of concordancing, who is most likely to be using what type of concordancer, the extent of the authentication burden posed by concordances, the extent to which the authentication burden of concordances can be manipulated, the extent of lexical variation and repetition in a concordance, nor the extent to which the extent of lexical variation and repetition in a concordance can be manipulated. But this does not mean that these facets of learner concordance use are not significant for research on learner use of concordances; their extensive discussion in the theoretical literature, practice literature and discussion sections of the research literature testifies as much. The gaps in the research literature that this thesis addresses are only gaps in the research literature in the sense that there has been no empirical research that directly addresses these issues, although they have frequently been discussed from theoretical perspectives.

The absence of a coherent framework for research on learner use of concordances is in itself a gap in the theoretical literature, and so drawing together the vast literature on learner use of concordances has value in its own right. However, more importantly, the literature review presented in this chapter develops two major themes. The first theme is the idea of models of use. As we will see, the literature on learner use of concordances abounds in recommendations and metaphors for how learners can use concordances. It is argued in this chapter that the various perspectives in the literature reflect three core underlying affordances of concordances: concordances as a research tool, concordances as a source of concentrated input and concordances as an index to ad hoc language learning resources. In combination with these theoretical perspectives on the affordances of concordances, the findings of the studies reported in this thesis are used to identify two empirically informed models of learner use in Chapter 7, the discussion chapter of this thesis. The second theme developed in this chapter emerges from the typology of affordances mentioned above: the authentication burden of concordances. The

authentication burden of concordances and how it can be addressed are perennial concerns in discussion of learner use of concordances: arguably, it is the driving force behind the majority of recommendations made and metaphors suggested. It is somewhat surprising then that there have been very few attempts to approach this issue quantitatively (through see the literature review in Chapter 5).

The review of literature presented here underlines the need for research on the construct of concordancing (Chapter 3), research on the relationship between the construct of concordancing and different types of uses (Chapter 4) and research on the authentication burden of concordances (Chapters 5 and 6). For each of the four studies reported in this thesis, literature that is only relevant to that study is presented in the relevant chapter (Chapters 3, 4, 5 and 6): that is, literature pertaining to methodological decisions made in that study or presenting findings that might have some indirect bearing on the research questions addressed by that study.

### **What are concordances and how have they developed?**

To develop a definition of concordances that can encompass the range of ways in which concordances have been used in pedagogical contexts, it is necessary to trace the development of concordances in the broader context of language work. This process reveals the significant changes that have taken place in regards to both what is referred to by the word *concordance* (a concordance's *ontological status*) as well as the epistemological claims that can be made about language on the basis of a concordance (a concordance's *epistemological status*). In combination, consideration of the changing ontological and epistemological statuses of concordances not only provides a basis for a cogent definition of concordance that allows concordances to be distinguished from other texts, but also points to dimensions of variability within the affordances of concordances that particular operationalisations of concordancing provide.

Sinclair (1991) provides an excellent introduction to concordancing and states that, in its most basic form, a concordance is simply an index, a way of locating a word in its textual environment. Indeed, the first concordances functioned just as a standard index does: a list of words with page numbers and line numbers citing where the words occurred (even now, each instance located can be termed a *citation* for this reason, though they are sometimes also termed

*lines* or *hits*). For example, a concordance of this document would reveal that the orthographic form *Chinese* appears on the following pages of this document: pages 26, 36, 84, 95 and 142. Historically, concordances provided citations for every occurrence of a word within a closed set of texts, such as the bible, or the works of Shakespeare, and thereby facilitated the making of authoritative statements as to what these particular texts had to say about a particular topic (G. Kennedy, 1998; M. McCarthy & O'Keeffe, 2010; Sinclair, 1991; Tribble, 2010). For instance, we could concordance the words *blessed are the* in the King James Bible, and thereby identify the eight times it occurs in the New International Version or the nine times it occurs in the Kings James Version (<https://www.biblegateway.com/keyword/>). There are at least two particularly important features to note about how concordances have developed since then: first, as concordances are no longer made by hand, but are instead processed by computers, the ontological status of concordances has become blurred; second, concordances are now also made of open sets of text, with corresponding implications for their epistemological status.

### **Changes in the ontological status of concordances**

Sinclair (1991, p. 34) states that a concordance is an index, and then suggests that “It is valuable to think of a concordance as a text in itself”. He makes this statement because of the effect of information technology on concordancing. Pre-computer, the labour involved in locating and recording the location of each occurrence of a word in a large quantity of text suggested the index as an appropriate format of presenting a concordance: copying out each occurrence of the word found along with a number of the words found around it (its *cotext*) would have been expensive in terms of time and effort. However, with modern computers, the reproduction of text has become almost effortless and instantaneous, and the qualities of the text displayed to the user can be changed easily, depending on exactly which functions are provided in the concordancing software’s interface: for example, the amount of cotext presented in each citation can be varied, and the citations can be arranged and rearranged in various ways (G. Kennedy, 1998; Stevens, 1991a). Consequently, a concordance is now very unlikely to be presented to a user as an index: the index is generated in a computer-readable format, but the modern user is presented with a text formed of one or more of the citations indexed. In Sinclair’s words, “It saves the researcher looking up each occurrence”. However, because a decent home computer can now process texts



that are millions of words long in seconds, the size of present day concordances has brought the concepts of *concordance as index* and *concordance as text* into conflict.

The conflict between the concept of concordance as index and concordance as text is a subtle one. On the surface, there would not appear to be any problem at all: all the occurrences of a word are first indexed, then extracted (along with a specified quantity of cotext), and then compiled to present a text to the user. The problem arises, however, when the size of the index produces a text of unwieldy proportions; as Sinclair (1991, p. 5) humorously remarks, “in the days of paper print-out the phrase ‘knee-deep’ was not always figurative!” It is no great issue for the corpus linguist sitting in front of a computer, able to scroll through citations screen-by-screen, because the electronic concordancer provides access to the index in toto, even if not all of the citations can be displayed simultaneously. Concordancing software provides the user a dynamic text. However, it is perhaps something of an issue within learner use of concordances where teachers and researchers have frequently warned that learners may become overwhelmed if they are presented with too much data (Allan, 2010; Levy, 1990; Sripicharn, 2010; Tribble, 1997; Varley, 2009), and have explored presenting learners with paper-based or otherwise static texts in what is sometimes referred to as *hands-off* concordance use (Barlow & Burdine, 2006; Boulton, 2009b, 2010a; Burdine & Barlow, 2008; Chambers, 2010; Johns, 1991a, 1991b, 2002; Stevens, 1991a; Thurston & Candlin, 1997, 1998; Tribble & Jones, 1997). Interest in learning from a static concordance text is hardly surprising when we consider that in applied linguistics settings, when we talk of concordances, we almost invariably mean a text; rarely is appeal made to the slightly abstract concept of an exhaustive index of a body of texts. However, because a concordance as index in its entirety will often create a concordance as text that exceeds the space available on a single screen or page, in many cases, these two senses of concordance will come into conflict. If a user only has a proportion of a concordance as index available as a concordance as text, they no longer have an exhaustive set of citations in relation to the source text; hence, some authors refer to a text produced by extracting only a proportion of the citations available in the index as *concordance extracts* (Renouf, 2003; Romer, 2008; Tribble, 2010), though not all authors make such a distinction or use the term consistently. In brief, there has been a change in the ontological status of concordances. This change can perhaps be most usefully conceptualised in terms of concordances having a dual ontological status (concordance as index and concordance as text), and this change in the ontological status of concordances has implications

for the epistemological status of the data presented to a user, and so also for a concordance's language learning affordances.

### **Changes in the epistemological status of concordances**

The notion of a concordance providing an exhaustive index of a body of text may seem irrelevant to language study where, typically, an open set of texts is examined. However, the notion of an exhaustive index does have relevance to open sets of texts within corpus linguistics. Indeed, the distinction between an exhaustive and a selective concordance is the first distinction between concordances made by Sinclair (1991, p. 43), who sees exhaustiveness as the principle affordance of concordances. In the context of corpus linguistics, he remarks:

- a. selection is only made when the number of instances becomes quite unmanageable otherwise;
- b. the criteria for selection must be very carefully chosen

This distinction and these remarks are made because corpus linguistic method has been developed to allow open sets of texts to furnish generalisable descriptions of language on the basis of the representativeness of a corpus and the exhaustiveness of its analysis (Leech, 1992).

In corpus linguistics, the term *corpus* is often used to refer to a collection of texts with particular properties, not just any collection of texts used as a text-database (Adolphs, 2006; Biber et al., 1998; Biber & Reppen, 2015; Cheng, 2012; Davies, 2015; G. Kennedy, 1998; Reppen, 2010a; Sinclair, 1991; Teubert & Čermáková, 2007). Specifically, in this technical sense, a collection of texts is a corpus when it has been designed to provide a representative sample of a language population. Of course, the adequacy of any particular sampling scheme is always open to critique, but the principle remains: a collection of texts should be viewed as a corpus in this technical sense only if its claim to be representative of a specified language variety is warranted. Indeed, to the extent that this claim is warranted, principled observations made from a corpus provide a basis for generalisation from an open set of texts to a corresponding language variety. Hence, the exhaustiveness of a concordance as index has value in corpus linguistics: if the corpus is representative, an exhaustive analysis of it represents an analysis of a random sample of the target population. The exhaustiveness of analysis provided by a

concordance as index combined with the construct of a corpus as a representative sample of a variety of language provides corpus linguistics with a methodology that can be construed as a scientific approach to the description of language (Leech, 1992).

As discussed above, in the context of corpus linguistics, Sinclair suggests that an exhaustive concordance can be reduced to a more manageable size via principled selection of citations. The principle he has in mind is random sampling, if the intention is to retain a claim to generalisability (Sinclair, 1991, 2003). As an open set of texts, the size of a corpus is essentially arbitrary: the determining factor is the sufficiency of the data available. As Sinclair explains, the researcher needs to be satisfied that all recurrent patterns are adequately represented in their data set; consequently, for some of the most common features of language use, even a small corpus of several million words will suffice, but for the least common, mega-corpora may well be needed to provide adequate coverage of the feature of interest. Because the size of an open set of texts is inherently arbitrary, a random sample of a large, representative corpus may have the same epistemological status as an exhaustive concordance of a smaller, representative corpus. The generalisability of either is warranted by the addition of more data failing to problematize the description based on the existing sample. Hence, in *Reading concordances: An introduction*, Sinclair (2003) recommends iteratively reading random concordance samples until no new patterns are observed in the data.

To reiterate, these claims to the generalisability of observations made from concordances depend on certain conditions being met: first, when an open set of texts is being used, that the claim to their being a representative corpus is warranted; second, that the concordance obtained is either being processed exhaustively, or that randomly selected concordance extracts are being used in a comparable manner. Note, if either the corpus concordanced is not representative or the citations selected have not been randomly sampled, the epistemological status of the concordance is problematized and the generalisability of any observations made are threatened.

These conditions of generalisability are an important issue in relation to learner use of concordances for two reasons: first, because of debate over whether or not learners should be using concordances *as researchers do* (see for instance Pérez-Paredes, 2010); second, because theoretical perspectives on learner use of concordances commonly include suggestions that contradict these conditions — suggestions that teachers should select which citations to retain in a concordance extract on a pedagogical basis (Aston, 2001; Boulton, 2008, 2009b; Braun, 2005,

2006, 2007; Flowerdew, 2012b; Frankenberg-Garcia, 2014a; Gavioli & Aston, 2001; Johns, 1991b; Kaur & Hegelheimer, 2005; Levy, 1990; Partington, 2001; Stevens, 1991b) or else suggestions that teachers should select source texts in relation to a pedagogical criterion, and not a criterion of representativeness (Aston, 2001; Braun, 2005, 2006, 2007; Cobb, 1999; Flowerdew, 2012a; Gavioli & Aston, 2001; Levy, 1990; Pérez-Paredes, 2010; Römer, 2011; Tribble, 2013). Clearly, once either of these decisions has been made, the epistemological status of the concordance extract has changed: it can no longer be claimed to afford generalisable observations.

As we will see later in this chapter, the traditional affordances of concordances have often been tied to the epistemological status of concordances derived from the notion of a concordance as an exhaustive index of representative sample of a language variety; however, in the context of language learning, concordances may have other affordances independent of such considerations. These other affordances draw upon the notions of concordance as text and as index in ways which are distinct from their epistemological status as sources of generalisable observation: one is concordance as a source of concentrated text, the other is concordances as an ad hoc index to language learning resources. Furthermore, once we recognise that concordances have affordances that are independent of the conditions of generalisable observation, it becomes clear that these independent affordances of concordances have different conditions that their affordances are dependent upon: it is the different conditions of these affordances that underlies the multidimensionality of the construct of concordancing examined in Chapter 3, the correlations between different types of concordance user and different operationalisations of concordancing observed in Chapter 4 and the rationale for manipulating the authentication burden of concordances discussed in Chapters 5 and 6. However, it also problematizes the construct of a *concordance* to some extent because each of these three different sets of affordances emphasises different aspects of the construct of concordancing: exactly what is, or is not, a concordance. In the next subsection, the relationship between the two ontological statuses of concordances is used to create a cogent definition of concordance.

## A definition of concordance

A useful definition of concordance for research on learner use of concordances should be able to encompass a wide range of interest in learner use of concordances as discussed in the literature. The discussion above has drawn attention to two different senses of concordance derived from the history of concordances and corpus linguistics: *concordance as index* and *concordance as text*. These two different senses of concordance underlie differences in the generalizability of concordances, what we might call the corpus linguistic research affordances of concordances, but they also underlie the typology of pedagogical affordances presented later in this chapter, distinctions that underpin the research reported in Chapters 3, 4, 5 and 6. At the same time, the notions of concordance as text and concordance as index intersect with other distinctions commonly appealed to in the literature on learner use of concordances, such as *hands-on* and *hands-off* concordancing (Boulton, 2010a), *hard* and *soft* approaches to concordancing (Gabreilatos, 2005), *convergent* and *divergent* use of concordances (Tono et al., 2014), and the *possibilities* and *feasibilities* scenarios (Alcaraz Calero & Pérez-Paredes, 2008; Pérez-Paredes, 2010). The relationship between these distinctions and the theoretical framework advanced in this chapter are discussed below.

Hands-on concordancing refers to learner use of concordancing software. If a learner has access to a representative corpus through a concordancer, the learner then also has access to a concordance that affords robust generalisations. Of course, the most important aspect of hands-on concordancing is not access to the exhaustive index generated, but the opportunity for the learner to specify their own search term, and there has been much fascinating research that has shed light on learner search patterns (Franken, 2014; Lai & Chen, 2013; Park, 2012; Pérez-Paredes, Sánchez-Tornel, Alcaraz Calero, & Jiménez, 2011; Pérez-Paredes et al., 2012). However, this thesis is concerned with the concordance itself (as text or index), and not its interface, and so the interesting topic of learner search patterns is not pursued further here. Instead, it will suffice to note that, in hands-off concordance use where a static concordance extract is presented, a learner is very unlikely to be presented with a concordance that supports a robust claim to generalisability. However, hands-on use of either a non-representative corpus or non-random selection of a representative corpus would not constitute use of a robustly generalisable concordance and, in theory at least, a robustly generalisable concordance could be

presented on paper. Hence, although related, the generalisability affordances of a concordance are distinct from whether the learner uses a concordancer hands-on or a concordance hands-off.

Gabreilatos (2005) makes a commonly cited distinction between *hard* and *soft* approaches to learner use of concordances. The main thrust of his distinction parallels the notion of hands-on and hands-off use, hands-on use of concordancing software being harder than hands-off use of pre-prepared texts, but it also draws upon the distinction between the related concepts of *convergent* and *divergent* learner use of concordances (Tono et al., 2014). In convergent use, a concordance has been generated, or a concordancing task constructed, in such a way that the observations made in regards to the concordance are predictable by the teacher. In divergent use, the language learning outcomes of learner use of concordances is not pre-specified. Raw concordance data frequently results in unexpected observations for both teachers (Johns, 2002) and researchers (Sinclair, 1991). Consequently, convergence effects are typically fostered by pedagogically motivated selection of either which citations to present in an extract or which texts are to be concordanced: what we can term *pedagogical mediation*. Gabrielatos' notions of hard and soft use appear to acknowledge the synergy between hands-off use and convergent learning outcomes and hands-on use and divergent learning outcomes, but this is again different to the generalisability affordances of concordances and their potential to offer other, distinct, pedagogical affordances.

Taking one step further back, in presenting the *possibilities* and *feasibilities* scenarios, Alcaraz Calero and Pérez-Paredes (2008) do not only see the software component of hands-on use or the learning outcome component of divergent use as constituting the hard end of learner concordance use, rather they see the wholesale transfer of the corpus linguistic paradigm as the issue. They term transfer of corpus linguistic methods to the classroom *the possibilities scenario* and suggest in its place *the feasibilities scenario*. In the feasibilities scenario, it is not the method that is appropriated, but the tools. The important point here is that they are not identifying the challenges of learner use of concordances with hands-on use necessarily; rather, they are calling for a pedagogical mediation of the paradigm of concordance use. They write:

In this new scenario language corpora are specifically compiled, annotated and exploited with a pedagogic intention. The main characteristics of this new scenario are defined against the present role of principled corpora in language education (Pérez-Paredes, 2010, p. 62)

Clearly, in the terms presented above, this suggestion is not really about the use of corpora at all; it is about the pedagogical use of corpus tools with other types of text (cf. Ballance, 2012). As we will see in the next section, this is not a novel suggestion. In fact, there would seem to be a growing consensus that such pedagogical mediation is necessary because learner concordance use that parallels researcher use too closely is unfeasible due to both the quantity of data the learner is presented with and the cognitive complexity of rigorous analysis (Ädel, 2010; Boulton, 2010a; Braun, 2007; Coxhead, 2014; Flowerdew, 2012a, 2012b; Franken, 2014; Gavioli, 2001; Hunston, 2002; C. Kennedy & Miceli, 2001; Levy, 1990; O'Sullivan & Chambers, 2006; Partington, 2001; Pérez-Paredes, 2010; Pérez-Paredes et al., 2012; Stevens, 1991a; Sun, 2003; Varley, 2009; H. Yoon & Hirvela, 2004).

In light of the various distinctions that have been made between different types of learner concordance use, a definition of concordances in pedagogical contexts should be able to encompass all of these types of concordance use but still be able to distinguish a concordance from a type of text that would not be considered a concordance by anyone working in the field. That is, it should also be able to distinguish hands-on use of exhaustive concordances and hands-off use of pedagogically mediated concordance extracts from lists of example sentences written via introspection, corpus-informed language learning materials, and other types of text such as novels, essays and post cards. It would seem that the key element in such a definition of concordance is the necessity of a text having been accessed or generated via an exhaustive index of a specified form in a body of text. Such an index is itself a concordance, a concordance as index, but for a text to be considered a concordance, it is necessary that there be an extant connection between the processes of indexing, extracting and compiling the occurrences of a form. This connection to an index can be considered both a necessary and sufficient condition for considering a text to be a concordance. If so desired, this basic definition could then perhaps be complemented by definition of a *robustly-generalisable concordance*, a special type of concordance with a special status in corpus linguistics, as discussed above.

In the next section, concordances as defined here are charted against the theoretical perspectives on the language learning affordances of concordances that have been discussed in the literature. As we identify the pedagogical affordances of concordances and elucidate their various conditions, it becomes clear that concordancing is likely to be a complex multifaceted

construct, that different affordances of concordances are likely to be facilitated by different operationalisations of concordancing, and that there is an urgent need for research on how the authentication burden of concordances can be manipulated at the point of concordance generation.

### **The language learning affordances of concordances**

The 1990s saw a surge of interest in using concordance extracts in classrooms (Johns, 1991a, 1991b; Stevens, 1991a, 1991b; Thurston & Candlin, 1997, 1998; Tribble & Jones, 1990). As access to computing facilities and language corpora was fairly limited at this time, researchers often had little choice but to present concordance extracts on paper, and they typically presented citations from texts that would not meet a strict definition of a corpus. However, limited resources was not a major barrier to their research as they believed that even static concordance extracts from ad hoc collections of text afforded learners important language learning opportunities. Indeed, although Johns is typically cited as promoting the idea of learner as researcher, his publications actually provide detailed and insightful accounts of how teachers can profitably mediate between concordance and learner and, in that sense, show that Johns was well aware of the difficulty a learner would have in using a concordance in a manner fully commensurate with that of a researcher (Johns, 1986, 1991b): Johns' concept of the learner as a researcher is in essence a metaphor. The pioneers of learner use of concordances did not equate the language learning affordances of concordances with their research affordances.

The most commonly cited studies of the effectiveness of learner concordance use are innovative studies by Tom Cobb that integrated concordance output within a computer-assisted vocabulary learning environment (Cobb, 1997, 1999). But this was not concordancing in the corpus linguistic sense either as the language database concordanced was not a true corpus; rather it was the integration of the affordances of a pedagogically appropriate language database within a computer-assisted language learning (CALL) environment via concordancing. Again, Cobb did not equate the affordances of concordances for language learners with their affordances for researchers.



Since then, the rate of publications in this field has rapidly increased (Boulton, 2010b), and so too has access to dynamic, exhaustive concordances of even very large corpora, resources which support robust generalisation and are increasingly becoming available online. However, some of the pedagogical motivations for presenting learners with concordances appear to have remained the same as in the early days of research on learner use of concordances and, in fact, exist alongside reservations about strongly researcher-like concordancing of large authentic corpora.

Considering the changes in the ontological status of concordances, as well as changes in their epistemological status, there is value in a review of the literature on learner use of concordances that considers the affordances of concordances in relation to both their status as indexes and their status as texts, and also the related concepts of robust-generalisability and pedagogical mediation. Such a review points to the fact that there are a wide range of possible affordances of concordances and the fact that these affordances are dependent on very different concordancing conditions being met. Chapter 3 of this thesis reports a study that finds evidence for such different operationalisations of concordancing. Chapter 4 then connects different operationalisations of concordancing to different user groups, taking a first step towards empirical evidence that different operationalisations of concordancing can be associated with different affordances. It will also be seen in this chapter how seeking different language learning affordances from concordances permits pedagogical mediation of concordances that researcher like exploitation would not allow; however, at the same time, it is shown that the drive to mediate between concordance and learner interacts with the affordances of concordances in complex ways and that some of the most distinctive and valuable affordances of concordances for language learners are strongly connected with the use of relatively unmediated concordances. For these reasons, it is vital that the authentication burden of concordances is quantified and approached as a significant variable in learner use of concordances. Chapters 5 and 6 begin to explore concordances from this perspective.

The next section of this chapter is organised around a typology of three main affordances of concordances derived from the definition of concordances above: concordances as research tools, concordances as sources of concentrated input texts, and concordances as ad hoc indexes to language learning resources. Theoretical perspectives on the affordances of concordances

discussed in the literature are organised within sub-categories within these three main affordances.

### **Concordances as a research tool**

It has been suggested that, from the historical perspective, exhaustive concordances of corpora have been important in much language work as they provide the basis for making generalisations about languages and varieties of languages, and that interest in the use of concordances in linguistic work has been a significant stimulus of interest in learner use of concordances (Bernardini, 2004; Braun, 2006; Mukherjee, 2006; Pérez-Paredes, 2010; Römer, 2011; Seidlhofer, 2003). Indeed, exhaustive concordances of corpora would appear to afford language learners the same affordances that they offer to corpus linguists: an opportunity to observe a set of generalisable language datum en masse. This can perhaps be thought of as the hard interpretation of Tim Johns's metaphor of learner as researcher, although it does not seem that Johns himself endorsed the idea of presenting learners with either exhaustive concordances or concordances of formal corpora (both conditions being requirements of robustly generalisable concordances, as discussed above). For instance, (Johns, 1991b) explicitly discusses preparing concordance extracts in which citations have been selected on pedagogical grounds; he also discusses presenting learners with concordances that have been extracted from a collection of popular science magazines, selected on the basis of its availability and relevance to his learners' interests and specialisations (Johns, 1991a), not its representativeness as a corpus. Indeed, while a good number of studies of learner use of concordances have made use of concordances of large, representative corpora (Boulton, 2010a; Daskalovska, 2013; Gordani, 2013; Kaur & Hegelheimer, 2005; Pérez-Paredes et al., 2012), none appear to have asked learners to examine concordances exhaustively or iteratively. In other words, although it is theoretically possible for concordances to afford learners robustly generalisable data sets, it is not entirely clear that this somewhat hard-line approach is actually being advocated, despite how commonly it is discussed.

Sinclair's excellent *Reading concordances: an introduction* does teach a principled approach to using a concordance to analyse a representative corpus, but it is worth noting that this book is specifically targeted at "students, teachers and researchers in the language industries" and emerged from the need to train masters level students in how to interrogate a large corpus, manage and interpret their results and then present their findings along with a

convincing body of evidence (Sinclair, 2003, p. ix). The book consists of 18 chapter-length units, and inducts readers in the process of working with smallish random samples through iterative stages of generalisation and degeneralisation. As the book is in print format, it presents randomly sampled concordance extracts (random selection being a cornerstone of scientific sampling), and contains corresponding disclaimers regarding the conclusiveness of the findings of the exercises in the book. Thus, although the book aims to provide readers with a method for reading concordances and not robust answers to the linguistic questions in the tasks set, it strongly emphasises the necessity of engaging with an exhaustive concordance in a principled manner and the concept of analysing a sufficiently large sample to warrant generalisation. Although an excellent introduction to how to use concordances, it is definitely not targeted at the typical language learner. It does however provide an indication of the extent of training a postgraduate language worker needs to become proficient in concordancing. Indeed, the observation that full-fledged, linguistic-researcher-like use of concordances might require considerable training is also a strong finding in the literature (Bernardini, 2004; Braun, 2007; Götz & Mukherjee, 2006; C. Kennedy & Miceli, 2001; Mukherjee, 2006; Pérez-Paredes et al., 2012; Tono et al., 2014).

An example of a study reporting on learners using concordances in an approach broadly comparable to that presented in *How to read concordances* is provided by C. Kennedy and Miceli (2001). They report on an undergraduate course that exposed language learners to concordancing in what they describe as an *apprenticeship approach*. Their approach revolved around the students modelling the teacher-researchers' own concordance use in relation to the work on grammar and writing undertaken in the classes; however, they found that, despite some successful usage, students often struggled to interrogate the corpus effectively and to generalise their results appropriately. They conclude that they had not provided their students with sufficient training in corpus research skills. Their conclusion is similar to the conclusion of a chapter in the *Routledge Handbook of Corpus Linguistics* that discusses how to prepare language learners to use concordances in detail and very much from a corpus linguistics perspective: learners need extensive training in how to generalise appropriately from the data (Sripicharn, 2010). Thus, while using a concordance of a representative corpus in a principled manner affords learners a data set from which rigorously generalisable observations can be made, there also seems to be considerable consensus that it is very difficult for language learners to use concordances in this way. A wide range of scholars subscribe to the position that the quantity of data involved and the

complexity of the research task mean that that such an approach to learner use of concordances would be too challenging (Ädel, 2010; Boulton, 2010a; Braun, 2007; Coxhead, 2014; Flowerdew, 2012a, 2012b; Franken, 2014; Gavioli, 2001; Hunston, 2002; C. Kennedy & Miceli, 2001; Levy, 1990; O'Sullivan & Chambers, 2006; Partington, 2001; Pérez-Paredes, 2010; Pérez-Paredes et al., 2012; Stevens, 1991a; Sun, 2003; Varley, 2009; H. Yoon & Hirvela, 2004).

### ***Pedagogical mediation and the authentication burden of concordances***

Given doubts as to the feasibility of researcher like use of concordances on the part of language learners, it is perhaps unsurprising that many researchers interested in learner use of concordances have suggested some form of teacher-mediation between the concordance and the learner in the form of pedagogical mediation. This subsection will first discuss the qualities of unmediated concordance data that are thought to make concordances difficult for language learners and then Henry Widdowson's (2000a, 2000b, 2003, 2004) commonly cited critiques of *authenticity*, deriving from the two a notion of the *authentication burden* of concordances, not as a constant, but as a variable in learner concordance use. The latter term not only serves to highlight the importance of Widdowson's critique of learner use of concordances, but also serves as a reminder that indicators of linguistic difficulty such as word frequency may have different implications in learner use of concordances as compared with other pedagogical activities. It is a reminder that we should be cautious in assuming that such factors will operate in exactly the same way in relation to learner use of concordances as compared with other pedagogical contexts. This subsection concludes with a summary of theoretical perspectives on how pedagogical mediation may be able to reduce the authentication burden of concordances.

From the perspective of pedagogical mediation and authentication burden, a distinction between concordance data as *samples* and concordance data as *examples* can be informative (Gavioli, 2005). Thinking of the concordance as a simple, unmediated sample of a language without consideration for the generalisability of any observations made, Gavioli argues that concordances provide a sample of occurrences because they are selected purely on the basis of orthographic form. In other words, the citations presented in a concordance are not exemplar uses that meet a pedagogical criterion of usefulness and so, Gavioli argues, they should not be considered examples. Although the equation of examples and exemplars can be disputed, the general point stands: concordances extracted randomly from a representative corpus represent

samples of language and, as such, they are very unlikely to represent exemplar instances of use. To this basic point can be added a number of more specific points.

First, a sample of concordance data will typically contain numerous exceptional, ambiguous, idiosyncratic, and sometimes even simply erroneous, occurrences (Sinclair, 1991). We can expect a proportion of such problematic citations to be presented to learners in either an exhaustive concordance or a randomly sampled concordance extract. Second, even when citations are not erroneous per se, few citations will turn out to be neat in the sense of exemplifying a desirable target structure without also containing extraneous features, meaning that the citations sampled may actually be more demanding on the learner than any individual structure contained in the concordance itself (Allan, 2009; Chujo, Oghigian, & Akasegawa, 2015; Frankenberg-Garcia, 2014b; Gilquin & Granger, 2010; Kilgariff, Husák, McAdam, Rundell, & Rychlý, 2008; Sinclair, 1991). To this it might be added that references to cultural or textual entities outside of the citation may also be particularly challenging for learners who may not be familiar with the cultural milieu of the texts or the conventions of the text type that the concordance is extracted from (Widdowson, 2004). Third, when working with a complete concordance, the number of citations retrieved may well exceed the amount of language data that it is reasonable to expect learners to cope with (Allan, 2010; Levy, 1990; Sripicharn, 2010; Tribble, 1997; Varley, 2009). Fourth, a single randomly compiled concordance may contain occurrences presenting more than one different pattern of use or usage, thereby complicating analysis (Gavioli, 2005; Sinclair, 2003). Fifth, the visibility of any particular pattern of use will depend on the proportion of citations that display that pattern and how dispersed they are (Frankenberg-Garcia, 2014b; Gavioli, 2001; Sinclair, 2004). That is, if there are only a small number of citations available, or the citations are very heterogeneous, underlying regularities in the data may be hard to ascertain. To summarise, an unmediated concordance, as a sample, may present language learners with data of considerable complexity, and for this reason, concordances can be seen as having the potential to put an unusually large burden on language learners.

Henry Widdowson has been a vocal critic of using concordances with language learners on the basis of his concept of authenticity as it pertains to data from a corpus or comparable language database (Widdowson, 2000a, 2000b, 2003, 2004). Widdowson's notion of *authenticity* can be usefully recast in terms of a variable in learner use of concordances, what I am terming a

concordance's *authentication burden*, an important variable in learner concordance use that has not been addressed in the research so far. However, before discussing the idea of *authentication burden* as a variable, it is important to be clear about what Widdowson's notion of *authenticity* refers to.

The crux of Widdowson's position is that the meaning conveyed by language is co-constructed with relevance to the context language is used in; therefore, the text entered into a language repository such as a corpus is shorn of its context with the consequence that it only retains a partial record of the meaning originally conveyed. To illustrate, a bystander witnessing a truck career erratically towards a bus stop is more likely to shout "*Hey, watch out!*" than "*Hey, person standing at the bus stop, watch out for that truck weaving all over the road!*" because the truck is visible, and the tone, volume and urgency of the utterance is sufficient to attract the intended recipients' attention. It is the context of the utterance that creates reference to the truck careering towards the message's recipient; it is contained within the textual trace of the communicative event. Even comparatively abstract texts are grounded in a social context that determines which information can be assumed as given and the various associations and connotations a particular language form will activate. In this way, Widdowson makes a distinction between *context* (the environment that language originates in) and *cotext* (the words that occur with another word). The language recorded in a corpus or language database is, according to Widdowson, only a partial record of a communicative event: texts record cotext, but omit context. In Widdowson's parlance, this means that the language found in a repository such as a corpus is *inauthentic*; language must be imbued with contextual meaning for it to be considered *authentic*.

Widdowson (1998, p. 711) critiqued calls for the use of what would conventionally be considered authentic language more widely than in relation to corpus or concordance use:

[What is commonly termed authentic] language cannot be authentic because the classroom cannot provide the contextual conditions for it to be authenticated by the learners ... Listeners can only authenticate it ... if they are insiders. But learners are outsiders, by definition, not members of user communities. So the language that is authentic for native speaker users cannot possibly be authentic for learners.

This position is somewhat hyperbolic: language learning would be a bafflingly solipsistic proposition if we were permanently excluded from any community of users in virtue of our

learner/outsider status. His point is that text can only be considered an appropriate input for learners to the extent that the learners are able to authenticate it: to paraphrase, to the extent that the classroom provides the contextual conditions for a text to be authenticated. This would appear to be the case whatever the texts origins. Thus, it is more useful to think of texts as presenting a relative *authentication burden* than as simply being authentic or not. If we think about texts as having variable degrees of authentication burden, it explains to some extent Widdowson's preoccupation with concordances as particularly unsuitable types of text for language learners.

At least two factors would appear to account for Widdowson's critique of learner use of corpora and, by extension, concordances. First, he appears to perceive a certain zealotry in the work of corpus-based pedagogues such as Sinclair, a sense that they were over-selling the value of conventionally-defined-authenticity (i.e. *not authored for pedagogical purposes*) at the expense of other factors such as comprehensibility and utility (Widdowson, 2003). Second, although Widdowson does not critique concordance extracts directly, his arguments would appear to be based on the idea that corpus-data was to be presented as concordance extracts, and so might present more of an authentication burden to language learners than the whole texts that composed the corpus itself. Three possible causes of a greater authentication burden can be identified within the literature on learner use of concordances: presenting insufficient contextual data about the language in the concordance; fracturing intra-textual contextualisation of language through the process of extracting and recompiling citations; and the possibility that citations that present short spans of cotext put significant additional strain on a reader's ability to contextualise language, as compared with either syntactically complete structures or longer spans of text more generally.

In terms of presenting insufficient contextual data about the language data in the concordance, in a worst case scenario, citations are extracted from a language database, and then presented without any indication as to where the citations originated from. Such decontextualisation is widely thought to be a significant barrier to the authentication of concordances as texts (Ädel, 2010; Braun, 2005, 2006, 2007; Flowerdew, 2012b; Götz & Mukherjee, 2006; Mishan, 2004). When decontextualized, the citations presented in a concordance may be thought to differ from whole texts in that whole texts will typically include contextualising information such as titles, an indication of authorship, publication details or

formatting indicative of text genre or text type (Mishan, 2004). However, the extent to which concordances are denuded of this contextualising information is highly variable. First, concordancers are extracted from indexes of extant texts, and thus many concordances provide users the ability to transition from citation to source text easily. Concordancing software will often allow a user to link to a complete source text at a click of a button (for example, WordSmith Tools Scott, 2016), and so when this functionality is available, concordances can be seen as providing as much contextual data as contained in the source texts themselves. At the same time though, it should be recognised that some concordance interfaces do not provide this functionality, and this functionality is unlikely to be available if static concordance extracts are used. Second, whether or not the source text for each citation is available, concordances vary in terms of how much information they provide about either the source texts as a whole or each citation individually: for example, the concordancer available online from [corpus.byu.edu](http://corpus.byu.edu) will present various types of meta-data about each of the citations it displays, information such as date of publication, genre, text name or publisher. Thus, although extracting citations from a text may result in readers losing contextualising information about a text, such as the title of the text, the author of a text or the genre of the text, this point should be balanced against that fact that concordances may also be used in close conjunction with the source texts that were concordanced and, even when they are not, they may also be annotated with recontextualising information. Indeed, the importance of providing learners with such information has been noted by a number of scholars (Ädel, 2010; Braun, 2005, 2006; Johns, 1986; Pérez-Paredes, 2010; Stevens, 1991a).

In terms of fracturing intra-textual contextualisation of language data through the process of extracting and recompiling citations, Widdowson points out that the meaning attributable to a language form develops over a text through co-textual relations within potentially broad spans of text (Widdowson, 2004). He gives the example of the interpretation of the words *close* and *shops*, words that can be used together in the description of a temporary state of closure reflecting business hours, but also the permanent closure of a business enterprise. To illustrate, in a text that has previously discussed the famous London shopping area *Oxford Street*, we are likely to read the former interpretation into an otherwise ambiguously phrased statement containing the words *close* and *shops*; if, however, we have previously read the phrase *industrial premises*, we are likely to interpret *close* and *shops* as reference to the latter sense. Intra-textual



connections provide a proportion of the context from which the meaning of the text at hand is constructed. Thus, because a concordance extracts citations from a wider textual environment, it divorces each extract from any meaning constituted in its relation to its wider cotext. Hence, a second possible reason that concordances may be more difficult to authenticate than whole texts is that each citation presented is stripped of the wider textual context it was generated within. Of course, as suggested above, this fracturing of intra-textual relations would not seem to be a major issue when access to the original texts is provided by a concordancer, but it may be more of an issue in hands-off use of concordances, especially if citations are presented with very short spans of cotext. It is also interesting to note that several scholars have suggested that learners use concordances of texts that they are already familiar with as this may make the citations in the concordance easier for learners to authenticate (Aston, 2001; Braun, 2005; Gavioli, 1997); being familiar with the texts concordanced may help the learner bring intra-textual connections to bear in the process of authenticating the concordance.

The third potential reason that concordances may be harder to authenticate than whole texts is the fact that citations are often extracted with short spans of cotext, and so the citations extracted may present incomplete syntactic structures or reference to text outside the span extracted. First, the most common concordance display format, key word in context (KWIC), presents a relatively limited span of cotext because it is typically limited to a single line of cotext and it will, thus, typically display syntactically incomplete structures. The potential for such citations to pose learners additional difficulties has been a common criticism of learner use of concordances (Aston, 2001; Boulton, 2009b; Cobb, 1999; Johns, 1986; Levy, 1990; Mishan, 2004; Stevens, 1991a; H. Yoon & Hirvela, 2004). Second, even if concordances are generated so as to present complete sentences, there may still be an issue with intersentential features of a text that affect comprehension of the citations presented: for example, pronouns (Kilgariff et al., 2008). Again, neither would appear to be a significant issue if the concordance used is dynamic and access to larger spans of text is available on demand, but it does seem likely to be an issue when using static concordances, especially when they are formatted to show relatively short spans of cotext, such as when KWIC formatting is used. Indeed, it has been noted that having access to even just slightly wider spans of cotext can have a significant effect on the interpretability of citations (Johns, 1986; Levy, 1990; Sinclair, 2004). Thus, as with the concern about intra-textual contextualisation, being limited to a very short span of cotext means there is

less information to aid the learner in authentication of the citation, while access to longer spans of context equates to there being more material from which to construe a plausible context for the citation to be interpreted within.

From the perspective of concordances as having the potential to present a greater authentication burden than other types of text, it is interesting to note that a study by Passapong Sripicharn found native speaker concordance users more likely to interpret concordances by drawing upon their intuition than non-native speaker concordance users, although it should be noted that this observation is confounded by the fact that only the non-native speakers in this study had been trained in an approach to reading concordances that did not draw on intuition (Sripicharn, 2004). Although attributing the differences observed to the participants' native and non-native speaker status is suspect as a consequence, it does indicate that, without training in inductive reasoning, concordance users may be predisposed to draw on their extant linguistic and cultural awareness to help them authenticate the language presented to them. Certainly, pre-existing linguistic and textual knowledge is commonly cited in connection to successful engagement with concordances (Aston, 2001; Braun, 2005, 2006; Gavioli, 2001; Gavioli & Aston, 2001; Johns, 1986; Partington, 2001; Sinclair, 2003, 2004; Tono et al., 2014). Gavioli and Aston (2001, p. 241) suggest:

A concordance does not make sense in itself: sense has to be attributed to it by the reader... a concordance can be viewed as text that provokes 'a pragmatic reaction' in the observer

Indeed, usage based theories of language such as Hoey's theory of Lexical Priming suggest that even a single word can prime its recipient with a wealth of information about features of usage such as register, syntax and collocation, garnered from their previous encounters with the word in context (Hoey, 2005). If this is the case, the burden of authenticating the kind of fragmented discontinuous text presented in a concordance may be far heavier for learners than for proficient language users because they are likely to find the same quantity of text activates far less information for them than it does for the more proficient language user; learners may be more reliant on explicit contextualising information than more proficient language users who may be more able to authenticate language extracts on the basis of priming: unconscious familiarity with linguistic, cultural and textual norms. If such a phenomenon does indeed exist, it would go some

way to explaining Widdowson's suspicion of concordances as an appropriate type of text for learner use (Widdowson, 2000a, 2000b, 2003, 2004); concordances may represent both a greater authentication burden than other text types in general as well as a type of authentication burden that learners are ill-equipped to deal with.

In sum, the issue of how to lower the authentication burden of concordances has been speculated upon by a number of scholars. It has been suggested that the authentication burden of concordances can be lowered by providing learners with access to the whole texts citations are drawn from, annotating concordance data with recontextualising information, or using concordances of texts that learners are already familiar with. Aston (2001) has suggested that, in some cases, it may be enough for learners to be using concordances of a type of text that they are already familiar with. Within the context of English for specific purposes, Charles (2012) and Lee and Swales (2006) both report that learners found it easy to use concordances that had been extracted from collections of texts that students themselves had identified as relevant to their own areas of expertise. More generally, it has been suggested that concordances of more narrowly defined collections of texts may be easier for learners to use than concordances extracted from broader, more heterogeneous corpora (Aston, 2001; Braun, 2005; Gavioli, 2005; Gavioli & Aston, 2001). In the broadest terms, Tribble (2013) argues that a match between learner and corpus concordanced is an essential element in fruitful learner use of concordances. These suggestions sit alongside the more general calls for the pedagogical selection of easier citations noted above. However, one thing seems very clear in the literature, concordances are not best viewed as simply authentic or inauthentic, but as presenting different degrees of authentication burden and, as a consequence, authentication burden is a variable in learner use of concordances that researchers should try to explore.

In the next section, the broader theme of this review is returned to, focusing this time on two pedagogically orientated categories of affordance: concordances as a source of concentrated input texts and concordances as ad hoc indexes to language learning resources. However, the important issues of pedagogical mediation and authentication burden are also explored in relation to each.

### **Concordances as a source of concentrated input text**

There are a number of concordance affordances that can be conceptualised as the affordances of a concordance as a type of text. Teachers and researchers have seen much potential in the kinds of concentrated texts produced by concordancing, and these potentials can be summarised as three main types of affordance: as an input text for making pedagogical generalisations, as an input text for providing massed encounters with a target form, and as an input text for observing the patterns of collocation and colligation that a target form occurs in. However, before discussing each of these affordance types in turn, it is important to note two important features of concordances as concentrated input texts: the affinity between concordances as concentrated input texts and pedagogical mediation, and the fact that equivalent concentrated texts could in theory be generated by processes other than concordancing.

One important feature of concordances as concentrated input texts is the affinity between concordances as texts and pedagogical mediation. If the affordances of a concordance as an input text are distinct from the affordances of concordances as a research tool, then the requirements of generalisability discussed above (exhaustiveness or random sampling combined with extraction from a representative corpus) are not relevant. If robust generalisability is not sought, it is possible to adopt pedagogical mediation strategies without compromising the affordances of concordances as concentrated input texts. The main two pedagogical mediation strategies would appear to be selecting which citations to present in a concordance, or selecting the texts to be concordanced on pedagogical grounds (see above). Each approach compromises the generalisability of the concordance generated, but does not compromise the affordances of a concordance as a source of concentrated input text as defined here. This is not to say that a robustly generalisable concordance cannot also be used as a concentrated input text; these two categories of affordance are not incompatible. However, it does mean that some of the difficulties presented to a learner by a robustly generalisable concordance sample can be addressed without jeopardizing concentrated input text affordances.

The second important feature of concordances as concentrated input texts is the fact that these affordances could be provided by any sort of similarly concentrated text, not only a concordance. Extracting occurrences of search terms along with limited spans of cotext and then compiling them with the extraneous, intervening cotext omitted produces a text with very high rates of occurrence of the search term: what we can call *concentrated text*. In theory at least,

texts could be directly authored so as to produce an equivalent level of concentration that could provide equivalent affordances as a source of input text; however, as the definition presented earlier in this chapter makes clear, for a text to be considered a concordance, it must have been generated through the process of indexing one or more texts, extracting citations, and then compiling them to form the concordance as text. In other words, this concentrated quality of concordances as texts is only indirectly related to the status of a concordance as an index, and generating a text with a high level of concentration is not dependent on concordancing. Thus, these affordances of concordances as a source of concentrated input text are primarily the ease with which such concentrated texts can be produced via concordancing, and this point has important practical implications in relation to pedagogical mediation. If concentrated texts are found to be pedagogically valuable but concordances require extensive pedagogical mediation, then we need to ask whether concordances are the best source of concentrated text or not.

There is a complex interaction between the affordances of concordances as concentrated input texts, the affinity between concentrated input text affordances and pedagogical mediation, and the practical value of concordancing as a source of input texts vis-à-vis other possible sources of input text. Arguably, the most valuable affordances of concordances as a source of input text are those that combine the concentrated input text affordances of concordances with affordances of concordances as indexes to ad hoc collections of language learning resources (discussed below). The reason for this is that concentrated input texts with appeal to broad demographics could be produced commercially, and therefore the convenience of the concordance as a source of concentrated input text may have less relevance. If a concentrated input text presenting an unusual search term in a narrow language variety is sought, the value of concordances as a means of producing such a concentrated input text would seem clear. However, by the same token, the combination of concentrated input text affordances and ad hoc index affordances would seem to preclude labour intensive forms of pedagogical mediation, such as manually choosing which citations to present in a concordance extract (see for example Boulton, 2009b; Johns 1991b). Consequently, there is a real need for the authentication burden of relatively unmediated concordances to be assessed and, hopefully, manipulated (see Chapters 5 and 6). However, in the subsections below, each subcategory of concentrated input text affordances is discussed without reference to the affordances of concordances as ad hoc indexes; the ad hoc index affordances are discussed separately later in this chapter.

### *Concordances as input texts for making pedagogical generalisations*

The first, and perhaps most commonly appealed to affordance of concordances is the capacity for concordances to provide an input text appropriate for making generalisations, regardless of the actual reliability of the generalisations made: that is, the process of making generalisations, not the product. In other words, one affordance of concordances as input texts is the suitability of the concordance for use in data-driven learning, as defined in Chapter 1.

In the classic data-driven learning task, the learner interrogates a concordance by answering a series of questions that scaffold the learner towards making a pedagogically appropriate generalisation. Johns suggests a basic rubric behind this kind of activity around the formula *identify-classify-generalise* (Johns, 1991b). Typically, the identify phase involves a series of questions that prompt the learner to examine multiple concordance lines, making note of particular language features: for instance, which preposition most commonly follows the search term. In the classify phase, learners are prompted to group citations on the basis of a formal or functional criterion: for example, which citations are in the active and which are in the passive, or which citations refer to ideas and which to institutions. In the final phase, learners are asked to go beyond the data and to generalise the pattern they have observed. While Boulton (2011) is correct to observe that data-driven learning defies a single watertight definition, this kind of exercise can be seen as the core, or prototypical, exercise type in data-driven learning: exercises in which concordances afford learners a concentrated input text that affords a basis for making what we may call *pedagogical generalisations*. The extent of teacher mediation in the process can vary considerably, from none through to manual selection of citations from a concordance of a collection of text compiled on a pedagogical criterion, but because this affordance focuses on the process and not the product, it does not require the strict methodological criterion required of concordances as a research tool for making *robust generalisations*. It does not require a sufficient quantity of data for reliable generalisation, an unbiased sample, not a representative source text. This means teachers are free to mediate between learner and concordance and thereby circumvent many of the problems posed by strict samples of language corpora that have been discussed earlier in this chapter.

In this regard, it is also worth noting that a randomly sampled concordance extract composed of a small number of citations is not guaranteed to provide sufficient exemplification

of a word to support a particular, pre-specified generalisation. I experienced this very clearly when attempting to replicate concordance exercises presented in *Exploring academic English: A workbook for student essay writing* (Thurston & Candlin, 1997). I had been attempting to design an experiment in which the guided research exercise segment of Thurston and Candlin's *Look-Familiarise-Practice-Create* exercise design is compared with two other common learner use of concordance exercises: a gap-fill exercise and an error correction exercise. To this end, it seemed desirable for the three exercises to be based on a single concordance generated independently of the three exercise types to be examined. However, it soon became clear that the questions guiding the learner towards generalisations in *Exploring academic English* could not be transferred to concordances of the same size and of the same word but generated with reference to a different corpus and randomly sampled. If I wanted to use the same questions, I would need to use the same concordance as presented in *Exploring academic English*, or else to select citations to present in the concordance on the basis of the questions in the exercise. That is, to choose citations that provided suitable answers to the questions in the exercise and, thereby, supported the making of pedagogically appropriate generalisations. This experience illustrates the difference between the robust generalisability of concordances used as research tools and concordances as concentrated input texts for making pedagogically appropriate generalisations: while the former affords a basis for scientifically meaningful generalisations, the latter affords an opportunity for engaging with language so as to make generalisations; the former precludes pedagogical mediation, the latter all but requires it.

This kind of concordance use is held to offer a range of interrelated language learning affordances: supporting constructivist learning, promoting learner autonomy and engagement and developing lifelong learning skills (Aston, 2001; Bernardini, 2004; Boulton, 2009b, 2010a; Braun, 2005, 2007; Chambers, 2010; Cobb, 1999; Cotos, 2014; Daskalovska, 2013; Flowerdew, 2012b; Götz & Mukherjee, 2006; Johns, 1986, 1991b; Kaur & Hegelheimer, 2005; Park, 2012; Stevens, 1995; H. Yoon & Hirvela, 2004). The important point here is that these affordances do not require learners to use random samples or to process concordances exhaustively, and nor do they require learners to be using concordances of source texts that meet any particular criterion of representativeness. In fact, learners can engage with a concordance in this manner even if the concordance is derived from graded reading materials (Allan, 2009, 2010; Aston, 2001; Boulton, 2009b; Cobb, 1999; Flowerdew, 2012b; Römer, 2011). The authenticity of the source texts is not

a prerequisite of this kind of affordance. Indeed, as discussed above, this kind of learning does not even necessitate use of a concordance, although concordances would appear to be a very convenient way of producing input texts that can facilitate this kind of approach to an input text — regardless of whether the learner is using a concordance hands-on but engaging with the concordance in a more ad hoc manner than a researcher would, or whether the concordance has been pedagogically mediated in one manner or another by their teacher. By bringing together otherwise dispersed occurrences, a concordance as a concentrated input text can facilitate comparison and contrast amongst citations, and thereby facilitate pedagogical generalisation.

***Concordances as an input text for providing massed encounters with a target form***

A second perspective on concordances as input texts emphasises concentrated exposure to a target form in context without emphasis on the metaphor of learner as linguistic researcher or the affordance of pedagogical generalisation. There are a number of slightly different perspectives on this affordance, but these perspectives can be glossed as concordances providing unusually rich exposure to the search term in context (Cotos, 2014; Daskalovska, 2013; Götz & Mukherjee, 2006; Levy, 1990; Römer, 2010, 2011; Thurston & Candlin, 1998; H. Yoon & Hirvela, 2004). This is thought to help make the search term more salient in the input text (Boulton, 2009b; Cotos, 2014), and this affordance of concordances resonates with theories of second language acquisition that emphasise the relationship between noticing and acquisition (see Flowerdew, 2012b, 2015). As with all of the affordances discussed, the affordances of concordances as input texts for massed encounters are not mutually exclusive of other affordances that may well be available simultaneously within a single concordance. Consequently, it may not be immediately obvious that the massed encounters affordances of concordances differ from the pedagogical generalisations affordances. The difference can however be effectively illustrated by considering two examples of massed encounter affordances that primarily see concordances as an input text for reading, and not as an input text for analysis.

Cobb (1997, 1999) argues that concordances have an important role to play in vocabulary learning. Drawing on research in reading and vocabulary learning, Cobb argues that multiple exposures to a word in context are vitally important for developing a nuanced understanding of a word in a foreign language: depth of vocabulary knowledge. He argues that, by providing opportunities to meet a word in multiple contexts, learners are able to read a wide range of words



in context at rates of encounter that would be extremely unlikely to be provided by whole texts. His emphasis is not on a processes of inductive reasoning (progressive stages of identification, classification and generalisation), but on simple reading exposure: not the conscious application of inductive learning strategies, but on naturalistic learning from an unusually concentrated input text. Cobb and Horst (2001) and Horst, Cobb, and Nicolae (2005) take a similar approach, positioning the learner as dictionary maker and providing the learner with a concordancer to use as a tool for identifying useful sentences. Again, the concordancer is not positioned as a resource for facilitating conscious comparison and contrast between citations; rather, it is a convenient means of providing massed encounters.

Bernardini's approach to learner use of concordances also emphasises the affordances of concordances in presenting a search term in context en masse (Bernardini, 2000, 2002, 2004). Bernardini does not explicitly locate her approach within vocabulary and reading research, but she does see reading a concordance as an engaging way of navigating text in an approach she describes as serendipitous, discovery learning. In Bernardini's approach, learners navigate a text via a concordancer, exploring instances of interesting or unexpected usage; the idea being that happening upon sufficiently illuminating instances of usage via increasingly autonomous browsing, the learners' understanding of the search term progresses beyond that of their initial starting point. Her metaphor for this approach is *learner as traveller*. It is not premised upon learners engaging with a concordance in a manner that explicitly draws on inductive reasoning as in the learner as researcher paradigm; rather, in her approach, the concordance is simply a means of providing sufficiently rich exposure to the search term in context for reading to facilitate learning.

As these examples illustrate, concordances as concentrated input texts can afford language learning opportunities distinct from those of pedagogical generalisations; concordances can afford learners rich exposure to a target search form in a broader sense. Interestingly, while Cobb's research has tended to use a limited range of source texts selected on pedagogical grounds often accessed through a concordancer embedded within a CALL environment, Bernardini has tended to expose her students to a wide range of corpora accessed through a variety of standalone concordancers. Hence, while both approaches have emphasised rich exposure via concordancing, it seems very likely that the specific qualities of this rich exposure may have differed significantly in terms of both the quantities and qualities of the input texts

provided. Thus, while any concordance can be seen as having the potential to provide rich exposure to the search term, there is no reason to think that the exposure provided is a fixed variable in learning conditions: the texts concordanced, the quantity of citations found and how the citations are presented to the learner are all potentially significant variables in concordances as sources of massed encounters.

### ***Concordances as input texts for observing patterns of collocation and colligation***

A third theoretical perspective on learner use of concordances as input texts focuses on concordances as an input text for observing patterns of co-occurrence (Aston, 2001; Boulton, 2009b; Braun, 2005, 2006; Charles, 2012; Daskalovska, 2013; Flowerdew, 2012b; Gavioli, 2001; Götz & Mukherjee, 2006; Johns, 1991b; Kaur & Hegelheimer, 2005; Levy, 1990; McKay, 1980; Mishan, 2004; Römer, 2010, 2011; Sinclair, 2004; Sun & Wang, 2003; Tono et al., 2014; H. Yoon & Hirvela, 2004). This perspective differs from that of concordances as an input text for providing massed encounters with a target form because affording learners opportunities to observe patterns of co-occurrence focuses on regularities in cotext, and not just rich encounters with the word. That is, if there is no recurrent pattern of usage in massed encounters, the absence of reoccurrence does not undermine that affordance: a pattern does not need to be discernible for the citations to provide rich exposure. Consequently, even a handful of citations can be seen as affording learners rich exposure through massed encounters. However, the situation is somewhat different with concordances as an input text for observing patterns of co-occurrence. The observability of a pattern depends on how frequently a particular co-occurrence occurs relative to others (Frankenberg-Garcia, 2014b; Gavioli, 2001; Sinclair, 2004). For many search terms, a small number of randomly selected citations are unlikely to make any particular pattern of collocation or colligation discernible. Thus, the minimum conditions for a concordance to be able to afford learners massed encounters are different to those of opportunities to observe patterns of collocation or colligation. This is an issue that is also explored in Chapter 6.

Affording the opportunity to observe patterns also differs from pedagogical generalisation affordances because concordances as inputs which afford learners opportunities to observe patterns emphasizes observation over inference, although the two concepts are indeed closely related and may be present in a single concordance simultaneously. The key difference between these two types of affordance is that the former is focused on concordances as affording

a visually beneficial text type (observation of patterns), the latter, a text type that supports particular types of interrogation (data-driven learning). That is, the pattern observation affordances of concordances are derived from the capacity for concordancers to present citations in display formats that visually highlight such patterns and so do not require regularities to be discovered via cognitive processes of interrogation while, in contrast, the conscious processes of data-driven learning can be used to find semantic regularities that might not be particularly visually salient: for example, co-occurrence with a particular word having a particular semantic prosody. Thus, pattern observation affordances are distinguishable from data-driven learning affordances and, furthermore, it is these observability of patterns affordances that are strongly identified with the key word in context (KWIC) format of displaying concordances (Barlow, 2004; Gilquin & Granger, 2010; Hunston, 2002, 2010; Sinclair, 1991; Tribble, 2010).

In the KWIC display format, the spatial relationships between words in the cotext of different citations is regular on the texts vertical axis, at least to the extent that word lengths are similar. The search term (or key word) of each citation is presented in a single column in the centre of the text. Consequently, the words that occur directly before and after each citation of the search term are also presented in columns as a result of their co-occurrence with the key word. Indeed, the same effect can be observed further from the key word until differences in word length break the pattern of vertical alignment (typically two to five words before and after the search term). In KWIC, the span of cotext presented with each citation is normally limited to a single line because, if citations that spanned more than a single line were presented, each line of spatially meaningful cotext would become interspersed with cotext that was spatially arbitrary in relation to the key word (Sinclair, 1991). Thus, the basic idea of the observability of patterns of co-occurrence is that the spatial regularities in the cotext of KWIC format concordances allow the immediate cotext to be read along the vertical axis of the text, revealing repeated syntagmatic relationships and the paradigmatic associations of the key word.

The capacity of KWIC format concordances to highlight patterns of co-occurrence has led many researchers to speculate on their potential as an aid to learning a word's typical patterns of collocation and colligation (Aston, 2001; Basturkmen, 2006; Bernardini, 2002; Boulton, 2009b; Braun, 2005; Cobb & Boulton, 2015; Hunston, 2002; Johns, 1986; Mishan, 2004; O'Keefe, McCarthy, & Carter, 2007; Sinclair, 2004; Tono et al., 2014; Warren, 2011; Willis, 1998), and KWIC format concordances presenting relatively short spans of cotext in particular

(Aston, 2001; Boulton, 2009b; Cobb, 1999; Lai & Chen, 2013; Levy, 1990). In other words, KWIC format concordances have been hypothesized to afford learners a powerful opportunity to learn a word's patterns of collocation and colligation through observation. However, it is also important to realise that KWIC formatting is neither a necessary nor sufficient condition of concordancing: concordances can be displayed in other formats, such as left-aligned sentence format, and in theory at least, a text could be manually authored so that a particular word occurred in the central position of each discontinuous line displayed. Again, in essence, it is the convenience of generating such a text via concordancing that would appear to constitute this affordance of concordances as input texts, as it is with the other concordance as concentrated input text affordances.

Furthermore, as suggested above, the extent to which any text is actually able to afford this type of language learning opportunity is dependent upon the extent to which the concentrated text does actually contain repetition. For example, it is well recognised that sorting citations alphabetically (for instance, by one word left of the search term, or three words to the search term's right) can powerfully enhance the observability of patterns of co-occurrence (Aston, 2001; Charles, 2012; Hunston, 2010; Johns, 1986, 1991b; Levy, 1990; Partington, 2001; Sinclair, 1991, 2004; Tognini-Bonelli, 2001; Tono et al., 2014; Tribble, 2010). This facility clearly has a synergy with hands-on concordancing, assuming a concordancer has this functionality and learners have been shown how to use it. The situation seems more complex in regards to use of static concordances and small random samples. First, a static concordance can obviously not be reordered; this is significant because different permutations of how citations are ordered are likely to highlight different patterns of co-occurrence. Second, as indicated at the beginning of this subsection, small random samples from large general corpora may contain too sparse a trace of patterns of co-occurrence for patterns to be observable, however they have been arranged. Of course, as with the other affordances of concordances as concentrated input texts, when this affordance is sought without reference to robust generalisability, pedagogical mediation can intervene between learner and concordance so as to ensure a concordance presents sufficiently observable patterning, and this once again raises the issue of teacher-use versus learner-use and whether the amount of effort involved in mediation becomes excessive.

### *Final comments on concordances as input texts*

These three affordances of concordances as concentrated input texts (an input text for making pedagogical generalisations, an input text for providing massed encounters, and an input text for observing patterns of co-occurrence) are technically separable from concordance use per se; consequently, concordances as a source of input text are perhaps best thought of as a convenient means of generating input texts that possess particular properties. Furthermore, these affordances are not necessarily in conflict with the affordances of a concordance as a research tool: a single concordance may afford any combination of these affordances if the criterion for affording each is met. Thus, if any of these affordances are sought without recourse to the affordances of concordances as a source of robustly generalisable language data, learner use of concordances can be pedagogically mediated to reduce the difficulty of concordance use for the learner. However, such mediation not only undermines claims to the generalisability of any observations made, but also problematizes the extent to which concordances are a convenient means of providing concentrated input texts: the more extensive the mediation, the less clear it is that the concordance provides convenience to the learner and not to the teacher or materials designer.

The issue of the trade-off between efforts spent in pedagogical mediation versus benefits to learners is an important issue (Boulton, 2010a; Braun, 2005; Frankenberg-Garcia, 2014a; Johns, 1991b), although it has so far received relatively little attention in the literature. While Gavioli and Aston (2001) emphasize the fact pedagogical mediation exists on a cline and that various degrees of pedagogical mediation are available, Tribble (2013) warns that in many situations extensive mediation may be required and the effort may actually outweigh the benefits. Hence, logically, it becomes necessary to ask at what point authoring input texts that offer opportunities for pedagogical generalisation, massed encounters or observation of patterns might prove more convenient, or even more effective, than mediating between learner and concordance. This issue is important for this thesis because, if concordances cannot always be extensively mediated for practical reasons, there is value in assessing the affordances and authentication burdens of concordances that have not been manually mediated (see Chapters 5 and 6). Furthermore, as will be seen in the next section, when we consider the affordances of concordances as indexes of ad hoc language learning materials (affordances that can be combined with the affordances of concordances as concentrated input texts), then it becomes clear that manual pedagogical mediation of concordances would severely limit the affordances of

concordances from this perspective: many of the most unique and valuable affordances of concordances depend on learners being able to overcome the authentication burden of relatively unmediated concordances.

### **Concordances as an index to ad hoc language learning resources**

A third major category of affordances is derived from the capacity of concordances to provide a window on text that is approached as a language learning resource: specifically, resources that exemplify authentic language use, varieties of language not exemplified sufficiently in commercial materials and reference resources that exemplify language use relevant to the language learner's context of language use when using the concordance. Before discussing each of these affordances in detail, two points should be noted. First, these affordances are not in conflict with the previously discussed affordances of concordances; in fact, they are primarily complementary. Second, the ad hoc status is defined in contradistinction to the availability of equivalent commercial resources. By facilitating ad hoc resource access, these affordances of concordances balance the trade-off between pedagogical mediation effort and benefit to learners in favour of the benefits brought to learners; there will always be a limit to how much pre-processed exemplification will be commercially available (Frankenberg-Garcia, 2014a), and the pre-processed materials commercially available will perforce be targeted at the most profitable consumer base. Thus, the ad hoc affordances of concordances cast the trade-off between efforts spent in pedagogical mediation and benefits to learners in a new light, and have strong affinity with the use of concordances in language for specific purposes (LSP) work (Bernardini, 2004; Charles, 2012; Coxhead, 2014; Gavioli, 2005; Lee & Swales, 2006; Levy, 1990; Park, 2012; Römer, 2011; Tribble, 2013; H. Yoon & Hirvela, 2004).

The capacity for a concordance to locate orthographic forms within a text allows concordances to provide specific qualities of text that may be otherwise unavailable, to turn any available text into a language learning resource with particular qualities: this is the ad hoc character of these affordances, derived from the essential indexing functionality of concordances. This affordance of concordances may be used to provide learners with concentrated input texts, as discussed above, but drawn from specific varieties of text; however it may also be seen as affording specific qualities of exemplification, quite aside from whether the concordance as input text is concentrated or not. To the extent that these types of concentrated text or qualities of

exemplification are chiefly available through concordancing, these are inherently the affordances of concordances. The implications of this point are subtle but important: there seems to be a tension between ad hoc access and pedagogical mediation. The more pedagogical mediation that is involved, the less ad hoc access would appear to be afforded to the learner, as opposed to the materials designer. As suggested earlier in this chapter, considerations such as this point to the importance of assessing the authentication burden of relatively unmediated texts, such as texts that have been mediated through choice of text concordanced, rather than the comparatively labour intensive manual selection of which citations to display.

In the next three subsections, three specific qualities of exemplification available through concordances as ad hoc indexes to language learning resources are discussed: access to authentic exemplification, exemplification drawn from specific domains of language use, and exemplification with task specific relevance. These specific qualities of exemplification can be seen as three of the most important affordances of concordances for language learners. However, as we will see, they are not dependent on using KWIC format concordances (cf. Chapter 3), they are not dependent on using concordances in a manner comparable to researchers (cf. Chapter 4), but they are likely to be amenable to pedagogical mediation through the selection of which texts to concordance (cf. Chapter 5 and 6).

### *Concordances as a source of authentic examples*

One of the original motivations for presenting learners with concordance extracts was the notion that language learning materials were not presenting language accurately; the idea being that concordances could overcome this problem by providing learners access to examples of authentic language use (Johns, 1986, 1991a, 1991b, 2002; Levy, 1990; Stevens, 1991a; Tribble & Jones, 1990). Johns put the case forcefully, suggesting that the grammars available at his time of writing were “usually rather than exceptionally incomplete, partial and misleading” (Johns, 1991b). He also argued that simplification of language input was also distortion of language input because language form and function are, in the final analysis, inseparable (Johns, 1991a). From the perspective of concordances as ad hoc resources, while authentic texts are all around us in the form of books, magazines, webpages and so forth, the indexing functionality of concordances offers a means of rapidly locating particular orthographic forms within authentic texts. A single instance of a target form in a huge collection of text can be located and extracted

almost instantaneously. Thus, concordances can facilitate authentic exemplification of a target form.

There are two important points to note about this view of the affordances of concordances. First, the notion of concordance extracts as authentic text has inspired a certain amount of controversy that still has implications for successful concordance exploitation today (Widdowson, 2000a, 2000b, 2003, 2004). As discussed above, Widdowson is primarily concerned with *authentic* in the sense of *decontextualized*, and so his arguments are not directly related to authentic in Johns' sense, that of *attested*. As argued above, we can see any text as presenting an authentication burden, and the authentication burden of concordances is surely an important variable in learner use of concordances (see Chapters 5 and 6). I will discuss Widdowson's notion of *authentic* in terms of a text having an *authentication burden*, reserving the unadorned use of *authentic* for the more common distinction between texts that contain attested use and texts authored for pedagogical purposes such as graded readers and language learning text books. This brings us to the second point, that learner use of concordances does not necessarily entail concordancing authentic texts. A number of scholars have indeed proposed using concordances of graded readers with language learners (Allan, 2009, 2010; Aston, 2001; Boulton, 2009b; Cobb, 1999; Flowerdew, 2012b; Römer, 2011). However, it is not entirely clear whether the process of compiling a concordance from such materials would still result in a text with a higher authentication burden than reading a comparable quantity of continuous text, because of the qualities of concordances as fragmented texts (see Pedagogical mediation and the authentication burden of concordances, p. 40).

Finally, it is worth noting that the capacity for concordances to afford learners access to authentic materials has perhaps less relevance to pedagogy than it once did. Although commercially available language learning materials have often been criticised for presenting a skewed or unrealistic picture of language use as attested in corpora (Allan, 2009, 2010; Aston, 2001; Boulton, 2009b; Cobb, 1999; Flowerdew, 2012b; Römer, 2011), there are also claims that language learning materials are now commonly produced with reference to corpus data (Boulton, 2010a; Braun, 2006; Cotos, 2014; Frankenberg-Garcia, 2014a; M. McCarthy & McCarten, 2012; Mukherjee, 2006; Römer, 2010), and so we might expect that the gap between the representations of language in authentic and inauthentic materials to be less of an issue than it once was. Although materials are still likely to have been pedagogically mediated, they are no



longer so likely to be misleading, from the corpus data perspective at least. Thus, while concordances may afford learners examples of authentic usage, they are not the only source of exposure to authentic examples (Frankenberg-Garcia, 2014a) and so, considered in isolation, this affordance of concordances would appear to be less attractive than it may have been in the nineteen-nineties.

### *Concordances as a source of domain specific examples*

The value of access to authentic examples in a simple, undifferentiated sense may be something of a moot point in the era of corpus informed language learning materials and resources, but the capacity to identify authentic examples from a specific domain or variety of language is perhaps more appealing. If a learner simply requires an example of a word in attested use, this can be easily accessed through a corpus-based dictionary, or even just a search engine such as Google. However, if a learner requires an example drawn from a more specific domain of language, how to access such examples is less clear. If, however, a learner has access to one or more texts of the specified category and a concordancer, a teacher or learner with moderate computer literacy can quickly and easily locate every example of a search term in those texts: concordances can afford learners authentic examples of domain specific use. Little wonder then that concordances are thought to have particularly clear affordances in the context of language for specific purposes (LSP) (Bernardini, 2004; Charles, 2012; Coxhead, 2014; Gavioli, 2005; Lee & Swales, 2006; Levy, 1990; Park, 2012; Römer, 2011; Tribble, 2013; H. Yoon & Hirvela, 2004). Concordances provide the potential for domain specific exemplification where market forces inhibit provision on a commercial basis. Hence, this affordance of concordances also has strong synergy with the affordances of concordances as concentrated input texts discussed above, even if concordance use does still require pedagogical mediation of some kind. If commercially available input texts are not drawn from a domain of language use with relevance to the learners' needs, a concordance provides a means of generating a concentrated input text that does. A parallel point can be made with regards to authenticity. The trade-off between pedagogical mediation and learner benefit seems likely to be more worthwhile when a concordance's capacity to provide exemplification of narrow domains of language use is operative: an affordance that is only curtailed by the availability of appropriate source text.

In the context of LSP use of concordances, Gavioli (2005) has argued for a different criterion of generalisability to the robust generalisability of observation at the extreme end of the learner/research paradigm of use. She argues that in LSP contexts in which concordance use is undertaken towards specific goals relevant to the learners' needs, generalisations need only satisfy a criterion of *local relevance*. The upshot of her suggestion is that the representativeness of the corpus used or the rigorousness of the analysis undertaken need only meet a criterion appropriate to the learners' end goals, that the sort of authority imparted by a rigorous corpus linguistic analysis might not be so relevant. Just as the degree of mediation required in learner concordance use can be seen to exist along a cline (Gavioli & Aston, 2001), so too may the value and rigour of generalisations made with reference to a concordance. Where pre-digested descriptions of a language variety are not readily available, there may still be value in the product of learner generalisations from concordances, even when the criterion of robust generalisability has not been met.

As should be clear by now, the affordances of concordances as a source of domain specific exemplification should not be seen as incompatible with the other affordances discussed so far. Although this affordance can be understood in the simple sense of the ability to find examples from particular domains of language use or text types (providing that is that one or more appropriate texts are available), it has considerable synergy with each of the other affordances discussed so far. In combination with other types of affordance, the affordances of concordances as a source of domain specific exemplification would appear to add the extra requirement on concordance use that text of an appropriate type is available in the required quantity; likewise, any conditions of concordance use entailed by another affordance sought simultaneously would still obtain. The important point here is that there do indeed appear to be multiple affordances of concordances, and the conditions for realising these affordances would appear to vary, suggesting that concordancing may well be best conceived of as a complex, multi-faceted construct (see Chapter 3) in which different affordances of concordances are realised by different operationalisation of concordancing and typically drawn upon by different groups of concordance users (see Chapter 4).

### *Concordances as a source of task relevant examples*

A different way in which the examples indexed by a concordancer may benefit language learners is the capacity for a concordancer to help a learner find examples that have particular relevance to the learner's immediate needs, such as a task they are engaged in (Charles, 2012; Franken, 2014; Frankenberg-Garcia, 2014a, 2014b; Gavioli, 2005; Johns, 1986; Lai & Chen, 2013; Lee & Swales, 2006; Park, 2012). For instance, a learner may want to find an example sentence that parallels a sentence they are composing or editing. This criterion is different to that of domain specificity because domain specific examples do not require the examples provided by the concordance to meet any criterion of relevance to a concordance-external text or context of use. It seems unlikely that static concordances or short concordance extracts could be sufficient to afford learners a concordance containing strongly task relevant citations (though see Frankenberg-Garcia, 2014b for studies that examine the selection of citations to support either production or reception). However, in hands-on concordancing where the search-term can be specified, processed, re-specified, and re-processed, with sufficient proficiency in constructing a search term a learner may well be able to iteratively filter the examples available to them through a concordance and thereby maximise their chance of identifying a citation with maximum relevance to their context of language use (cf. Pérez-Paredes et al., 2012). Indeed, there is some evidence in the literature that this is so, especially when learners are using concordances of texts drawn from their areas of academic expertise (Charles, 2012; Gavioli, 2005; Lee & Swales, 2006; Park, 2012).

There has been much excellent research on learner use of concordances that has used techniques such key tracking, screen capture, and logs of learner use to examine how learners construct search terms and correlate learner searches with editing of their texts (Franken, 2014; Lai & Chen, 2013; Park, 2012; Pérez-Paredes et al., 2011; Pérez-Paredes et al., 2012). It has shown that learners typically struggle to form the advanced search terms that would most effectively help them to find citations with the most relevance to their task (Park, 2012; Pérez-Paredes et al., 2011; Pérez-Paredes et al., 2012). However, studies such as these have also shown that learners can indeed construct simple search terms appropriate to their needs and use the citations accessed effectively (Franken, 2014; Park, 2012; Pérez-Paredes et al., 2012).

Interestingly, after their 2001 study in which they took a research orientated approach to learner concordance use (discussed above under the heading Concordances as a research tool, p.

39), C. Kennedy and Miceli (2010) took a somewhat different approach in their follow up study. Whereas their 2001 paper took a *learner as apprentice* approach in which the learners were inducted in research style concordance use, their 2010 study emphasised the affordances of concordances as a source of examples relevant to the specific context of use in the learners' writing. They conclude from their studies that:

We should think in terms of having a set of distinct reference-resource functions at our disposal, rather than simply a suite of reference resources, for what matters is not mastery of a resource but of the function(s) it can be used for [p. 40]

Another take on this point would be that concordances appear to have multiple affordances for language learners, that many of these affordances would appear to be constituted in a concordance as an ad hoc language learning resource, and that the affordances of concordances for language learners are not simply an extension of the affordances of concordances in corpus linguistics. These are all issues taken up in the studies reported in Chapters 3, 4, 5 and 6. In the next section, these issues are discussed in relation to variables in learner use of concordances.

### **Variables in learner use of concordances**

When we look at the theoretical perspectives on learner use of concordances discussed in the literature, we see that the affordances sought (which can themselves be seen as one variable in learner use of concordances) exist alongside at least five others: the task the learner engages in, the texts concordanced, the selection of the citations presented in the concordance, the format the concordance is presented in, and whether the learner has access to a static or dynamic concordance text. In this subsection, the main levels within each of these variables is discussed, and connections are made between these variables and the studies reported in Chapters 3, 4, 5 and 6.

### **The affordances sought**

The literature shows that concordances can be seen as presenting a wide array of language learning affordances, and it has been suggested that the affordances of concordances can be conceptualised in terms of three main types: opportunities for making robust generalisations, opportunities for engaging with concentrated input texts and the opportunity to use texts as ad hoc language learning resources. In Table 1, these three categories of affordance are presented as levels, along with notes outlining the main features of each level.

One important thing to note about affordances as a variable in learner use of concordances is the varying degree of opportunity for pedagogical mediation that is possible. Robust generalisation affordances would appear to preclude mediation altogether. Ad hoc access affordances seem likely to severely limit the extent of mediation possible. Thus, it seems several levels within this variable are contingent upon learners being able to overcome the authentication burden of concordances. Consequently, Chapters 5 and 6 begin to quantify aspects of authentication burden.

### **The task the learner engages in**

Another variable that emerges in the literature is the task that the learner is envisioned engaging in. This is similar to the notion of affordances, but whereas affordances are focused on the outcomes of learner concordance use, task is focused on the manner of concordance use. Although there are many different tasks learners can be asked to engage in with concordances, a potentially key theoretical distinction can be made on the basis of the learners' pattern of engagement with a text. It is suggested here that, in a broad sense, a distinction can be drawn between tasks that require analysis of the concordance and tasks that require reference to the concordance.

*Table 1 Affordances sought*

Variable	Main levels	Key points
Affordances sought	Robust generalisation	Requires complete or principled sample of representative corpora; likely to be very difficult
	Concentration	A convenience of concordancing, no formal requirements, and so these affordances remain available even when a concordance has been pedagogically mediated; note, texts could also be authored to achieve same effect, calling into question the efficiency of using concordances to provide these affordances in isolation from other affordances
	Ad hoc access	Access to appropriate texts is the only formal requirement. Concordancing as indexing is essential to the characterisation of these affordances, but limited room for pedagogical mediation implicit in notion of ad hoc access

*Table 2 Task engaged in*

Variable	Main levels	Key points
Task engaged in	Analysis	Focused on comparison of multiple citations. In combination with a concordance suitable for robust analysis, likely to be very challenging; if undertaken with pedagogically mediated concordances, task difficulty can be reduced, but authenticity and generalisability threatened. Note, if it is the process of analysis and not the product that is seen as the main affordance, equivalent tasks could be undertaken with authored texts
	Reference	Focused on relevance of citations to a concordance-external context of language use: for instance, relevance to an essay being edited, a text being translated or a sentence being corrected. Note, in some circumstances, a single useful citation may suffice.

There are a good many theoretical perspectives on how input texts such as concordances facilitate second language acquisition: the best known of which are probably the noticing hypothesis (Schmidt, 1990, 1992), which emphasises the role of consciousness in attending to an input in acquisition; related work by Nick Ellis that emphasises the interaction between explicit attention and implicit learning (Ellis, 1995, 2008, 2015); and the depth of processing hypothesis ( Craik, 2002; Craik & Lockhart, 1972), which contends that an input can be processed at different levels of processing within memory with correspondingly different effects on retention. In the context of learning vocabulary, Schmitt (2008) usefully summarises the dominant theme of such theories of second language acquisition: more and better engagement with the target language leads to more and better learning. The important point to note in regards to learner use of concordances is that there would appear to be at least two very different patterns of engagement with a concordance. In the first pattern of engagement, the citations are compared and contrasted with each other; this pattern of engagement is typical of classic data-driven learning tasks as discussed above, and has been glossed in the table below as *analysis*. The second pattern of engagement involves comparing citations to an external context of use, such as a text being translated or an essay being edited, and not the comparison of the citations within the concordance with each other. This has been glossed as *reference* in the table below. Whether they provide equivalent quantities and qualities of engagement with the concordance is an empirical question, but it would seem they have the potential to differ in this regard, and so they would appear to warrant further investigation as potential variables in learner concordance use, albeit that these broad patterns of engagement could be further subdivided into a narrower typology of concordance based tasks (for an early attempt at a typology of concordancing tasks, see Honeyfield, 1989).

Again, treated as a variable, the authentication burden of the concordance used would seem likely to be of importance, especially in *reference* type tasks in which concordances would seem to need to be processed in much the same as ordinary texts. In *analysis* type tasks, we might expect learners to be engaging with the text via *vertical reading*: a term used by some scholars to denote engagement with the cotext of a concordance along its vertical axis (Boulton, 2009a, 2009b; Braun, 2005; O'Keeffe & McCarthy, 2010; Pérez-Paredes et al., 2012; Tognini-Bonelli, 2001, 2010). To the extent that *analysis* type tasks require learners to engage in vertical reading, we might not expect traditional reading skills to be drawn on as much as they would

appear to be in reference type tasks, but sufficient proficiency in the language for word recognition would presumably still be an advantage to the learner, even when “reading” a concordance vertically. As such, the research reported in Chapters 5 and 6 is also relevant to *task engaged in* as a variable. This variable is explored directly in Chapters 3 and 4, helping us to move beyond this a priori account of differences in how concordances are engaged with and towards richer accounts of how concordancing is typically operationalised in regards to each.

### **The texts concordanced**

Another potentially significant variable in learner concordance use is the type of text concordanced (see Table 3). There are many possible ways that corpora can be classified but, as we have seen, some commonly appealed to categories with theoretical implications are large general corpora, small focused corpora, or locally compiled collections of texts and, within learner use of concordances specifically, perhaps also corpora of graded readers. Large general corpora are the most defensible means of providing learners with a basis for making broad generalisations about a language in general, but they are also likely to present the most heterogeneous concordances, and this might reflect a higher authentication burden than concordances of other available texts. It has been claimed that smaller, more narrowly focused corpora or learner made corpora may be easier to authenticate (Aston, 2001; Braun, 2005; Gavioli, 2005; Gavioli & Aston, 2001), and though such corpora might not always afford robust generalisability affordances, they would still seem to be offering many potentially valuable language learning affordances, especially in the context of language for specific purposes. In contrast, while concordances of collections of graded texts might present a lower authentication burden (any authentication burden from text fragmentation aside), the affordances of such concordances would seem to be limited to the affordances of concordances as concentrated text.

The basic distinction between large general corpora and small focused corpora is explored in Chapters 3 and 4. In Chapter 5, concordances from three medium sized corpora are compared with each other, as well as with concordances extracted from a corpus of graded texts. The studies reported in Chapter 5 highlight the significance of corpus concordanced as a variable in learner concordance use from the perspective of authentication burden. In Chapter 6, the homogeneity-heterogeneity issue is addressed, with concordances from a large general corpus being compared with concordances from a medium sized corpus and a small, learner-compiled-



type corpus, again highlighting the importance of texts concordanced as a variable in learner use of concordances.

*Table 3 Texts concordanced*

Variable	Main levels	Key points
Texts concordanced	Large, general	Potentially generalisable to the general variety of a language. Likely to produce large and varied concordances (for frequent words at least), and so may be difficult for learners as a consequence
	Small, narrow	Likely to be less generalisable. Likely to produce smaller, more homogenous concordances. Potentially easier to authenticate as a consequence
	Learner made	Questionable generalizability. Possibly easiest to authenticate. Likely to require of learners more technical skill than many less hands-on approaches to concordance use.
	Graded materials	Concordances of graded materials afford easy access to a linguistically concentrated text while controlling for the linguistic difficulty of citations. However, they would appear to have limited suitability in relation to many of the purported affordances of concordances in general.

### **Which citations are presented**

Recalling that concordances are generated via exhaustively indexing a text, there are several important issues in learner use of concordance pertaining to which citations are presented: chiefly though, how many citations are shown, which citations are shown, and the resulting complexity of the concordance obtained. Although a potentially significant variable, it is not however a variable that is explored further in this thesis. If we think about the three main levels of this variable (see Table 4 below), it seems random sampling has the most relevance to direct learner use of concordances. It does not seem exhaustive processing of large concordances is

recommended by anyone working in the field; warnings about exposing learners to too much data are common place. However, it has been argued here that manual selection of citations both limits the affordances of concordances to concentrated text affordances, and significantly blurs the distinction between learner use of concordances and teacher use of concordances in materials design. Whether or not input texts inspired by the incidental qualities of concordances as texts have an important role to play in language learning is an interesting topic and well worth researching, but it is not the topic of this thesis. In terms of the sort of direct use required to make ad hoc use of texts as language learning resources, random sampling can control for the quantity of data presented, if not the quality. Research on automated selection of citations is discussed in Chapter 5, but it would seem many of the most important affordances of concordances for learners cannot be mediated by pedagogical selection of concordances, at least not in the sense of *learner use* intended here. Consequently, the studies reported in Chapters 5 and 6 analyse concordance extracts based on random samples from the corpora used.

*Table 4 Citations shown*

Variable	Main levels	Key points
Citations shown	Exhaustive	Basis for robust generalisation. All citations are available, even if not all shown simultaneously. Likely to include a proportion of erroneous or otherwise difficult citations.
	Randomly sampled	Provides control over number of citations shown but is still a basis for robust generalisation if of sufficient size. Likely to include a proportion of erroneous or otherwise difficult citations. Additional citations available in context of hands-on concordance use.
	Pedagogically selected	Potentially labour intensive. Blurs distinction between learner use of concordances and materials writing. Allows for maximal control over difficulty and more convergent learning outcomes

### **The format citations are presented in**

In discussion of the potential for concordances to afford learners opportunities to literally observe patterns of co-occurrence, it seemed clear that the format and order that citations are

shown in might be a significant variable in learner use of concordances, in regards to this affordance at least. There are potentially many subtle distinctions that could be made between different options for displaying concordances, but there appear to be two main categories (see Table 5): the key word in context (KWIC) format and conventional sentence style formatting. As suggested in the table below, the former may make patterns of co-occurrence more observable, but also make the citations harder to authenticate; the latter, vice versa. Format is explored as a variable in concordancing in Chapters 3 and 4.

*Table 5 Format citations are presented in*

Variable	Main levels	Key points
Format shown in	Key word in context (KWIC)	Associated with observability of patterns of co-occurrence and sorting of concordance lines. Short spans of cotext may increase authentication load, but facilitate pattern observation. Additional cotext is typically available on request in hands-on use.
	Sentence format	Likely to be easier to authenticate due to familiar text format and presenting syntactically complete fragments of text. Does not represent co-occurrences spatially, though may still increase their noticeability to some extent via proximity.

### **Dynamic versus static access**

I have framed this variable slightly differently to the more conventional distinctions of *hands-off* and *hands-off, hard* and *soft*, prevalent in the literature. The terms I use are intended to highlight the significance of access to the indexing functionality of concordancing. It might also be noted that online concordancing software aimed at learners often presents relatively static concordances, albeit the concordance is accessed through a web page, blurring the distinction between hands-on and hands-off use to some extent. However formulated, this is undoubtedly an important variable in learner use of concordances; however, it is not pursued further here.

*Table 6 Dynamic or static access*

Variable	Main levels	Key points
Dynamic or static access	Dynamic	May involve learner control over the search term, the texts concordanced, the citations shown or the format the citations are shown in
	Static	The learner may not have control over the search term, the texts concordanced, the citations shown or the format the citations are shown in

## Overview

The theoretical perspectives on learner use of concordances discussed in this chapter highlight the range of affordances offered by concordances, several important variables in how concordancing is operationalised, and their inter-relation with the underlying variable of the authentication burden of concordances. The next four chapters report studies that engage with the variables identified through empirical research.

The next chapter seeks evidence of concordancing as a complex, multi-faceted construct, reflecting the multiple affordances that concordances are purported to afford and the potential for these affordances to interact with variables in how concordancing is operationalised. It explores this issue via a survey of concordance users' concordancing preferences. On the assumption that systematic differences would not be arbitrary but would in fact be motivated, observing systematic differences in respondents' concordancing preferences would provide a clear indication that the construct of concordancing is indeed complex and multifaceted; it also helps to identify psychologically salient variables in concordance use from the perspective of concordance users (Chapter 3). Indeed, the literature review suggests that the affordances and constraints of concordance use are likely to differ in relation to how concordancing is operationalised and, therefore, users seeking particular affordances or attempting to overcome particular constraints are likely to be using concordances in characteristic ways. The thesis

explores this issue by looking for correlations between various categorisations of the concordance users surveyed and their concordancing preferences (Chapter 4).

The second major issue raised in the literature review is the perennial issue of the difficulty that learners may face in authenticating concordances: a concordances *authentication burden* (see *Pedagogical mediation and the authentication burden of concordances*, pp. 40-47). Chapters 5 and 6 report studies that begin to address this issue. Chapter 5 analyses the authentication burden of concordances in terms of word frequency measures of vocabulary load, using corpus genre as an experimental variable; Chapter 6 examines repetition and variation in concordances, exploring corpus homogeneity/heterogeneity as a variable.

Please note, because the studies reported in this thesis are methodologically diverse, there is no methodology chapter in this thesis. Instead, each chapter reporting a study presents a separate methodology section of its own, along with a rationale for the approach adopted in that particular study.



### Chapter 3. Analysing concordancing: A simple or multifaceted construct?<sup>1</sup>

#### Introduction

A review of relevant literature indicates that concordances are the primary focus of papers addressing data-driven learning (Boulton, 2010b; Cobb & Boulton, 2015), but that the language learning conditions described as *concordancing* can differ in a number of potentially significant ways (see Tables 1 to 6 presented above in *Variables in learner use of concordances*, pp. 63-71). This suggests that, within an affordances framework (Levy & Stockwell, 2006), it may not be appropriate to view concordancing as a simple, unidimensional construct in which concordancing offers equivalent opportunities for learning irrespective of the details of how learner concordance use is realised; rather, concordancing may actually be a coverall term, a term which can mask potentially important differences between a range of concordancing technologies that offer a range of language learning affordances, each to different extents. The latter view of concordances implies that considerable care needs to be taken in designing, interpreting and reporting on learner use of concordances. This chapter reports an online survey that was used to explore the construct of concordancing from the perspective of concordance users and their concordancing preferences.

The survey targets as many concordance users as could be contacted for a number of reasons. First, the underlying hypothesis behind the study is that concordances can be used in a variety of ways: if only one type of concordance user were surveyed (for example, language learners), then the chances of identifying multiple orientations to concordancing would be reduced. Second, although Tribble (2013) does provide statistics on teacher use of concordances, there is no equivalent data on other groups of concordance users such as researchers, translators and language learners. Consequently, this survey also provided an opportunity to find out just who concordance users are. Indeed, the results of the survey suggest that some groups (such as language learners) genuinely are minority user groups at present. Third, if the survey were limited to only collecting responses from language learners, not only would this have resulted in

---

<sup>1</sup> A paper based on this chapter has been published in the journal *Computer Assisted Language Learning*, currently available through early online access as: Ballance, O. J. (2016). Analysing concordancing: a simple or multifaceted construct? *Computer Assisted Language Learning*. DOI:10.1080/09588221.2016.1209527

a very small sample, it would have ignored an important aspect of the rationale behind the thesis: the possibility that the limited uptake of concordancing amongst language learners may be explained by learners being introduced to operationalisations of concordancing that are suboptimal from a language learning perspective. In sum, to get a full picture of whether or not concordancing is a simple, monolithic construct or whether it is a complex, multifaceted construct that may offer users variable and competing affordances, it is desirable to survey as full a range of concordance users as possible.

Using an online survey allowed data to be collected on a wide and varied sample of concordance users, and thus facilitated relatively generalisable empirical investigation of the construct. The survey asked questions addressed to five potentially significant variables in concordance use and a series of factor analyses were used to provide a psychometrically robust account of the dimensions of variance present in the responses collected: evidence of systematic differences in concordancing preferences constitutes good evidence for there being different operationalisations of concordancing, concordancing as a multi-dimensional construct, and not just more or less prototypical ways of using a concordance, concordancing as a simple construct. In addition, the dimensions of variance identified represent psychologically real differences in the operationalisation of concordancing, and therefore also represent potentially significant variables in learner use of concordances.

This chapter first describes the variables that were explored, and then the research questions addressed. It then provides details of the methodology employed, followed by the results obtained. The chapter concludes with a discussion of the results and the limitations of the study.

### **Variables in concordance use with theoretical significance**

As the literature review indicated, there are of course more than 5 variables in concordance use that could have been explored. However, a psychometrically credible survey must use multi-item scales to ensure the reliability of survey responses and this in turn limits the number of questions that can be presented in a survey: if a survey attempts to examine too many variables through multi-item scales, it is likely to become too long and negatively affect the respondents' quality of response and willingness to participate (Dörnyei & Taguchi, 2010; Punch, 2003). This practical limitation meant that the variables to be explored had to be carefully chosen. The criterion used



balanced a number of relevant factors: first, some variables are easier to ask about in a self-report questionnaire than others (piloting showed that some potential variables, such as linguistic difficulty level or search term used, were perhaps too abstract to be included in a self-completed questionnaire); second, variables that have been operationalised differently in published studies of learner concordance use were targeted; third, variables that are likely to have pedagogical significance were preferred over theoretically uninteresting variables; fourth, given that a quasi-experimental questionnaire aims to detect variance in the sample surveyed (Punch, 2003), variables that were amenable to conceptualisation in terms of mutually exclusive operationalisations were preferred. This led to development of a survey that sought to gather data on five hypothetical dimensions of concordancing: the corpus used, two aspects of the format citations are displayed in (alignment and size), the order citations are displayed in and the concordance-based language learning task engaged in. Each is discussed below.

### **Variable: The corpus used**

One way in which two concordances can offer divergent affordances is the corpora that they are generated from (see Chapter 2): in combination with the search term used, the choice of corpus is a major determinant of the language contained in the concordance. This is important because studies have used a wide range of corpora, with decisions as to which corpus to use in a given context generally being made in relation to the learners' proficiency level or language learning goals. For instance, some studies use large general corpora, such as the BNC (Boulton, 2010a; Daskalovska, 2013; Gordani, 2013; Kaur & Hegelheimer, 2005). A concordance of the BNC affords the learner examples drawn from a wide range of genres and domains and, because the BNC was designed to be representative of British English in general, there is some justification for interpreting a phenomenon well attested in the BNC as being typical of British English in general. However, as a large general corpus, it is also likely to present a relatively heterogeneous group of citations, and so it might also present a greater authentication burden for a learner (Aston, 2001; Gavioli & Aston, 2001). In contrast, a number of studies have used corpora composed of materials explicitly targeted at language learners (Allan, 2010; Cobb, 1997, 1999). The idea here is that concordances of materials designed for learners will present less vocabulary burden, and as such, will be easier for learners to use (Allan, 2009, 2010; Aston, 2001; Boulton, 2009b; Cobb, 1999; Flowerdew, 2012b; Römer, 2011). Several studies from Taiwan based

researchers have drawn upon a parallel English-Chinese newspaper corpus, Sinorama (Chan & Liou, 2005; Lai & Chen, 2013). Arguably, a corpus of this type affords learners a more readily accessible context for the data in the concordance, as well as restricting data to a situationally appropriate variety of English. For these kinds of reasons, Boulton (2009b) decided to present his learners with concordances of contemporary British newspapers. The point here is that, while each type of corpus would appear to afford learners some benefit, these affordances are divergent: for instance, the affordances of a large general corpus are different to and essentially incompatible with those of the smaller, more focused corpora. Indeed, the distinction between large, general corpora and small, focused corpora is very common in the literature (Aston, 2001; Braun, 2005; Gavioli, 2001; Gavioli & Aston, 2001; Park, 2012; Pérez-Paredes et al., 2012; Römer, 2011), and proved to be readily understood by informants during survey piloting.

#### **Variable: The format of the citation**

A second potentially important dimension within the construct of concordance use is how the data is displayed (Boulton, 2009b; Flowerdew, 2015). Sinclair (1991) states that the *key word in context* (KWIC) format has been the favoured way of displaying concordances from the beginning of their use in linguistic research, and as Chapter 2 makes clear, the KWIC format of concordance is strongly associated with observing patterns of co-occurrence within the cotext of the search term. From the KWIC format of concordances we can derive the notion of the cotext presented in a concordance being *spatially meaningful*: the KWIC format can allow patterns of co-occurrence to be represented spatially, making them visually salient. This is a slightly formal account of the general comment that KWIC concordances make co-occurrence patterns more observable (Aston, 2001; Boulton, 2009b; Braun, 2005, 2006; Charles, 2012; Daskalovska, 2013; Flowerdew, 2012b; Gavioli, 2001; Götz & Mukherjee, 2006; Johns, 1991b; Kaur & Hegelheimer, 2005; Levy, 1990; Mishan, 2004; Römer, 2010, 2011; Sinclair, 2004; Sun & Wang, 2003; Tono et al., 2014; H. Yoon & Hirvela, 2004). However, as discussed in Chapter 2, KWIC is also often criticised for putting a higher authentication burden on learners than longer, syntactically complete citations (Aston, 2001; Boulton, 2009b; Cobb, 1999; Johns, 1986; Levy, 1990; Mishan, 2004; Stevens, 1991a; H. Yoon & Hirvela, 2004). For this reason, sentence format citations may be preferable, for novice concordance users at least (Tribble, 2010). At the same time though, it should be recognised that contextual co-occurrences in sentence format citations

are likely to be *spatially arbitrary* if the sentences are not aligned to a central column of the screen/page or spanning more than a single line each. Thus, the affordances of the readable but spatially arbitrary sentence format citations and the less readable but spatially meaningful KWIC format citations would appear to be in conflict.

Many studies of learner use of concordances do appear to have employed KWIC format concordances (Allan, 2010; Boulton, 2010a; Cobb, 1997, 1999; Daskalovska, 2013; Kaur & Hegelheimer, 2005). However, Gordani (2013) reports students using the concordancer available at <http://www.natcorp.ox.ac.uk/>, which, when used in *simple search* mode, presents citations aligned to the left-margin of the page, and in full-sentences, as does the TOTALrecall concordancer, and the concordancer integrated within the TANGO collocation finder (Chan & Liou, 2005; Lai & Chen, 2013). To complicate matters further, when we look at the static concordances available at <http://www.just-the-word.com>, we find that citations are presented with the search word aligned to in a single column in the centre of the screen, but many of the citations are complete sentences or else very long, blurring the distinction between KWIC format citations and sentence format citations. If we consider the concordance display in the professional software WordSmith Tools (Scott, 2016), citations are again aligned to a central column by the key word, but the user can increase the span of cotext available by either dragging the column width of the citation column or by adding in additional lines of cotext. Thus, the apparently simple distinction between sentence format citations and KWIC format citations is more complicated than it first appears. For this reason, the format of the citation was operationalised as two separate citation display variables: *alignment* and *size*.

### ***Variable: Alignment***

The alignment variable concerns how each citation is placed in relation to the others. There are two incompatible alignment types that are commonly used in the display of citations: central alignment and left aligned. In central alignment, the search term of each citation forms a column in the text. Central alignment is the core feature of KWIC formatting. In left alignment, each citation is positioned so that it starts from the left margin of the screen or page without reference to the position of the search term: left alignment is typical of sentence format citations, though sentence format citations could also be centrally aligned in theory. Consequently, the variable

alignment was conceptualised as being separable from size, and consisted of two mutually exclusive categories: centrally aligned and left aligned.

***Variable: Size***

The size variable concerns how much cotext is presented: the span of cotext presented. There are several ways that citation size can be conceived of: for example, how many characters are presented, how many words are presented, how many pixels or millimetres the citations spans. As discussed previously, in terms of KWIC formatting, the size of citations is usually set to however much cotext can be displayed on a single line, by default at least: as noted above, this is to preserve the visibility of spatially meaningful relationships between citations. Of course, sentence format citations vary in length; however, typically, a sentence format citation will exceed a single line of text. Due to the wide range of ways citation size can be conceived of, this variable was approached in terms of a preference for either more or less cotext being presented.

**Variable: The order of citations**

As well as how the citations themselves are formatted, concordances may also differ in terms of how the citations are ordered. Unfortunately, this aspect of concordancing is rarely reported in published research, although some indication of how citations are ordered is often evident from screen shots or examples of experimental materials. However, it is not without theoretical significance, as discussed in Chapter 2. Being able to sort citations alphabetically by positions in the cotext is thought to powerfully enhance the observability of patterns of co-occurrence (Aston, 2001; Charles, 2012; Hunston, 2010; Johns, 1986, 1991b; Levy, 1990; Partington, 2001; Sinclair, 1991, 2004; Tono et al., 2014; Tribble, 2010). A further indication of the value of being able to sort citations by cotext is provided by Breyer (2011) in her design for a language learning concordancer, My Concordancer. Despite wishing to present the simplest, most intuitive interface possible and dispensing with as much research orientated functionality as possible, My Concordancer retains a basic alphabetical sort function. Clearly, Breyer sees cotext-based sorting as indispensable.

Ordering citations by cotext is, however, not the only possible way of ordering citations. Johns (2002) suggests ordering citations manually so as to facilitate the process of hypothesis formation and testing, a key tenet of his theory of data-driven learning. Looking at the design of

concordancers, it would seem that another significant method of ordering citations is to display them in the order in which the concordancer finds them in the corpus: i.e. in text order. Although this has not been discussed in the literature as far as I am aware, this has possible implications for language learning in that citations that occurred near to each other in the corpus may have more intra- or inter-textual coherence and thereby reduce the authentication burden of the citations presented (cf. Aston, 2001; Braun, 2005). More generally, presenting citations derived from the same text or section of a corpus contiguously presumably facilitates comparisons of citations across authors or sub-corpora.

In sum, although this aspect of concordancing has so far received little research interest, it seems clear that there are a number of distinct approaches to ordering citations available, and that different ways of ordering citations are likely to present divergent affordances. The survey approached this variable in terms of a distinction between cotext ordering of citations and source text ordering of citations. While these two options are mutually exclusive, it should be acknowledged that alternative ordering schemes are possible, such as random ordering and manual ordering.

### **Variable: The concordance-based language learning task**

One final aspect of learner concordance use that seems likely to present divergent affordances is the concordance-based task that the learner engages in. There is a tendency in the literature for studies to frame the variable of interest in terms of the resource provided (the concordance) at the expense of research that explores the potential for an interaction effect between the language learning resource and the language learning task. Indeed, there have been calls for research to directly address the extent to which the benefits of learner concordance use are mediated by what it is that learners are asked to do with a concordance (Aston, 2001; Franken, 2014; Gavioli & Aston, 2001).

One type of task commonly used in studies is the classic data-driven learning task (Allan, 2010; Boulton, 2010b; Daskalovska, 2013; Gordani, 2013). In this kind of task, learners are presented with a concordance accompanied by a set of questions that aim to guide the learner through a series of research-like tasks: identifying some feature of the language in the concordance, classifying the language identified, and then forming a generalisation on this basis (Johns, 1991b; Thurston & Candlin, 1998). This kind of task requires the learner to focus on

information that is repeated in multiple-citations: a process of comparing and contrasting the citations within the concordance. In contrast, Boulton (2009b), Cobb (1997), Chan and Liou (2005) and Kaur and Hegelheimer (2005), present concordances alongside gap-fill tasks. Here, the learner's attention is focused on a context outside of the concordance (the sentence to be gap-filled), and they may approach the task by comparing individual citations to this concordance-external context. Importantly, they are not required to simultaneously compare and contrast multiple citations, and a single suitably informative citation may be sufficient to inform their answer. Chan and Liou (2005) report that their treatments also included sentence translation exercises, while Kaur and Hegelheimer (2005) also asked learners to use concordances in what they term sentence building exercises. Again, these tasks do not require comparison of citations and may be tackled on the basis of individual citations that are felt to be particularly informative. In fact, this distinction would seem to underlie the research/research-like use of concordances and reference resource type use of concordances discussed in Chapter 2. Hence, this variable was approached in terms of a preference for looking for information repeated in the concordance (engagement with the concordance based on comparing and contrasting citations) or looking for relevance to an outside context of use (engagement with a concordance on the basis of comparing individual citations with an external context, such as an essay), although it must be acknowledged that these two broad categories of usage may in actual fact be interwoven in practice.

### **Research questions**

Research on learner use of concordances has used concordances that have been generated, displayed and used in diverse ways, and there is reason to think that these differences may correspond to divergent language-learning affordances, indicating that *concordancing* may best be thought of as a catch all term for a range of distinguishable language learning technologies. This stands in stark contrast to the majority of research which has tended to discuss concordancing as a unidimensional construct. Consequently, the research questions addressed by this chapter are:

1. Does concordancing appear to be a unidimensional or multidimensional construct?

2. If it is the case that concordancing appears to be a multi-dimensional construct, what are the factors that underlie the construct of concordancing?

## Methodology

The design of this study is based on the premise that, if different types of concordancing offer genuinely divergent affordances, then this should be observable in concordance users' preferences: that divergent concordancing preferences are an indicator of divergent affordances. Furthermore, as psychometric theory provides appropriate techniques for exploring abstract mental variables such as preferences, and these techniques can be utilized within a questionnaire, an online survey of concordance users represents the most effective way of collecting a large enough sample of concordance users for psychometric analysis of the factors underlying the data collected.

### Conceptual framework of the survey

This paper reports on the sections of a survey conducted in 2015 that collected data on users' concordancing preferences. Punch (2003) emphasizes the importance of a quantitative survey being designed to tap into variability. Thus, on the basis of the literature review, and to this end, the survey hypothesized five dimensions of concordancing preference that could be treated as if they were dichotomous: corpus type preference; citation alignment preference; citation size preference; citation order preference; concordance task preference (Table 7).

### The survey instrument

A key consideration in designing a reliable survey instrument is the concept of multi-item scales (Dörnyei & Taguchi, 2010; Punch, 2003). Differences in how people interpret questions makes individual responses to survey questions inherently unreliable. For example, changing the wording of a question from *should ban* to *should not allow* can have a significant effect on how some respondents answer the question, despite the two versions of the question being propositionally equivalent (Dörnyei & Taguchi, 2010). Fortunately, the use of multi-item scales

*Table 7 Hypothetical dimensions of concordancing conceptualised as dichotomies*

Dimension	Dichotomy
Corpus type preference	Small, focused corpora
	Large, general corpora
Citation alignment preference	Aligned by citation to the left-margin
	Aligned by key word to the centre
Citation size preference	Large spans of cotext
	Small spans of cotext
Citation order preference	Ordered by the texts citations are extracted from
	Ordered by the cotext contained in the citation
Task preference	Read to locate relevant examples
	Read for relationships between citations

is able to counter-act this kind of problem to some extent: by asking several questions that each address the same variable and aggregating the scores, measurement error introduced by idiosyncratic interpretations of one or two questions is minimised. With this principle in mind, the survey instrument was piloted extensively with a range of corpus users, as well as with a number of experts in corpus linguistics.

On the basis of piloting, the survey's questions, response scales and presentation were revised until informants appeared able to complete the survey confidently and consistently. The final set of survey questions consisted of 36 statements that targeted five hypothetical dimensions of concordancing and were answered on a single five-point agreement scale, along with a sixth response category, *I do not understand the question*. One particularly important outcome of the piloting process was recognition of the fact that many concordance users use concordances differently depending on their field of use: for instance, using concordances for translation or using concordances for teaching. The field of concordance use envisioned by the respondent could have a strong bearing on how they rated the statements leading to hesitation and inconsistent responses. Consequently, respondents were asked to choose one of several fields of concordance use that they would think of while rating the statements; this appeared to facilitate more confident and reliable completion of the survey. Screenshots of the survey are presented in Appendix 1, and an outline of the survey instrument is presented in Table 8 below.



Table 8 Survey outline

Section	Topic	Question types
Introduction	Information and ethics sheet	<i>Text display, but with a tick box to indicate consent</i>
	Global instructions to respondents	<i>Text display only</i>
Section 1: Use	User experience	<i>Multiple choice question – single answer permitted</i>
	Frequency of use	<i>Multiple choice question – single answer permitted</i>
	Linguistic relation to concordance: yes, occasionally, no	<i>Matrix question – multiple answer permitted</i>
	Field of use: yes, occasionally, no	<i>Matrix question – multiple answers permitted</i>
	Type of teaching	<i>Multiple choice question – multiple answers permitted – question only displayed if respondent indicated that they used concordances for teaching above</i>
	Direct or indirect use	<i>Multiple choice question – multiple answers permitted – question only displayed if respondent indicated that they used concordances for teaching above</i>
	Field of use: relative time estimations	<i>Constant sum question – response categories available determined by answers to above</i>
	Field of use: referred to in remainder of survey	<i>Multiple choice question – single answer permitted – response categories determined by answers to above</i>
	Linguistic relation to concordance: referred to in remainder of survey	<i>Multiple choice question – single answer permitted</i>
	Aspect of language explored: yes, occasionally, no	<i>Multiple choice question – multiple answers permitted</i>
	Aspect of language explored: relative time estimations	<i>Constant sum question - response categories available determined by answers to above</i>
Task	Syntagmatic or paradigmatic reading	<i>10 statements responded to on a 5 point scale of agreement</i>
Section 2: Corpus used	Corpus type	<i>7 statements responded to on a 5 point scale of agreement</i>
	Definitions of terms	<i>Text display only</i>
Section 3: Alignment	Citation alignment	<i>6 statements responded to on a 5 point scale of agreement</i>

Section 4: Size	Citation size	<i>7 statements responded to on a 5 point scale of agreement</i>
Section 5: Citation ordering	Citation ordering	<i>6 statements responded to on a 5 point scale of agreement</i>
Section 6: Biographical data	Ethnicity	<i>Text entry</i>
	1st language	<i>Text entry</i>
	Profession/subject area	<i>Text entry</i>
	Contact details	<i>Text entry</i>

Note, in Section 1, respondents were asked to think of a single field of use (teaching, translating, linguistics, and so forth) and linguistic relationship to the concordance (their first language, a second language or a bilingual concordancer), before providing any information pertaining to the five variables discussed in this study: *task*, *corpus*, *alignment*, *size* and *order*. Please also note that this same survey instrument was used to collect data for the follow-up study reported in Chapter 4.

### **Data collection and participants**

Before gathering data, the project was reviewed and approved by the research ethics board at Victoria University of Wellington (see *Appendix 1: Screenshots of the questionnaire* for a screenshot of the ethics documentation). The survey was then distributed and completed online. It was not possible to employ a formal sampling scheme for the survey because it was not possible to delineate the target population, *concordance users*, a priori. In such circumstances, it is appropriate to use snowball sampling (Dörnyei & Taguchi, 2010): that is, to disseminate the survey to unknown concordance users through known concordance users. This was achieved through word of mouth, emails, Internet bulletin boards and social media. The survey obtained 181 usable responses. In Section 6 of the survey, biographical data was collected on the respondents' ethnicity, first language and profession or subject area of study. In brief, the data collected indicates that the respondents were from many different regions of the world, and typically associated with higher education through either their profession (EAP teachers, lecturers and researchers) or study (postgraduate students). Thus, the respondents to this survey would appear to be similar to the teachers surveyed in a 2012 survey of teaching and language

corpora, which also found the majority of respondents to be working in higher education (Tribble, 2015). More detailed summaries of the biographical data obtained are presented below.

### ***Respondent ethnicity***

The majority of participants indicated their national or ethnic background in geographical terms. Frequency counts of their answers are tabularised below (Table 9). The category of *other* includes responses from people who did not indicate a geographical identity (e.g. white or Caucasian) as well as respondents who indicated nationalities that are difficult to place in standard continent categories (i.e. Russian and Turkish). As can be seen, the largest proportion of respondents appear to have been European, and it is also perhaps noteworthy that none of the respondents identified themselves as African. Nevertheless, the sample obtained does include responses from concordance users from many regions of the world.

*Table 9 Respondents' backgrounds*

Background	<i>n</i>	Percent	Cumulative Percent
Europe	81	44.8	44.8
North America	27	14.9	59.7
Oceania	15	8.3	68.0
Asia	14	7.7	75.7
South America	11	6.1	81.8
Other <sup>a</sup>	17	9.4	91.2
Not given	16	8.8	100.0

<sup>a</sup> including ethnic identifiers such as white (4), Caucasian (2), Hispanic (2) and Latin (1), as well as Eurasian nationalities: Russian (2) and Turkish (2)

### ***Respondent first language***

Table 10 shows frequency counts of how respondents identified their first language or languages. A large proportion of the respondents indicated that English was their first language (77 out of 181 possible responses accounting for 42.5%), with the other answers consisting of a wide variety of languages (32 other languages or language varieties accounting for 33.7% of

Table 10 Respondents' 1st language/s

Rank	Language/s	<i>n</i>	Percent	Cumulative
				Percent
1	English	77	42.5	42.5
2	Italian	17	9.4	51.9
3.5	German	8	4.4	56.3
3.5	Spanish	8	4.4	60.7
5	Chinese	6	3.3	64.0
6	Portuguese	5	2.8	66.8
7	Greek	3	1.7	68.5
8	Brazilian-Portuguese; Dutch; Japanese; Polish; Russian; Swedish; Turkish	2 (14) <sup>a</sup>	7.7	76.2
9.5 (monolingual)	Arabic; Bulgarian; Cantonese; Croatian; Czech; Danish; Finnish; French; German-Swiss; Greek; Hebrew; Korean; Lithuanian; Macedonian; Norwegian; Persian; Serbo-Croatian; Singaporean English; Uzbek	1 (19) <sup>a</sup>	10.5	86.7
9.5 (bilingual)	Dutch & German; English & French; German & English; Italian & German; Russian & Italian; Slovene & Hungarian; Spanish & Catalan; Spanish & English; Spanish & Galician; Welsh & English	1 (10) <sup>a</sup>	5.5	92.2
Not given		14	7.7	99.9

<sup>a</sup> Number in brackets indicates the total frequency of responses from all the languages listed within the category

responses) and answers indicating bilingualism (10 responses accounting for 5.5%). The bias towards native English language speakers may well reflect the fact that the survey was only available in English.

### ***Respondent profession or subject area***

The respondents' answers to the question about their field or work, research or study were, as one might imagine, diverse. Using WordSmith Tools (Scott, 2016), responses were analysed by identifying content words that occurred in the responses 10 times or more and then identifying content-word collocates that occurred 5 times or more (Table 11). The table reveals a strong tendency for the respondents to be connected to academic life, either as researchers, educators or students (particularly postgraduate students). For example, the most frequent content word used, *language* (48), collocated with *PhD* (18), *linguistics* (16), *lecturer* (10), *researcher* (8), *professor* (7), *university* (5), *MA* (5) and *degree* (5). Furthermore, the majority of references to teaching appear to be within higher education contexts, and frequently references to translation are made in connection with graduate study. This is in keeping with earlier research that has found corpus use to be mainly within the academy (Tribble, 2015). It is, however, not possible to discern whether this is because concordances are rarely used in other contexts outside of the academy, or simply because the survey was unable to obtain responses from a wider concordance using community, if one exists.

*Table 11 Content words occurring 10 or more times and their collocates*

Rank	Word	<i>n</i>	Collocates
1	Language	48	teacher (54), PhD (18), linguistics (16), English (14), lecturer (10), teaching (9), researcher (8), professor (7), student (6), corpus (6), applied (6), university (5), working (5), MA (5), degree (5)
2	Teacher	45	language (54), university (15), EAP (14), PhD (13), linguistics (13), researcher (13), English (12), degree (6), applied (6), trainer (6), lecturer (5), MA (5), student (5)
3	Linguistics	44	researcher (19), applied (19), professor (17), corpus (17), language (16), lecturer (13), teacher (13), PhD (12), working

			(9), student (8), English (7), teaching (13), university (6), EAP (5), retired (5), research (5)
4	Researcher	31	linguistics (19), working (17), teacher (13), lecturer (11), language (8), corpus (8), university (7), professor (6), applied (5), field (5), English (5)
5	English	29	language (14), teacher (12), student (7), teaching (7), linguistics (7), professor (7), lecturer (6), university (5), researcher (5)
6	Student	28	PhD (24), translation (12), graduate (10), postgraduate (8), linguistics (8), English (7), language (6), translator (5), studying (5), teacher (5)
7	Professor	24	linguistics (17), university (9), language (7), English (7), applied (6), researcher (6), teaching (5), linguist (5)
8	PhD	22	student (24), language (18), teacher (13), linguistics (12)
9	Translation	21	student (12), postgraduate (9), working (5), studies (5), specialised (5)
10	Working	21	researcher (17), linguistics (9), translation (5), university (5), field (5), language (5), studies (5)
11	Corpus	20	linguistics (17), researcher (8), language (6), lecturer (6)
12	University	20	teacher (15), professor (9), lecturer (7), researcher (7), linguistics (6), working (5), assistant (5), applied (5), language (5), English (5)
13	Lecturer	18	linguistics (13), researcher (11), language (10), university (7), corpus (6), English (6), teacher (5)
14	Teaching	16	language (9), English (7), linguistics (6), professor (5), academic (5)
15	Applied	15	linguistics (19), professor (6), teacher (6), language (6), university (5), researcher (5)
16	EAP	12	teacher (14), tutor (8), linguistics (5)
17	Graduate	12	student (10)

---

Although it might have been preferable to collect fixed response data, it was not possible to confidently predict who the respondents would end up being and, therefore, which response categories would have been appropriate. As a consequence of using open response questions, the

data obtained cannot be neatly categorised in terms of, say teachers versus lecturers, not least because many lecturers might also identify themselves as teachers (e.g. *I am a lecturer who teaches translation studies*). However, arguably, the table above provides a richer summary of how respondents identified their profession or subject area than data obtained from tick boxes. For instance, it seems clear that a large number of respondents would have ticked a box labelled teacher, but without further elaboration, we would have no indication of the level taught at or the subjects taught. Table 11 provides a rich account of the respondents' answers and shows that respondents were typically connected to academic life through either research, teaching or study connected to language, linguistics or translation.

### **Data analysis**

Having collected data on the 36 survey items intended to tap the five hypothetical dimensions of concordancing discussed above, factor analysis was used to identify the underlying structure of the data. Factor analysis is a powerful statistical technique, but prone to misuse (Matsunaga, 2010; Plonsky & Gonulal, 2015). The approach to factor analysis taken in this study follows guidance in Matsunaga (2010). He suggests three stages of factor analysis, each performed on an independent data sample, if a sufficient quantity of data is available. The total number of responses required per independent analysis is a function of the number of items entered into the analysis, the number of factors obtained by an analysis and the strength of the loadings of items onto factors. In an ideal situation, first, principal component analysis (PCA) would be used to reduce the item pool and identify a preliminary factor structure; second, principal axis factoring (PAF) would be used to confirm the underlying factor structure of the data; and third, a confirmatory factor analysis (CFA) would then rigorously test the model obtained by the PAF. However, in the present study, the total number of responses obtained was only sufficient for two independent analyses, a PCA and a PAF. Fortunately, in lieu of a CFA, convergent factor solutions from a PCA and PAF also provide strong evidence for the factor solution obtained (MacCallum, Widaman, Zhang, & Hong, 1999; Matsunaga, 2010).

To determine the appropriate eigenvalues for factor extraction, SPSS syntax was used to conduct parallel analyses of the data: the rawpar syntax available from Brian O'Connor's website, an earlier version of which is reported on in O'Connor (2000). Rawpar generates

random permutations of a data set and thereby determines the eigenvalues that would be expected to emerge from the data by chance: the eigenvalue that is the smallest value in the original data set that exceeds the equivalent value at the same root level in the random permutations is the appropriate eigenvalue to use. It can be considered an even more reliable approach to parallel analysis than analyses based on invented data and as being able to produce much more reliable eigenvalues than arbitrary eigenvalues such as one (Matsunaga, 2010; O'Connor, 2000).

Most factors in applied linguistics research can be assumed to correlate to some extent (Plonsky & Gonulal, 2015). Consequently, Promax rotation ( $k=4$ ) was used to determine factor solutions in this study as it allows oblique solutions, meaning that it does not minimise correlations amongst factors artificially (Matsunaga, 2010). The relatively stringent .6/.3 rule was used for determining which items to retain in the PCA: items were retained if their primary loading on a factor was above .6, and they did not have any secondary loadings above .3. Matsunaga (2010) describes this as constituting something of a norm within social scientific studies. Having reduced the data set, the more conventional .4 cut-off point was used for PAF. Finally, because the measurement instrument collected interval data via response scales and, as such, was not truly continuous data, it was necessary to check that the factor solution was not the result of the measurement instrument. This was done by inspecting means and standard deviations, following guidelines from (Bernstein, Garbin, & Teng, 1988).

Before analysing the data, scores for each item were inspected for missing data. No individual item was missing more than 20% of the data, and no individual respondent was missing data on more than 20% of the items, so mean scores were calculated for missing data points. Before beginning the factor analyses, the data set was randomly divided into two halves, the first consisting of ninety-one cases, the second consisting of ninety: this provided a good number of responses in each data set considering the number of variables and items specified in the survey design. These are hereafter referred to as Data Set A ( $n = 91$ ) and Data Set B ( $n = 90$ ). Data Set A was used to perform a principal component analysis (PCA), Data Set B was used to perform principle axis factoring (PAF). The factorability of the data was confirmed through the KMO test (.65 and .69, respectively) and Bartlett's test of sphericity ( $p < .001$ ).



## Results

In this section, three factor analyses are reported: a principal component analysis of Data Set A, principal axis factoring of Data Set B, and then principal axis factoring of the combined data sets.

### Principal component analysis of Data Set A

The PCA was set to extract factors with an eigenvalue over 1.75 and rotated with Promax rotation ( $k=4$ ). The structure matrix generated was inspected, and items with a factor loading over .6 were retained as long their next highest factor loading was at least .3 less than their highest factor loading. However, an exception was made for one item that loaded onto a single factor with a loading of .594 and no other significant factor loadings. This decision was made because the cut-off point of .6 is a very conservative figure, and removing the item would have reduced the number of items retained in the analysis detrimentally. One further item was removed because it was the only item remaining on a factor and, therefore, that factor would not have been able to form a multi-item scale. The final number of items retained was 18 in total, with the factor solution consisting of four factors (see Table 12).

*Table 12 Principal component analysis of Data Set A ( $n = 91$ )*

Factor	Initial eigenvalues			Extracted sums of squared loadings		
	Total	% of variance	Cumulative % of var.	Total	% of variance	Cumulative % of var.
1	7.04	19.56	19.56	7.04	19.56	19.56
2	3.78	10.51	30.07	3.78	10.51	30.07
3	2.95	8.21	38.28	2.95	8.21	38.28
4	2.33	6.74	44.76	2.33	6.47	44.76

The first factor consisted of items hypothesised to tap into *alignment preferences* and *size preferences*. Inspecting the signs of items, it was observed that a preference for centrally aligned key words corresponded to a preference for short spans of context, while preferences for left-margin alignment corresponded to preferences for longer spans. As such, Factor 1 was identified

as *citation format preference*: a preference for KWIC format concordances, in contrast to a preference for non-KWIC format concordances, presumably sentence format citations. The other three factors emerged as hypothesised: Factor 2, *corpus type preference*; Factor 3, *citation order preference*; Factor 4, *task preference*. It should be noted, however, that Factor 1 and Factor 3 were fairly strongly correlated with each other (.47).

### **Principal axis factoring of Data Set B**

PAF was performed on the reduced set of 18 items using Data Set B. This analysis extracted factors with an eigenvalue above .95, used Promax rotation ( $k=4$ ), and retained items with a factor loading over .4. Reviewing the structure matrix for the analysis, it was apparent that partial factor loadings were significant in interpreting the factor structure identified. Ostensibly, the solution obtained had indicated six factors; however, inspecting the data more thoroughly revealed patterns of correlation that undermine this interpretation (see Table 13).

Inspecting correlations between individual items in the structure matrix suggested that, instead of six distinct factors, the analysis actually identified the same factors as before, although items from two factors were now spread over multiple oblique factors. Specifically Factors 2 and 4 in PAF consisted of items that had been written to tap into a single hypothesised factor (*corpus type preference*), had formed a single factor in the PCA, and remained strongly correlated with each other in the PAF. Similarly, Factor 5 in the PAF solution consisted of a single item from the factor *citation formatting preference* in the PCA, and this item remained strongly correlated with those items. Thus, the differences in the number of factors identified does not reflect fundamentally different factor solutions, but are a result of differences in how PCA and PAF handle measurement error, and the relatively small sample size available (Matsunaga, 2010; Woods et al., 1986). While the PCA identified four largely orthogonal factors, PAF identified six factors, with pronounced oblique relationships between factors that had formed single factors in the PCA. Thus, taking these item correlations into account, the PCA and PAF actually produce very similar factor solutions and, as such, constitute strong support for the reliability for the factor solution obtained. The only theoretically significant difference between the solutions obtained is the loading of one item (*Size 2*) onto a different factor: in the PCA, it had formed a factor with the other size and alignment items (the *citation format preference* factor), but in PAF it had migrated to the partially correlated factor, *citation order preference*. Indeed, this migration

of an item in the PAF is not too surprising or worrying given the correlation of the two factors that was observed in the PCA.

*Table 13 Structure matrix of principal axis factoring Data Set B (n = 90)*

Item Code	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Alignment 1	<b>.95</b>		.44			
Alignment 2	<b>.90</b>		.44		.44	
Alignment 3	<b>.77</b>					
Size 1	<b>.53</b>		.41		.46	
Corpus 1		<b>.91</b>		.47		
Corpus 2		<b>.81</b>		.42		
Corpus 3		<b>.77</b>				
Order 1			<b>.91</b>			
Order 2			<b>.72</b>			
Order 3			<b>.59</b>			
Size 2			<b>.56</b>			
Corpus 1		.45		<b>.94</b>		
Corpus 2				<b>.79</b>		
Corpus 3		.46		<b>.70</b>		
Size 3					<b>.97</b>	
Task 1						<b>.67</b>
Task 2						<b>.52</b>
Task 3						<b>.41</b>

Note: items are labelled by the dimensions they were hypothesised to tap, and numbered by the order they are listed in.

### Clarifying the factor solution

To help clarify the exact structure of the factor solution, the analysis was repeated using the complete data set. Unlike the original PAF analysis of Data set B, conducting a PAF on the complete sample does not constitute a stringent test of the factor solution because it is no longer performed on a completely independent sample. However, being based on a larger data set ( $n = 181$ ), it is more robust to measurement error. This analysis initially produced a five factor solution. The factor *format preference* emerged once again as it had in the PCA; it was composed of all of the items in the analysis that had been hypothesized as *alignment* and *size* items. However, the items hypothesised as *corpus type* remained split across two strongly correlated factors, as in the PAF of Data Set B (see Table 13 above). Taking both theoretical considerations

and partial item correlations into account, two items were removed from the analysis, thereby simplifying the factor structure (Matsunaga, 2010). The final factor structure obtained was thus based on 181 respondents' scores on 16 items (see Table 14).

*Table 14 Principal axis factoring of the complete sample (n = 181)*

Factors	Initial eigenvalues			Extracted sums of squared loadings		
	Total	% of variance	Cumulative % of var.	Total	% of variance	Cumulative % of var.
1 Format	4.40	27.49	27.49	3.99	24.92	24.92
2 Corpus	2.39	14.90	42.39	2.05	12.79	37.71
3 Order	1.92	11.98	54.37	1.49	9.28	46.99
4 Task	1.48	9.26	63.63	0.88	5.48	52.47

As with the PCA, the four factors were identified as *format preference*, *corpus type preference*, *citation order preference* and *task preference*, each consisting of the same items as in the PCA, minus the two items that had been removed from the analysis. Again, as in the PCA, the *format* factor and the *order* factor were partially correlated (0.31). Together, these four factors accounted for 52.47% of the variance in the data, which can be considered a credible amount of variance accounted for in a study of this kind (Plonsky & Gonulal, 2015). The table below shows the factor loading of each of the items that loaded onto each of the factors identified.

Table 15 Factor loadings of items for complete sample ( $n = 181$ )

Item Code	Item	Factor Number	Factor Loading
Alignment 1	I find it most useful when the node of each citation is aligned to the centre of the page <sup>a</sup>	1	.88
Alignment 2	In my experience, concordances are easiest to use when the node of each citation is aligned to the centre of the page <sup>a</sup>		.79
Alignment 3	It is important to me that the node is presented in a single column in the middle of the concordance <sup>a</sup>		.74
Size 1	In my experience, concordances are easiest to use when each citation is presented on a single line of text <sup>a</sup>		.70
Size 2	In my experience, concordances are easiest to use when each citation presents multiple lines of text		.50
Size 3	I find concordances most useful when sentences that are longer than the screen width are cut short <sup>a</sup>		.45
Corpus 1	I find it most useful when the citations in the concordance are from a large general corpus	2	.91
Corpus 2	I like to use concordances of a large general corpus		.80
Corpus 3	In my experience, concordances are easiest to use when the citations in the concordance are from a large general corpus		.76
Corpus 4	I find it most useful when the citations in the concordance are from a narrow, focused corpus <sup>a</sup>		.54
Order 1	I find it most useful when concordances are organised in relation to the texts that citations come from	3	.90
Order 2	In my experience, concordances are easiest to use when the order of citations is determined by the text each citation comes from		.82
Order 3	Concordances that present citations from related texts together suit my purposes best		.63
Task 1	I am searching for general rules or descriptions	4	.60
Task 2	I am looking for patterns		.53
Task 3	I focus on relationships between citations		.49
Note: items are labelled by the dimensions they were hypothesised to tap, and numbered by the order they are listed in			
<sup>a</sup> Items that were reverse coded (i.e. high agreement was scored as low agreement)			

## Discussion

Factor analysis of the data provided strong support for viewing concordancing as a multi-dimensional construct. Respondents' concordancing preferences were distinguishable in relation to *citation formatting preference*, *corpus type preference*, *citation order preference* and *task*

*preference*, although it should be remembered that *citation formatting preference* and *citation ordering preference* were partially correlated. Correlation between these two factors is not particularly surprising given the observation that ordering citations by cotext facilitates observation of patterns of lexical co-occurrence (Hunston, 2010; Sinclair, 1991; Tribble, 2010), and that KWIC formatting can make patterns of lexical co-occurrence visible via the creation of spatially meaningful relationships between citations. However, as these two factors did not converge to form a single factor, if we accept the idea that concordance users' preferences will be related to affordances, then this also indicates that KWIC format concordances may still provide users affordances when ordered by the texts citations come from. A similar point can be made vis-à-vis task: as *citation format preference* and *task preference* emerge as distinct factors, we can infer that a preference for looking for repetition amongst citations is not dependent on a preference for KWIC formatting. The identification of four underlying factors suggests that differences in concordancing conditions is not best described as variation upon a single prototypical activity, but rather as a complex set of factors, each of which may be providing users with distinguishable affordances. As such, care needs to be taken to identify the specific conditions of learner concordance use when comparing or generalizing from studies.

The four factors identified as underlying respondents' concordancing preferences emerged, in the main, as hypothesised. Aspects of concordancing hypothesized to offer divergent affordances were found to be distinguishable. The design of the survey had, however, implicitly hypothesized *citation alignment* and *citation size* as separate underlying factors because, in theoretical terms, these two aspects of citation display are separable. This study demonstrated that, in terms of concordance users' preferences, these two hypothetical factors are in fact indistinguishable. Items from the two hypothetical dimensions converged to form a single observable factor, *citation formatting preference*: a preference for centrally aligned citations correlates with a preference for short spans of cotext; a preference for left-aligned citations correlates with a preference for longer spans of cotext. This can be thought of as a preference for either KWIC format concordances, or non-KWIC format concordances, sentence format citations perhaps.

As the first factor extracted, *citation formatting preference* explains the highest proportion of variance. This importance of citation formatting is congruent with speculation about the affordances of KWIC formatting in the research literature (Basturkmen, 2006;

Bernardini, 2002; Cobb & Boulton, 2015; Hunston, 2002; Johns, 1986; O’Keeffe et al., 2007; Sinclair, 2004; Warren, 2011; Willis, 1998), highlighting the fact that whether studies employ KWIC format concordances or not is a potentially significant variable. Furthermore, it would be rash to assume that KWIC formatting is the most appropriate format to use in all pedagogical contexts because a significant proportion of concordance users in this study preferred to use non-KWIC format concordances – perhaps because they offer affordances more suited to those users’ needs. As suggested above, in the context of learner use, one possible rationale may be a lower authentication burden for sentence format citations. However, in terms of the respondents to the survey, the great majority of whom did not identify themselves as language learners, the affordances of non-KWIC format citations may be related to greater contextualisation of the search term more generally, or else interest in co-textual relationships to the search term at greater remove than those typically available in KWIC format citations. Similar observations could be made in regards to the other three factors; the type of concordance used, the order citations are presented in, and task orientation were all identified as dimensions of variation in concordance users’ concordancing preferences and therefore each needs careful consideration in the context of affordances and learner use of concordances.

In regards to *corpus type preference* in particular, the results demonstrate that different users can be distinguished on the basis of preferences for either larger, more general corpora, or smaller, more focused corpora and, as the second largest dimension of variance identified, this suggests that language proficiency may not be the only factor that needs to be considered when determining which corpus to use. The results indicate that corpora at different levels of generality provide concordance users with different affordances, and so this is likely to also be an important variable in learner concordance use. Given that many non-learner users of concordances have a preference for using small, narrowly focused corpora and the potential for small, narrowly focused corpora to reduce the authentication burden of concordances (Aston, 2001; Braun, 2005; Gavioli, 2005; Gavioli & Aston, 2001), the decision to use very large, general corpora with learners should be carefully evaluated. It would perhaps be unwise to use such corpora with learners by default. Corpus type certainly appears to be a variable in learner concordance use that deserves more research attention (though see Charles, 2012, 2015).

In regards to *task preference*, although users could be distinguished on the basis of a preference for looking for repeated information or not, it accounted for the smallest amount of

variance amongst the factors identified. It is perhaps surprising that this factor did not converge with *citation order preference* or *citation format preference* given that these two preference types are strongly associated in the literature with the identification of patterns of lexical occurrence (Barlow, 2004; Gilquin & Granger, 2010; Hunston, 2002, 2010; Sinclair, 1991; Tribble, 2010). This result may be an indication that concordances afford an opportunity to observe repeated features of language that are not always realised through a simple lexical pattern: for instance, more complex aspects of semantics and usage. This may account for these preference emerging from the analysis as a weak but distinguishable dimension of variance amongst concordance users. At the same time, this result does suggest that the theoretical perspective on concordance use as involving comparison and contrast between citations or else comparison of individual citations with an external context of use may have some currency with concordance users and, as such, it also represents a promising line of enquiry for future research.

### **Limitations**

The decision to use snowball sampling was justified by its potential to collect a larger data set than other sampling schemes, but it also prohibits confident generalisation of the results obtained. Even though the convergence of factor solutions provides good grounds upon which to consider the solution identified as reliable, it is not possible to define the population the sample was drawn from, raising the question of whether the sample obtained represents a biased sample of concordance users in general. As such, the findings reported here apply most properly to *the kinds of respondents this survey was able to obtain responses from*. Whether this sample of concordance users is typical of concordance users is an empirical question that can only be evaluated by further research. However, that being said, biographical data collected from respondents does indicate that the data was collected from a similar population of respondents' to those reported on in Christopher Tribble's 2012 survey (Tribble, 2015), although respondents to the survey reported here answered from a far wider range of orientations to concordancing: not only teaching, but also learning, materials design, translating, lexicography and linguistics.

It might also be noted that the survey instrument employed in this study was not exhaustive of all aspects of concordancing that might be of research interest, and that presenting the dimensions of concordancing that were explored as dichotomous prevented respondents'



from indicating alternative preferences, such as preferences for randomly ordered citations, or medium sized corpora. Both of these issues would merit further investigation in future studies.

### **Summary**

This study has identified four vectors of potentially significant difference in how concordancing is operationalised. The four factors identified in this study should all be considered to be potentially significant variables in relation to how concordancing is operationalised with learners, which concordancer interface options to provide them with, and the design of language learning tasks that use concordances more generally. If we may assume that concordance users' concordancing preferences are not arbitrary, but are in fact motivated by different affordances that correspond to the different configurations of various aspects of concordancing (citation format, corpus type, citation ordering and task), then it appears that there is much work to be done in identifying which aspects of concordancing naturally coalesce, and how they may be able to inform the most effective exploitation of concordances in various pedagogical contexts. While experienced concordance users may well be adept at adapting concordances to their needs at any given time (Sinclair, 1991), developing models of different types of concordance use may have an important role to play in the design of pedagogical concordancers and concordance-based language learning materials. In the next chapter, these issues are explored further by looking for correlations between the respondents' concordancing preferences and different categorisations of concordance user: by identifying types of concordance preference with types of concordance user it sheds some light on the motivations that lay behind concordancing preferences, helping us to understand how concordancing preferences interact with the affordances they provide, helping us to match the configuration of a concordance with a particular goal or orientation to concordance use.



## Chapter 4. Correlations between concordance users and concordancing preferences

### Introduction

This chapter reports an extension of the factor analysis reported in Chapter 3. It uses factor scores derived from the factor analysis of survey items reported in Chapter 3, and explores them in relation to data that was not used in the analysis presented in Chapter 3. Namely, four different ways of categorising the survey respondents. The study reported in this chapter explores the extent to which each of these four categorical variables is able to predict the respondents' scores on each of the four continuous, factor score variables derived from the analysis in Chapter 3 and reported here. To avoid unnecessary repetition, details of the survey and factor analysis itself are not presented in this chapter: please refer to pages 81-89 of Chapter 3 for details about the survey instrument developed, research ethics, distribution of the survey and biographical details about the survey respondents, and pages 91-95 for the factor analysis results.

To briefly summarise the factor analysis reported in Chapter 3, it showed that there are at least four distinguishable factors which underlie concordance users' concordancing preferences: *format preferences*, *corpus preferences*, *order preferences* and *display preferences*. It was also asserted that these systematic differences in response patterns are likely to be motivated and not arbitrary. However, factor analysis on its own does little to illuminate what these motivations may be. The study reported in this chapter explores this issue by examining four potential predictors of concordancing preferences by looking for correlations between four ways of categorising respondents and factor scores derived from the factor analysis. The four ways of categorising respondents are: how long they have used concordances for (*length of use*), how frequently they use concordances (*frequency of use*), the linguistic relationship to the concordance that they were thinking of when completing the survey (*language of use*) and the field of use that they were thinking of when completing the survey (*field of use*).

This chapter begins by explaining the theoretical significance of the four ways of categorising concordance users that are explored in this study and providing some basic descriptive data in relation to each. It then provides a brief account of how factor scores were generated and descriptive statistics for the factor scores produced as a whole (please see Chapter 3 for a more theoretical discussion of the significance of these variables in relation to learner use

of concordances). This chapter then states the research questions addressed by the study reported here and provides details of the methodology employed. The results of inferential statistical tests are then reported, followed by a discussion of the results' implications, as well as the limitations of the study reported.

### **Four ways of categorising concordance users**

There are of course more than four ways in which concordance users could be categorised. The four ways of categorising users that are explored in this study were selected because they were of theoretical significance, and they could be treated as simple variables that could be operationalised via individual survey items: after all, due to considerations pertaining to survey length and completion time, not every variable can be treated as a complex variable and addressed through a multi-item scale (Punch, 2003). Thus, each of these four categories was operationalised through a simple, single answer, multiple choice question (see Table 8, p. 85 for an outline of the survey, and Appendix 1 for screenshots of the survey instrument). For this reason, the data collected on these four ways of categorising respondents was treated as category data, meaning the responses presented to respondents did not represent continuous or interval scales. This is appropriate, because self-report data such as that gathered through an online survey reports people's assessments of their own behaviour, and such assessments can differ markedly from objective observation of behaviour — much as our intuitions about language differ markedly from language use as attested in corpus data, our intuitions about how often we do something or how likely we are to do something are poor indicators of the publically observable facts. Survey data captures facts about perceptions, not facts about actions. In brief, it should be borne in mind that data collected about the four variables discussed below applies to the respondents' self-perceptions, not objective facts about their behaviours, and was used to create categorical response groups within each variable.

#### ***Length of use as a variable***

One variable of interest was how long respondents have been using concordances for. In particular, there is a sense in the literature that key word in context (KWIC) formatting is off-putting for inexperienced users, but that as experience grows, so too does comfort with this

formatting choice. For instance, making an appropriately nuanced point, Tribble (2010, p. 172) states:

There are advantages and disadvantages to this kind of display [KWIC] ... but in the early stages of introducing concordancing to learners and other students of language it is sometimes the case that a sentence view of concordance data can be more useful than the KWIC display as it presents fewer reading challenges to newcomers to corpus analysis.

Tribble's point is clearly that the format's inherent advantages and disadvantages notwithstanding, it may be sensible to start users off with sentence format citations. Hence, it seemed quite possible that the distinguishable preferences for KWIC and non-KWIC formatting observed in Study 3 might be, in part at least, a function of user experience: less experienced respondents may prefer non-KWIC format concordances, and more experienced respondents KWIC. This point parallels the extensive discussion of the importance of training in the literature (Bernardini, 2004; Braun, 2007; Götz & Mukherjee, 2006; C. Kennedy & Miceli, 2001; Mukherjee, 2006; Pérez-Paredes et al., 2012; Tono et al., 2014).

Data was collected on this variable via a single answer, multiple choice question with the following response categories: *less than a year*; *1 to 3 years*; *3 to 10 years*; *10 to 20 years*; *more than 20 years*. Note, as these variables were to be used as categorical variables, it was not necessary for the time frames referred to, to be numerically equal.

*Table 16 Frequency count for scores on Length of use*

Response category	<i>n</i>	Cumulative <i>n</i>	Percentage	Cumulative percentage
Less than a year	18	18	9.9	9.9
1 to 3 years	40	58	22.1	32.0
3 to 10 years	66	124	36.5	68.5
10 to 20 years	32	156	17.7	86.2
More than 20 years	23	179	12.7	98.9
/ missing responses	2	181	1.1	100

Frequency counts for *length of use* are shown in the Table 16 above. It can be seen that *3 to 10 years* of concordance use is the mode average for concordance users in this data set and

that the category with the lowest number of respondents was respondents who had used concordances for *less than a year* with 18 responses.

### ***Frequency of use as a variable***

The second independent variable addressed was how frequently respondents use concordances. In many ways this variable complements *length of use*, as a concordance user's total experience using a concordance is the combination of both how long they have used concordances for and how frequently they have used concordances during this time. However, there is also a sense in which *frequency of use* reflects how central concordancing is to a respondents' work or study. This perspective on frequency of use resonates with discussion in the literature of the place of learner use of concordances within their broader educational experience: whether learner use of concordances represents a paradigm shift in how learners approach language study, or whether concordances are simply a supplementary resource, complementary to other more traditional forms of study. The latter view appears to be gaining traction in the literature (Boulton, 2009b; C. Kennedy & Miceli, 2010; Pérez-Paredes, 2010; Tono et al., 2014), and if concordance use is best thought of as supplementary to more traditional language learning techniques, then it might be felt more reasonable for learner use of concordances to reflect a model of more casual concordance use than the more intensive, should significant differences between more frequent and less frequent users be observed. In many ways, this distinction parallels the distinction made between strongly researcher-like and more reference-resource-type use discussed in Chapter 2.

Frequency of use is a difficult variable to address because, not only is our ability to accurately recall how often we use concordances suspect, but so too is the unit of counting: it is not clear how one should quantify quickly checking something in a concordance vis-à-vis a period of extended corpus linguistic analysis. The trick here was to develop a response scale that respondents could confidently answer (Dörnyei & Taguchi, 2010; Punch, 2003). This was achieved by pairing standard response category wordings (very rarely, occasionally, quite often, very often) with richer descriptive illustrations (see bullets points below). During piloting, the combination of these two levels of response categorisation appeared to facilitate confident and reliable responses from informants. The response scale developed consisted of the following four categories:

- **very rarely:** *I have used concordances a few times*
- **occasionally:** *I use concordances every now and again, but they are not a regular part of my work or study*
- **quite often:** *it's not unusual for me to be using concordances for one reason or another*
- **very often:** *concordances are a key component of my work or study*

Frequency counts for *frequency of use* are shown in Table 17 below. Perhaps unsurprisingly, there are substantially fewer *very rare* users than others.

*Table 17 Frequency count for scores on Frequency of use*

Response category	<i>n</i>	Cumulative <i>n</i>	Percentage	Cumulative percentage
very rarely	22	22	12.2	12.2
occasionally	46	68	25.4	37.6
quite often	54	122	29.8	67.4
very often	58	180	32.0	99.4
/ missing responses	1	181	0.5	99.9

### ***Language of use as a variable***

Language of use refers to the linguistic relationship between the concordance user and the concordance text: for instance, use of a concordance that presents text in one's L1, use of a concordance that presents text in an L2, or use of a concordance that presents text in two languages simultaneously, a bilingual concordance. In Chapter 2 (pp. 39-47), it was noted that there is reason to think that less proficient language users may be particularly prone to the type of authentication burden presented by concordances because successfully interpreting concordances is thought to involve intuition and pre-existent knowledge of cultural, textual and linguistic norms. For this reason, it seemed likely that differences in concordancing preferences may be predicted by whether a concordance user is using a concordance of their first language or a second language: for instance, if we accept that KWIC format concordances have a greater authentication burden than non-KWIC format concordances (Aston, 2001; Boulton, 2009b; Cobb, 1999; Johns, 1986; Levy, 1990; Mishan, 2004; Stevens, 1991a; H. Yoon & Hirvela, 2004),

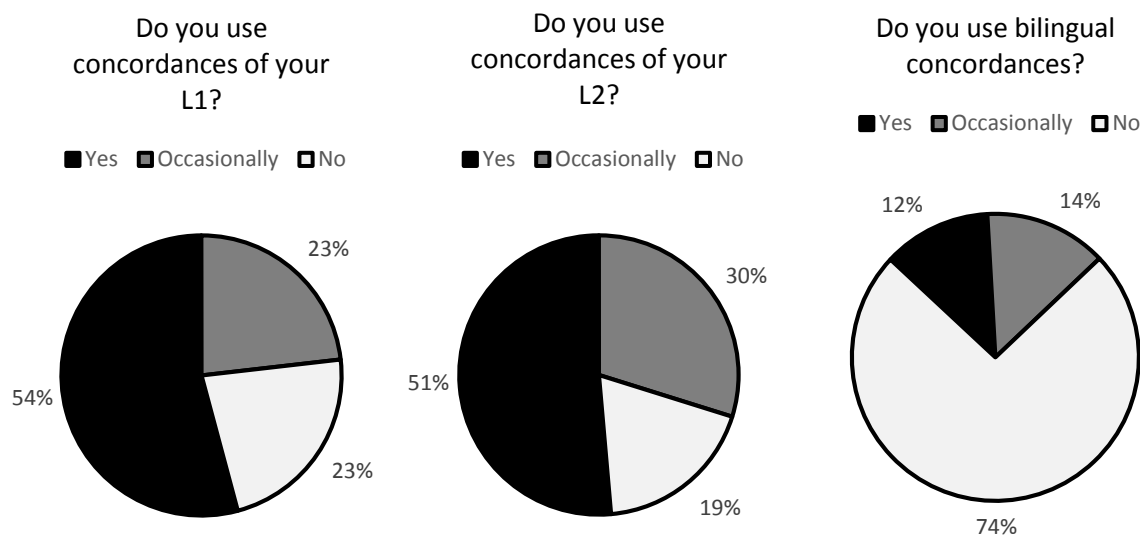
then using KWIC format concordances may compound the authentication burden of using concordances in a second language, and this might be detectable as a preference for using non-KWIC formatting. A parallel point could be made in regards to having preferences for large general corpora or small focused corpora, as these are also thought to present different authentication burdens (Aston, 2001; Braun, 2005; Gavioli, 2005; Gavioli & Aston, 2001). However, the situation is confounded somewhat by the option of using parallel concordances; parallel concordances typically being indexes of two corpora of two different languages that present citations from both corpora side by side (see for example Chan & Liou, 2005; Lai & Chen, 2013). Using a parallel concordance might involve using a concordance of a first language and a second language, two second languages, or possibly even two first languages, depending on how you define a first language.

In terms of how this variable was operationalised as a survey item, respondents were first asked to complete a matrix question about which types of concordances they used. The matrix consisted of three language of use categories (concordances of a first language, of a second language and bilingual) and three response categories (*yes*, *occasionally*, *no*). This allowed respondents to indicate a fairly full range of linguistic relationships to the concordances they use without the question becoming overly complex and unwieldy. This data is summarised in Figure 1 below. This descriptive data indicates that the vast majority of respondents use concordances of both their L1 and an L2, at least occasionally (77% and 81% respectively), but only around 50% in each of these categories characterised their use with an unqualified *yes* response. In contrast to the L1 and L2 concordance responses, a comparatively small proportion of respondents indicated that they use bilingual concordances (*yes*, 12%; *occasionally*, 14%). Exactly why this may be is not entirely clear: however, one possible explanation is the comparatively limited availability of resources for bilingual concordances. At its simplest level, you need twice as many corpora to use a bilingual concordance, but perhaps more interestingly, bilingual corpora can be used with parallel texts (resources consisting of two corpora, each consisting of texts that are paired with other: texts where one is a translation of another, or else both are translations of a third), and these can then be aligned so that accessing a citation in one simultaneously brings up the translation equivalent in the other. Of course, both the availability of such paired texts and the costs of aligning the corpora severely restrict the availability of such resources. The interested reader is referred to Kenning (2010) for a detailed account of the functionality of bilingual



concordancers in the context of corpus linguistics, but see also Kübler and Aston (2010) for discussion of bilingual concordances in the context of translation, a field of concordance use in which bilingual concordances feature prominently.

*Figure 1 Percentages for Language of use: Yes, Occasionally, No [n = 181]*



Taking a broad view of the data collected, it is clear that most of the survey respondents use concordances in more than one linguistic relationship. Consequently, as described in Chapter 3, respondents were asked to think of only one linguistic relationship to a concordance while answering the survey questions pertaining to concordancing preferences, as it is not clear that they would use concordances in the same way in each condition, for the reasons discussed above. Asking respondents to complete the survey while thinking of only one linguistic relationship to the concordance allows respondents' answers to concordancing preferences questions to be interpreted as pertaining to only the type of linguistic relationship chosen. The choice was operationalised through a straightforward single answer, multiple choice question, the answers to which are summarised below. The number of respondents who chose to answer from the perspective of bilingual concordance use is slightly disappointing, though perhaps not altogether unexpected given the relatively small proportion of respondents who indicated that they use bilingual concordances at all.

Table 18 Frequency count for Language of use

Response category	<i>n</i>	Cumulative <i>n</i>	Percentage	Cumulative percentage
I use concordances of my 1 <sup>st</sup> language	93	93	51.4	51.4
I use concordances of a 2 <sup>nd</sup> language	75	168	41.4	92.8
I use bilingual concordances	13	181	7.2	100

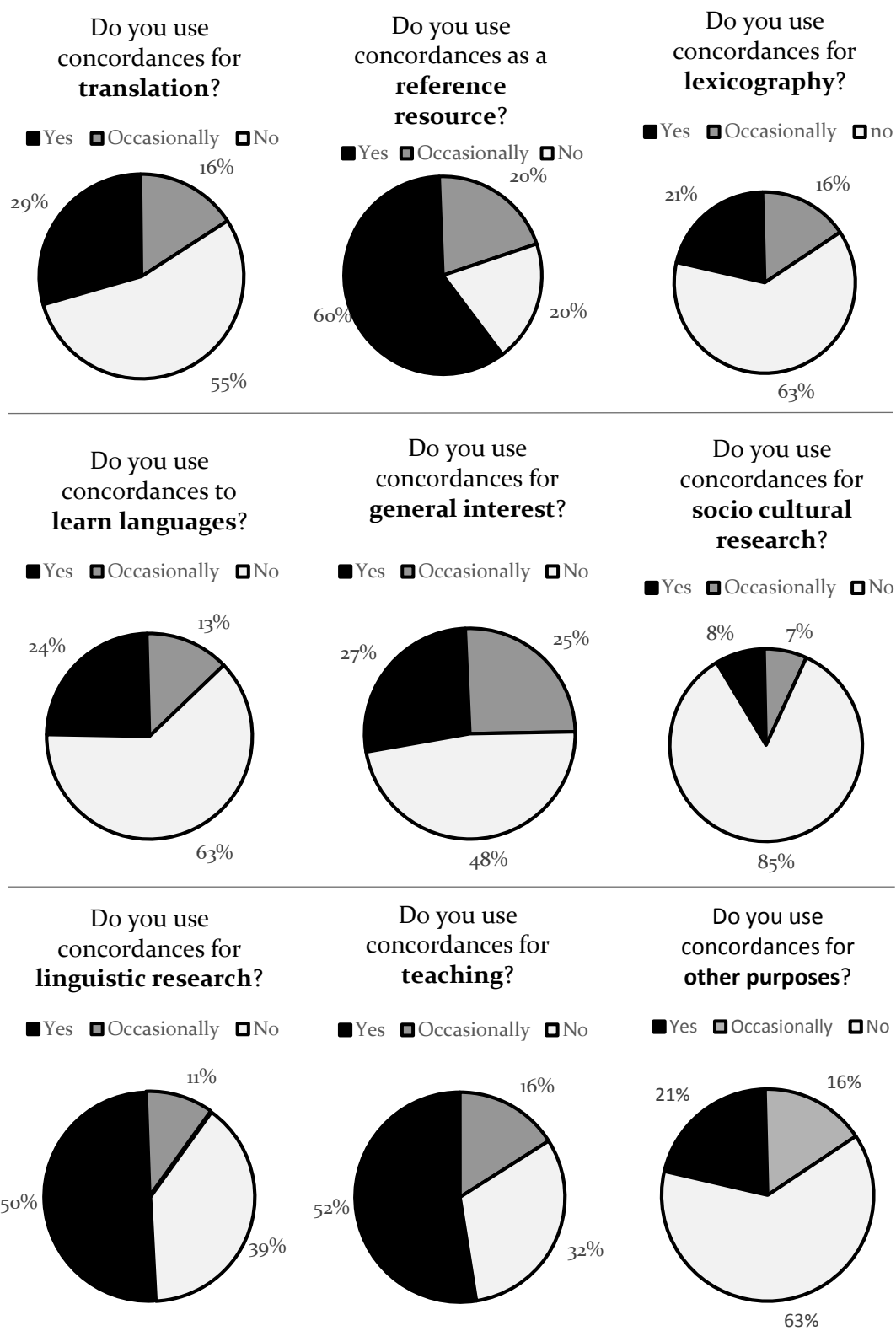
### *Field of use as a variable*

The last variable of interest investigated in this study is field of concordance use. As discussed in Chapter 2 (pp. 35-63), there is good reason to think that the conditions of concordance use in corpus linguistic research will differ from more casual use of concordances in, say, using a concordance as a reference resource. On the basis of the literature, a provisional list of fields was developed, and then developed over several rounds of piloting with both expert and non-expert informants. Eight response categories were eventually identified for *field of use*, along with a ninth catch-all category, *other*:

- for linguistic research
- for socio-cultural research
- for translation
- for lexicography
- in teaching
- to learn languages
- as a reference resource
- for general interest
- other (*please specify*)

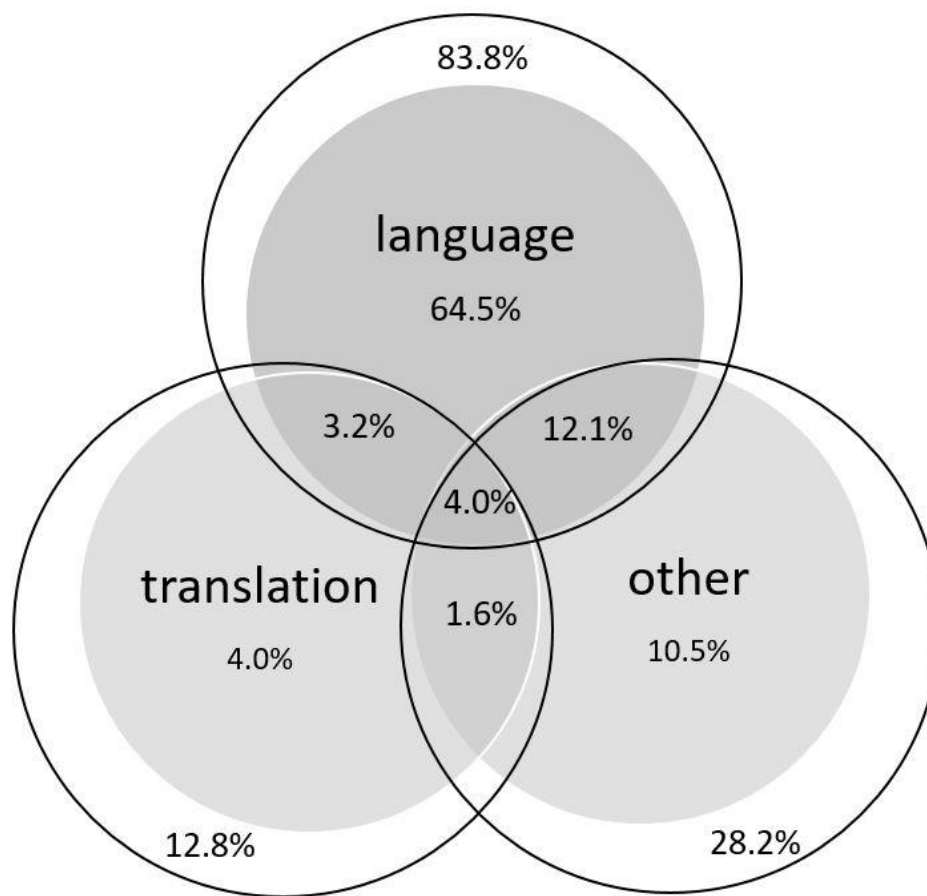
As with language of use, field of use data was first collected via a matrix, allowing respondents to indicate the range of fields of use that apply to them, as summarised in the pie charts in *Figure 2*. The data collected with *yes*, *occasionally*, and *no* as response categories paints a broad picture of how respondents use concordances. For instance, it shows that less than 50% of respondents use concordances *for translation, for lexicography, to learn languages, for socio-cultural research or for other purposes*, and that over 50% of respondents reported using concordances *for general interest, for linguistic research, for teaching and as a reference resource*, the latter appearing to be the most common field of use overall.

Figure 2 Percentages for Field of use: Yes, Occasionally, No [ $n = 181$ ]



Of the 181 respondents asked whether they use concordances for teaching (answering *yes* or *occasionally*), 124 respondents indicated that they did. This means that 68.5% of the survey respondents indicated that they use concordances in teaching, at least occasionally. These 124 respondents were then asked to indicate whether they used concordances to teach *language*, *translation* or *other*. The response category *other* also prompted the respondents to specify the domain of teaching concerned. In the figure below, the shaded areas form a Venn diagram that shows the percentage of respondents in each category or combination of categories, the sum equalling 100% of the 124 respondents who use concordances in teaching. In contrast, the circle outlines in the figure present the percentage of the 124 respondents who use concordances for teaching in each domain separately: of course, given that some respondents use concordances when teaching in two or more domains, the figures in the circle outlines cannot be summed.

Figure 3 Percentages of concordance use in teaching by domain [ $n = 124$ ]



These results are perhaps surprising. First, around one fifth of respondents who use concordances in teaching use concordances in more than one domain (20.9%). Second, rather unexpectedly, only a small proportion of respondents use concordances to teach translation (12.8%), and an even smaller proportion only taught translation with concordances (4%). However, most unexpected is the high proportion of respondents who indicated they used concordances to teach something other than language or translation: over a quarter of the respondents (28.2%).

The *other* response category had included a text input box and requested respondents to indicate what *other* referred to in their case. Answers are summarised in Table 19 below. Clearly, teacher training/TEFL could be classified as applied linguistics, and hence nested under the category of linguistics/applied linguistics; however, as there was no indication whether or not the teacher training was in an academic context or not, it was decided to list them separately. However, either way, it would still seem that a clear majority of using concordances in teaching *other* responses referred to using concordances in teaching linguistics, applied linguistics, and perhaps by extension, teacher training and TEFL.

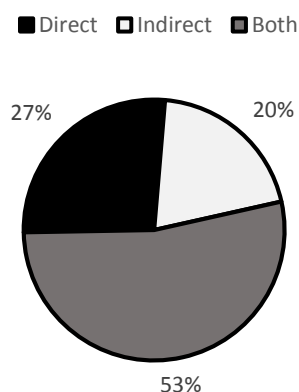
*Table 19 Frequency count of other domains of concordance use in teaching (n = 35)*

Response category	<i>n</i>	Cumulative <i>n</i>	Percentage	Cumulative percentage
Linguistics or applied linguistics	23	23	65.7	65.7
Teacher training/TEFL	5	28	14.3	80.0
Cultural and religious studies	2	30	5.7	85.7
Paper editing	1	31	2.8	88.5
/ no answer	1	32	2.8	91.3
/ uncategorised answers <sup>a</sup>	3	35	8.6	99.9

<sup>a</sup> corpus-based translation studies (PhD research project); ESP; language for science

The same 124 respondents who had indicated that they used concordances in teaching were then asked whether they used concordances *directly*, *indirectly* or *both* (Figure 4). Interestingly it seems that the majority of respondents who use concordances for teaching use them both directly and indirectly (53%), and more respondents used them just directly (27%) than just indirectly (20%).

Figure 4 Percentages for using concordances in teaching: Directly, Indirectly or Both [ $n = 124$ ]



Answers about the domain of teaching referred to were not followed up further in the analysis reported here, but answers pertaining to direct and indirect use were carried forward in the survey, bringing the total possible number of field of use categories to nine plus *other*.

In the next stage of the survey items about *field of use*, answers to earlier field of use answer categories were carried forward to the next question. That is, if a respondent had answered *yes* or *occasionally* to a response category on the previous question, it was presented as a response category in this next question; with *direct*, *indirect* or *both* being used to decide which response categories to provide for the two new response categories, *teaching: direct* and *teaching: indirect*. Respondents were then asked to estimate the relative amount of time they spent using concordances across each of the response categories that they had indicated were relevant to them: to distribute 100% of their time over the categories that were displayed to them on the basis of their previous answers. This type of question provides a much more nuanced answer than simply *yes* or *occasionally*, but a proportion of respondents did not answer it (16 out of 181). Some basic descriptive data is displayed in Table 20. The maximum scores show that none of the respondents only use concordances *to learn languages, for socio-cultural research, for general interest, for lexicography or for other purposes*. The mean scores show that, on average, concordance users spend more of their time using concordances *for linguistic research* than for any other purpose, even if the *yes, occasionally, no* data shows that using concordances *as a reference resource* is a more common use in an absolute sense.

Table 20 Field of use by percentage of time spent

Field of use	Minimum	Maximum	Mean	SD
for linguistic research	.00	100.00	29.00	31.61
as a reference resource	.00	100.00	18.56	20.63
for translation	.00	100.00	12.07	24.15
in teaching: indirectly	.00	100.00	11.96	20.15
in teaching: directly	.00	100.00	11.24	18.72
for general interest	.00	35.08	4.87	7.18
for lexicography	.00	90.00	4.68	13.26
to learn languages	.00	80.00	4.66	10.51
for socio-cultural research	.00	86.00	2.36	9.10
other	.00	81.33	0.71	6.56

Finally, as discussed for *language of use* above, it was necessary for respondents to choose a single field of concordance use to think of while answering the questions pertaining to their concordancing preference. Again, this was achieved by feeding forward respondents' answers to earlier *field of use* questions. Table 21 below shows a frequency count of answers to this question. Because the number of respondents answering from the perspective of *for general interest* and *other* were each under five, this meant that these two categories of use could not be included in the analyses reported in this study.

Table 21 Frequency count for Field of use as a variable in this study

Response category	<i>n</i>	Cumulative <i>n</i>	Percentage	Cumulative percentage
for linguistic research	63	63	34.8	34.8
in teaching: indirectly	29	92	16.0	50.8
for translation	26	118	14.4	65.2
in teaching: directly	22	140	12.1	77.3
as a reference resource	19	159	10.5	87.8
for socio-cultural research	8	167	4.4	92.2
for lexicography	6	173	3.3	95.5
to learn languages	5	178	2.8	98.3
for general interest	2	180	1.1	99.4
other	1	181	0.5	99.9

## Factor scores

The aim of this follow-up study was to explore the extent to which different ways of categorising concordance users could predict respondents' concordancing preferences. To this end, factor scores were produced for each of the factors found to underlie the respondents' concordancing preferences in Chapter 3. For each combination of dependent variable and respondent, weighted scores on the relevant items in the survey were combined to produce a single factor score for that respondent on that variable. These scores were calculated in SPSS, producing factor scores with means of zero across the sample as a whole (see *Table 22*).

Table 22 Descriptive statistics for factor scores

	Range	Minimum	Maximum	Mean	SD
Factor 1: Format	4.52	-3.17	1.35	.00	.94
Factor 2: Corpus	4.09	-2.24	1.84	.00	.95
Factor 3: Order	4.61	-2.33	2.28	.00	.94
Factor 4: Task	4.85	-2.68	2.17	.00	.79

## Research questions

This study addresses the following research question:

1. To what extent does each of the independent variables in this study (*length of use; frequency of use; language of use; field of use*) predict each of the dependent variables (*formatting preferences; corpus type preferences; citation ordering preferences; task preferences*)?



## Methodology

In this study, four ways of categorising concordance users are used as independent variables, with factor scores derived from the factor analysis reported in Chapter 3 providing the dependent variables. The dependent variables were continuous score variables (see *Table 22* above), produced by SPSS as part of the final factor analysis reported in Chapter 3. All four independent variables in this study were categorical, as described above. As noted at the beginning of this chapter, for details about the survey instrument developed, research ethics, distribution of the survey and the survey respondents, see pages 89-97 of Chapter 3. Thus, the method section of this chapter is focused on the data analyses undertaken in this follow-up study.

### Data analysis

MANOVA tests were used to test for significant differences in regards to each of the independent variables and the four dependent variables. Discriminant function analysis was then used to quantify the relationship between significant independent variables and the dependent variables. In planning statistical analyses, it is important that both the potential for type I errors and for type II errors are considered (Larson-Hall, 2010). Because the total number of responses available for some categories of some independent variables was quite small and this could result in a loss of power if too many variables were tested at once, it was decided to conduct four separate MANOVAs, thereby preserving statistical power and reducing the risk of type II error. Because MANOVAs are in and of themselves a way of controlling for type I error, this is a good compromise in view of the sample sizes available for analysis. Perhaps more pertinently, the selection of a statistical significance level for results to be compared with is a fairly arbitrary decision (Larson-Hall, 2010), especially as compared with the correlation coefficients obtained in a discriminant function analysis. Given that larger correlation coefficients equate to smaller  $p$  values, the risk of type I error in this study is minimal as long as correlation coefficients are attended to.

## Results

The section reports MANOVA tests for each independent variable and the four dependent variables. Where a MANOVA test identifies significant differences, data is further explored via discriminant function analysis.

### MANOVA for *Length of use*

Factor scores were analysed in relation to the respondents' answers to the question: *How long have you been using concordances for?* The response categories provided were: i) *Less than a year*; ii) *1 to 3 years*; iii) *3 to 10 years*; iv) *10 to 20 years*; v) *More than 20 years*. Two cases were not included in the analysis because respondents had not answered this survey question. Descriptive statistics for the variables are provided in Table 23 below.

*Table 23 Descriptive statistics for Length of use in relation to Format, Corpus, Order and Task*

	Less than a year ( <i>n</i> = 18)		1 to 3 years ( <i>n</i> = 40)		3 to 10 years ( <i>n</i> = 66)		10 to 20 years ( <i>n</i> = 32)		More than 20 years ( <i>n</i> = 23)	
	M	SD	M	SD	M	SD	M	SD	M	SD
Format	-0.31	0.83	-0.02	0.78	0.08	0.95	0.26	0.80	0.26	1.32
Corpus	0.05	0.82	0.13	1.08	-0.11	0.87	0.05	0.89	-0.04	1.14
Order	-0.35	0.61	-0.21	0.89	0.19	0.98	0.07	0.87	0.06	0.94
Task	-0.14	0.90	-0.21	0.80	-0.00	0.73	0.22	0.64	0.17	1.00

The samples were sufficiently large for the assumption of multivariate normality to be met, and the assumption that population covariance matrices were equal was supported by testing Box's *M* (59.04, *p* = .055). MANOVA was used to test the hypothesis that length of use was a predictor of respondents' concordancing preferences (factor scores: *format*, *corpus*, *order* and *use*), but the null hypothesis was supported, Pillai's *F*(16, 179), = 1.53, *p* = .083. These results suggest that *length of use* may not have a significant effect on concordance user's concordancing preferences.

### MANOVA and discriminant function analysis for *Frequency of use*

Factor scores were analysed in relation to the respondents' answers to the question: *How often do you use concordances?* The response categories provided were: i) *Very rarely (I have only used concordances a few times)*; ii) *Occasionally (I use concordances every now and again, but they are not a regular part of my work or study)*; iii) *Quite often (It's not unusual for me to be using concordances for one reason or another)*; iv) *Very often (Concordances are a key component of my work or study)*. One case was not included in the analysis because one respondent had not answered this survey question. Descriptive statistics for the variables are provided in Table 24 below.

Table 24 Descriptive statistics for *Frequency of use* in relation to *Format, Corpus, Order and Task*

	Very rarely ( <i>n</i> = 22)		Occasionally ( <i>n</i> = 46)		Quite often ( <i>n</i> = 54)		Very often ( <i>n</i> = 58)	
	M	SD	M	SD	M	SD	M	SD
Format	-0.22	0.80	-0.27	0.94	0.21	0.76	0.12	1.08
Corpus	0.22	0.84	0.02	0.90	0.15	0.99	-0.24	0.96
Order	-0.12	0.71	-0.13	0.96	0.23	0.91	-0.04	1.00
Task	-0.46	-0.80	-0.21	0.87	0.20	0.67	-.16	0.73

The samples were sufficiently large for the assumption of multivariate normality to be met, and the assumption that population covariance matrices were equal was supported by testing Box's M (30.75,  $p = .505$ ). MANOVA showed that concordancing preferences were related to frequency of concordance use, Pillai's  $F(12, 180) = 2.87, p = .001$ . A direct discriminant analysis was carried out using the four factor scores as predictors of *frequency of use* to determine which of these variables best discriminate between the different frequencies of use. Three discriminant functions were calculated, explaining 78.1 per cent, 19.3 per cent and 2.6 per cent of the variance, respectively (Table 25). Wilk's lambda was significant for the combined functions  $X^2(12, N = 180) = 34.38, p = .001$ , but was not significant when the first function was removed  $X^2(6, N = 180) = 7.88, p = .247$  (Table 26). The first discriminant function maximally discriminated the least frequent concordance users (*very rarely*) from the more frequent users (*quite often* and *very often*) (Table 27), and correlated most strongly with respondents' scores on *task* preferences (.79) and *format* preferences (.52) (Table 28).

*Table 25 Discriminant function analysis eigenvalues: Frequency of use*

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.16	78.1	78.1	.38
2	.04	19.3	97.4	.20
3	.00	2.6	100.0	.07

*Table 26 Discriminant function analysis Wilks' Lamda: Frequency of use*

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.82	34.38	12	.001
2 through 3	.96	7.88	6	.247
3	.99	.96	2	.619

*Table 27 Discriminant function analysis functions at group centroids: Frequency of use*

Frequency of use	Function 1
Very rarely	-.73
Occasionally	-.39
Quite often	.27
Very often	.33

*Table 28 Discriminant function analysis structure matrix: Frequency of use*

Factors	Function 1
Task	.79
Corpus	-.22
Order	.27
Format	.52

### MANOVA and discriminant function analysis for *Language of use*

Factor scores were analysed in relation to the category indicated by respondents in response to the prompt: *Please choose the type of concordance the rest of your answers will refer to*. This question determined group membership for the variable *language of use*. The possible response categories were: i) *concordances of my 1<sup>st</sup> language*; ii) *concordances of a 2<sup>nd</sup> language*; iii) *bilingual concordances*. Descriptive statistics for the variables are provided in Table 29 below.

Table 29 Descriptive statistics for *Language of use* in relation to *Format, Display, Corpus, Order and Task*

	Concordances of my 1 <sup>st</sup> language (n = 93)		Concordances of a 2 <sup>nd</sup> language (n = 75)		Bilingual concordances (n = 13)	
	M	SD	M	SD	M	SD
Format	-0.00	0.97	0.18	0.75	-1.03	1.13
Corpus	0.05	0.96	-0.01	0.97	-0.27	0.69
Order	-0.05	0.93	0.15	0.96	-0.50	0.71
Task	-0.00	0.80	-0.02	0.74	0.17	1.04

The samples were sufficiently large for the assumption of multivariate normality to be met, and the assumption that population covariance matrices were equal was supported by testing Box's M (30.04,  $p = .118$ ). MANOVA showed that concordancing preferences were related to *language of concordance use*, Pillai's  $F(8, 181) = 3.13$ ,  $p .002$ . A direct discriminant analysis was carried out using the four factor scores as predictors of *language of use* to determine which of these variables discriminated most successfully. Two discriminant functions were calculated, explaining 95.4 per cent and 4.5 per cent of the variance, respectively (Table 30). Wilk's lambda was significant for the combined functions  $X^2(8, N = 181) = 24.99$ ,  $p = .002$ , but was not significant when the first function was removed  $X^2(3, N = 181) = 1.22$ ,  $p = .749$  (Table 31). The first discriminant function maximally discriminated *bilingual concordance use* from both *1<sup>st</sup> language* and *2nd language concordance use* (Table 32), and correlated most highly with respondents' scores on *format* preferences (.89) and with *order* preferences (.51) (Table 33).

*Table 30 Discriminant function analysis eigenvalues: Language of use*

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.14	95.4	95.4	.36
2	.01	4.6	100.0	.08

*Table 31 Discriminant function analysis Wilks' Lamda: Language of use*

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.87	24.99	8	.002
2	.99	1.22	3	.749

*Table 32 Discriminant function analysis functions at group centroids: Language of use*

Language of use	Function 1
1st language concordances	-.00
2nd language concordances	.23
Bilingual concordances	-1.29

*Table 33 Discriminant function analysis structure matrix: Language of use*

Factors	Function1
Format	.89
Task	-.15
Corpus	.21
Order	.51

### MANOVA and discriminant function analysis for *Field of use*

Factor scores were analysed in relation to the category indicated by respondents in response to the prompt: *Please choose which use you will think of while answering the rest of the survey.*

This question determined group membership for the variable *field of use*. The possible response categories were: *for translation*; *as a reference resource*; *to learn languages*; *in teaching - directly*; *in teaching - indirectly*; *for general interest*; *for lexicography*; *for linguistic research*; *for socio-cultural research*; and *other*. Box's M tests the assumption of MANOVA that there is equality of covariance matrices between groups, but requires that each group in the analysis consists of more cases than there are dependent variables. Because of this, two categories (*general interest* and *other*) were removed from the analysis to satisfy the criteria for Box's M. This resulted in three cases being removed from the analysis. Descriptive statistics for the variables are provided in Table 34 below.

Table 34 Descriptive statistics for *Field of use* in relation to *Format*, *Corpus*, *Order* and *Task*

	For translation ( <i>n</i> = 26)		As a reference resource ( <i>n</i> = 19)		To learn languages ( <i>n</i> = 5)		For teaching: directly ( <i>n</i> = 22)	
	M	SD	M	SD	M	SD	M	SD
Format	-0.47	1.00	0.12	0.90	0.02	0.53	0.22	0.79
Corpus	-0.19	0.93	0.01	0.99	-0.57	1.49	0.53	0.84
Order	0.42	0.89	0.15	0.77	-0.22	0.50	0.15	0.96
Task	-0.56	0.88	-0.33	0.74	-0.23	0.47	0.22	0.70

	For linguistic research ( <i>n</i> = 63)		For socio-cult. research ( <i>n</i> = 8)		For lexicography ( <i>n</i> = 6)		For teaching: indirectly ( <i>n</i> = 29)	
	M	SD	M	SD	M	SD	M	SD
Format	0.17	0.98	0.04	0.88	-0.28	0.96	-0.06	0.78
Corpus	-0.13	0.94	-0.56	0.83	0.13	0.67	0.24	0.82
Order	0.07	1.00	-0.32	0.70	-0.46	1.52	-0.81	0.83
Task	0.21	0.78	0.06	1.06	0.03	0.61	0.07	0.55

The samples were sufficiently large for the assumption of multivariate normality to be met, and the assumption that population covariance matrices were equal was supported by testing Box's M (60.42,  $p = .972$ ). MANOVA showed that concordancing preferences were related to *field of use*, Pillai's  $F(28, 178) = 1.80$ ,  $p .006$ . A direct discriminant analysis was carried out using the four factor scores as predictors of *field of use* to determine which of these variables

best discriminate between the different fields of use. Four discriminant functions were calculated, explaining 55.7 per cent, 30.8 per cent, 9.0 per cent and 4.4 per cent of the variance, respectively (Table 35). Wilk's lamda was significant for the combined functions  $X^2(28, N = 178) = 50.04, p = .006$ , but was not significant when the first function was removed  $X^2(18, N = 178) = 22.72, p = .202$  (Table 36). The first discriminant function maximally discriminated thinking of using a concordance *for translation, as a reference resource* and *for language learning* from thinking of using a concordance *for teaching: directly* and *for linguistic research* (Table 37). It correlated most strongly with respondents' scores on *use* preferences (.93) and *format* preferences (.54), and to a lesser extent with *corpus* preferences (.30) (Table 38).

*Table 35 Discriminant function analysis eigenvalues: Field of use*

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.17	55.7	55.7	.38
2	.10	30.8	86.5	.30
3	.03	9.0	95.6	.16
4	.01	4.4	100.0	.12

*Table 36 Discriminant function analysis Wilks' Lamda: Field of use*

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 4	.75	50.04	28	.006
2 through 4	.88	22.72	18	.202
3 through 4	.96	7.08	10	.717
4	.99	2.35	4	.672



Table 37 Discriminant function analysis functions at group centroids: Field of use

Field of use	Function 1
for teaching: directly	.41
for linguistic research	.28
for teaching: indirectly	.10
for socio-cultural research	.02
for lexicography	-.02
as a reference resource	-.33
to learn languages	-.39
for translation	-.83

Table 38 Discriminant function analysis structure matrix: Field of use

Factors	Function 1
Task	.93
Format	.54
Corpus	.30
Order	.07

## Discussion

Factor scores for users concordancing preferences (*format*, *corpus*, *order* and *use*) were analysed in relation to each of the independent variables in the study: how long respondents have been using concordances for (*length of use*), how frequently they use concordances (*frequency of use*), their linguistic relationship to the concordance (*language of use*), and the field of work or study they use concordances within (*field of use*). MANOVA tests identified statically significant patterns of variance for *frequency of use*, *language of use* and *field of use*, and each was then further analysed through discriminant function analysis. The results for each independent variable are discussed in turn below.

### **Length of use**

The MANOVA test for the dependent variables in relation to length of use did not find statistically significant patterns of variation in the data. This could possibly be interpreted as standing in stark contrast to the body of literature that claims that difficulty using concordances is a major barrier to use of concordances in pedagogical contexts (Breyer, 2011; Chambers & O'Sullivan, 2004; Leńko-Szymańska, 2014; Tribble, 2015). On the assumption that length of concordance use is a good indicator of how proficient a user is, it might be expected that there would be a marked contrast in concordancing preferences between users with different scores for *length of use*, but this was not the case. Of course, an alternative explanation could simply be that inexperienced users are yet to develop any clear concordancing preferences, and the preferences of more experienced concordance users are determined by factors other than familiarity. The absence of a pattern of concordancing preferences in relation to length of use does not contradict the observation that inexperienced users find concordances difficult to use; only that concordancing preferences do not reflect how familiar with concordancing a user is (in this sample of concordance users at least). In many ways, this finding provides further indication that there is not a prototypical form of advanced concordance use that concordance users progress towards as they become more experienced, suggesting perhaps that concordancing preferences are indeed determined by the affordances provided by different operationalisations of concordancing.

### **Frequency of use**

MANOVA detected a significant pattern of correlation between frequency of concordance use and the dependent variables, Pillai's  $F(12, 180) = 2.87, p .001$ . In a discriminate function analysis, the means of group centroids for this function indicated that the concordancing preferences of the least frequent concordance users (*very rarely*,  $-.73$ ) differed in significant ways from those of the more frequent users (*quite often*,  $.23$  and *very often*  $.33$ ). The structure matrix indicated that this difference pertained to respondents' scores on *use* preferences ( $.79$ ) and *format* preferences ( $.52$ ). In other words, respondents who very rarely used concordances were very strongly disinclined to reading concordances for repetitions and strongly disinclined to use concordances in a KWIC style format, while the more frequent users were comparatively likely to be positively inclined towards both these types of concordance use.

It is important to avoid any assumptions about causality in interpreting these results. It seems likely that the correlation between formatting preferences and use preferences in relation to frequency of use are related to the notion of vertical reading, but the direction of causality cannot be deduced. It could be that more frequent use of concordances inclines users towards the KWIC format because of its purported benefits to identifying patterns of repetition, but it could equally be that concordances are particularly useful for finding patterns, especially in the KWIC format, and so concordance users who are interested in finding patterns use concordances more often than users with other interests. It does however seem to validate to some extent the observation that KWIC format concordances are particularly useful for identifying recurring patterns in text (Barlow, 2004; Gilquin & Granger, 2010; Hunston, 2002, 2010; Sinclair, 1991; Tribble, 2010). In sum, frequent concordance users have a slight preference for this use of concordances in this format, while infrequent users are strongly disinclined to use concordances in this way or in the KWIC format. This has implications for the place of concordances in wider language learning programmes (Boulton, 2009b; C. Kennedy & Miceli, 2010; Pérez-Paredes, 2010; Tono et al., 2014) and raises questions about which formats and task learners might most profitably engage in if they are unlikely to be using concordances very frequently. If infrequent users are characterised by a dislike for looking for repetition and KWIC format concordances, and language learners are not encouraged to be using concordances as a central platform for language learning, then these results suggest that concordancing formatting and task defaults for language learners might be best targeted at operationalisations of concordancing that encourage looking for relevant examples in concordances that present sentence format citations.

### **Language of use**

In the final analysis, MANOVA detected a significant pattern of correlation between language of concordance use and the dependent variables, Pillai's  $F(28, 181) = 3.13, p .002$ . In the discriminant function analysis performed, the means of group centroids indicated that the concordancing preference *bilingual concordance use* (-1.29) differed very significantly from *1st language* (-.00) and *2nd language concordance use* (.23). Unsurprisingly, the structure matrix indicated that this difference pertained to respondents' scores on *format* preferences (.89), as well as *order* preferences (.51).

The extremely pronounced negative loading of bilingual concordance use, the main contrast with the other categories of 1<sup>st</sup> and 2<sup>nd</sup> language concordance use, is not surprising given that bilingual concordances emphasise comparison between citations from different corpora, not between citations within a concordance of a single corpus. As discussed in Chapter 2, ordering citations alphabetically by positions in a concordance's cotext is thought to highlight patterns of co-occurrence within a single concordance (Aston, 2001; Charles, 2012; Hunston, 2010; Johns, 1986, 1991b; Levy, 1990; Partington, 2001; Sinclair, 1991, 2004; Tono et al., 2014; Tribble, 2010), not to facilitate comparison of two separate concordances as presented in a bilingual concordance. Furthermore, because translation equivalents are often realised at the sentence or discourse level and not always at the word level (Kenning, 2010; Kübler & Aston, 2010), it is not surprising that bilingual concordance users might prefer non-KWIC citation formats that present larger spans of cotext.

It is perhaps less obvious that there should be no clear distinction between the preferences of 1<sup>st</sup> and 2<sup>nd</sup> language concordance users: the 2<sup>nd</sup> language concordance users in this sample only displayed a marginally higher likelihood of preferring KWIC format concordances and citations ordered by cotext than the 1<sup>st</sup> language users. If anything, we might have expected the scores to have been inverted, as KWIC format concordances are sometimes thought to be harder to authenticate than non-KWIC formats (Chapter 2). However, we should also note that we have no way of judging the language proficiency level of the respondents who answered from the perspective of 2<sup>nd</sup> language use. Given the biographical data reported in Chapter 3, it seems the majority of respondents were highly educated language workers in academic contexts and, as such, we might expect many of the respondents to have very high proficiency levels in their second languages and the descriptive data reported earlier in this chapter shows that only 37% of respondents reported using concordances for language learning (24% yes, 13% occasionally). This might account for the lack of any clear difference between 1<sup>st</sup> and 2<sup>nd</sup> language concordance users in this study. It may be that few of the respondents who completed the survey from an L2 user perspective would encounter the kind of authentication difficulties discussed in the literature because their proficiency levels may have been sufficient for them to overcome such difficulties. In retrospect, it would have been desirable for concordance users who answered from the perspective of 2<sup>nd</sup> language use to provide an indication of their proficiency level in the L2 they were thinking of, as this may have facilitated more confident interpretation of these findings.

### Field of use

MANOVA detected a significant pattern of correlation between frequency of concordance use and the dependent variables, Pillai's  $F(28, 178) = 1.80, p .006$ . The means of group centroids in the discriminant function analysis indicated that the concordancing preferences of thinking of using a concordance *for translation* (-.83), *for language learning* (-.39), and *as a reference resource* (-.33) differed in significant ways from those thinking of using a concordance *for teaching: directly* (.41) and *for linguistic research* (.28). Interestingly, the structure matrix indicated that this difference pertained to respondents' scores on *task* preferences (.93) and *format* preferences (.54) and, to a less pronounced extent, with *corpus* preferences (.30). Basically, concordance use *for translation*, *for language learning* and *as a reference resource* is indicative of being disinclined to reading concordances for repetitions, using concordances in the KWIC format and using concordances of large, general corpora; in contrast, for *teaching: directly* (i.e. where learners are presented with concordances) and, to some extent *for linguistic research*, are indicative of a general preference for these types of concordance use.

The alignment of results *for language learning* with those of *for translation* and *as a reference resource* indicates that the language learners surveyed are using concordances in a way more similar to translators or editors than linguistic researchers. Indeed, the correlations with *task*, *format* and *corpus* suggest that the corpus linguistic, pattern identifying method of concordance use is not typical of the five learners surveyed, but it is typical of the teachers answering the survey from the perspective of using concordances directly with learners. In terms of the discussion of affordances presented in Chapter 2, while teachers appear keen to use concordances in researcher like fashion with their learners, the small number of learners this survey collected responses from were not. A few more points about this seemingly incongruous result can be made.

First, it should be remembered that a proportion of those answering from the perspective of *for teaching: directly* may not have had language learning in mind (see *Field of use as a variable* above). However, it does seem likely that the majority of responses will have been informed by the experience of using concordances directly in language teaching because approximately 84% of respondents who indicated they had used concordances in teaching indicated they had used concordances in language teaching.

Second, these findings bear comparison with the findings of a study of trainee teachers reported in Breyer (2011). Examining the data she reports, it seems that the trainee teachers in her study were relatively unlikely to recommend using concordances as a writing resource to learners (preferring more traditional grammar or vocabulary focused DDL activities), despite the fact that they themselves were very keen to use concordances as resources for their own writing (Breyer, 2011). One might wonder why what's thought good for the goose is not thought good for the gander. This result also recalls research that introduced learners to guided research activities but found learners most impressed with the potential of concordances as a reference resource to be used while editing their written work (Chambers & O'Sullivan, 2004; H. Yoon & Hirvela, 2004). On a slightly more optimistic note, we might add to this list the change of perspective from *learner as apprentice* to *learner as chef* reported by C. Kennedy and Miceli (2001, 2010); these results lend support to such a change of perspective.

Third, while these results are certainly provocative, a word of caution is also in order. Not only is the sample of learners rather small (five), learners may not always be the best judges of pedagogical effectiveness: learner preferences are not always the best index of pedagogical value. However, these results do suggest it might be worth exploring the extent to which models of concordance use that are not based around KWIC format concordances, looking for repetition, and concordancing large general corpora might have more immediate value to learners than the guided research type exercises that have tended to dominate the DDL literature until quite recently (see Chapter 2). For instance, we might consider a model of learner concordance use that reflects the preferences of concordances users who are using concordances for translation or as a reference resource.

### **Limitations**

As discussed in Chapter 3, the small scale of the survey data obtained and the sampling frame used represent obstacles to the reliability and generalisability of the results obtained in this study. A second concern is the desirability of revising some of the core mechanics of the survey instrument. In particular, it was apparent from text entry answers that some respondents who use concordances for discourse analysis categorised themselves as using concordances for linguistic research while others categorised themselves as using concordances for socio-cultural research. Indeed, the category of *for linguistic research* was not only the largest response category for field

of use but also appeared the most heterogeneous in terms of the indications respondents gave as to their field of research. The attempt to restrict respondents' responses to a limited context of use was only partially successful. Furthermore, in retrospect, it would have been useful to have data on the proficiency level of responses from the perspective of L2 concordance use and the type of teaching being thought of in relation to field of use. Of course, the danger of taking such an approach in this kind of macro-level quasi-experimental survey research is that it would fragment the pool of responses obtained to such an extent that the power or statistical analyses would be reduced and the reliability of findings jeopardised. Hence, in many respects, these kinds of issues are part and parcel of survey research. The survey was moderately successful from the point of view of identifying reliable relationships between some predictor variables and some outcome variables, quantitative questionnaire research is a macro-level research design that will necessarily gloss over many individual differences (Punch, 2003). Indeed, as Punch remarks, low-stake small-scale surveys of this kind are best thought of as making contributions to a field of study than as presenting definitive answers. As such, the results obtained in this study could be usefully followed up through data from a more naturalistic methodology, such as interviews with concordance users or observation of concordance use; triangulation with such findings would provide a much stronger foundation for interpreting the findings reported here.

### Summary

We have seen in this chapter that the concordancing preferences of users can be predicted by a number of different variables: *frequency of use*, *language of use* and *field of use*. It is interesting that *frequency of use* and *field of use* were both predicted by the same two dependent variables: *task preferences* and *format preferences*. Together, these results provide support for the literature driven analysis of the affordances of concordances presented in Chapter 2: reference type use of concordances is distinguishable from research like use of concordances on the basis of the relevance/repetition distinction and the KWIC/non-KWIC distinction, with the preferences of learners in this study similar to the preferences of people using concordances for translation or as a reference resource. The results for *language of use* were in some respects disappointing, dominated as they were by rather obvious features of bilingual concordance use: bilingual concordance users have little use for KWIC formatting or ordering citations by cotext, but then, it did not seem at all likely that they would. On the other hand, it is interesting that none of the

independent variables tested identified corpus type as the main type of preference distinguishing different types of user. In the next two chapters, concordance type is explored in relation to measures of authentication burden. The results reported in Chapters 5 and 6 are then triangulated with those of Chapters 3 and 4 in Chapter 7.



## Chapter 5. Exploring the authentication burden of concordances through word frequency measures

### Introduction

Concordances can be viewed as a type of text (Chapter 2) and, as such, in the context of learner use, it is important to evaluate the kinds of textual demands concordances place on users as readers. Indeed, the linguistic difficulty of concordances is a major concern in the literature (Boulton, 2009b; Braun, 2005, 2006; Breyer, 2009; Charles, 2012; Cobb, 1999; Daskalovska, 2013; Flowerdew, 2012a; Franken, 2014; Frankenberg-Garcia, 2014b; Gordani, 2013; Kaur & Hegelheimer, 2005; C. Kennedy & Miceli, 2010; O’Keeffe et al., 2007; Osbourne, 2004; Römer, 2011; Thurston & Candlin, 1998; Tono et al., 2014). While it should be recognised that not all learner corpus use depends on learners being able to read citations in their entirety (Sripicharn, 2010), some of the affordances offered learners by concordancers may involve reading citations in a fairly traditional sense, and even the more concentrated analysis of a search word’s immediate context that characterises vertical reading is surely facilitated by a basic level of word recognition. We might also note that the results reported in Chapter 4 indicate that not all concordance users are using concordances in ways that are congruent with vertical reading, and so it appears likely that a proportion of concordance users will need to be able to process concordances in a manner that draws heavily upon more traditional reading skills. Indeed, survey respondents who completed the survey from the perspective of using concordances *for language learning* were one of the categories of survey respondent who were found to have concordancing preferences congruent with traditional reading of concordances.

In this chapter, concordances are analysed from a vocabulary profiling perspective. It is argued that concordances composed of higher frequency words will present less of a vocabulary burden on concordance users, and so less of an authentication burden as well, all other factors being equal. As discussed in Chapter 1, the *language as fixed effect fallacy* (Clark, 1973) would seem just as relevant to the language presented in concordances as the language presented in experimental prompts, what we might think of as *the concordance as fixed effect fallacy*. The study reported in this chapter starts off by testing the hypothesis that the language contained in concordances will differ significantly on a key index of readability: word frequency. Finding

differences on such a measure is clear evidence that concordances should be treated as random effects, not fixed. It then moves on to test concordances from the three different authentic corpora against concordances of graded readers. The results provide an indication of both the extent of variation between different types of authentic concordance on measures of word frequency, as well as the extent of such variation vis-à-vis concordances of graded materials. This is important because it has been suggested in the literature that graded readers may represent an effective way of lowering the authentication burden of concordances.

## **Literature review**

In this section, existing research on the linguistic difficulty of concordances is reviewed. The common thread in the studies reviewed is the significance of word frequency measures in assessing the authentication burden of concordances. Word frequency as a gauge of readability is then briefly discussed. The final subsection reviews existing research on word frequency as a measure of vocabulary load and explains why results pertaining to word frequency measures of whole texts are not optimal for measuring word frequencies in the context of concordances, vocabulary load and authentication burden.

### **Existing approaches to concordances as difficult texts: grading outputs and inputs**

The issue of the authentication burden of concordances is reflected in two main streams of research: research on pedagogical filtering of concordance outputs (citations) and research on the grading of concordance inputs (the corpora to be concordanced). Each is discussed below.

A number of researchers have explored pedagogical filtering of concordance outputs as a means of improving the readability of the citations presented to learners: that is, the extent to which concordance outputs can be graded so as to present learners with concordances that present a lower authentication burden than non-graded outputs. Wible, Chien, Kuo, and Wang (2002) explored using a lexical difficulty filter. In their system, on the principle that the words that occur more frequently in a language present less difficulty than less frequent ones, the user sets a word frequency threshold that reflects their level of vocabulary knowledge; the filter then prioritises display of citations that contain a higher proportion of words that are ranked below

this threshold. However, they note that their approach still has the potential to “overwhelm learners with examples that are too difficult to be useful” [p. 148]. Their example data illustrates this point. Their example data consists of 50 sentence length citations from the BNC, but only four of these 50 citations passed the lexical difficulty threshold they set. Kilgarriff et al. (2008) report on a similar project called GDEX or, *the Good Dictionary Examples project*. GDEX is an algorithm that was originally developed to automatically extract ‘good’ citations that could be used as examples in dictionaries, but it has also been used within the pedagogically orientated *Sketch Engine for Language Learning* (SKELL) concordancer (Kilgarriff, Marcowitz, Smith, & Thomas, 2015). As with the lexical difficulty filter discussed above, GDEX uses word frequency data to score citations, but it also uses a range of other measures as well: sentence length, inter-sentential reference, collocate frequency and collocate position. However, the conclusions of the GDEX project are comparable to those of Wible et al. (2002): few citations from a corpus of naturally occurring text meet the criterion of easy readability that their project imposed. This is also the conclusion reached by another group of researchers aiming to develop a corpus based tool for remedial learners in an EFL context (Chujo et al., 2015).

A second approach to limiting the difficulty of concordance data is represented by grading the texts in the corpus to be concordanced: that is, lowering authentication burden by grading the input. Allan (2009) examines a corpus of graded reading materials in order to establish whether or not such a corpus can provide learners authentic examples of language use; she compares word clusters found in the graded corpus with clusters observed in a portion of the BNC, and finds that many of the clusters observed in the graded corpus also occur in the authentic corpus. On this basis, she argues that graded corpora can be a useful resource for learners because they present a lower vocabulary load than an authentic corpus but still present patterns of co-occurrence that are attested in authentic corpora. Her (2010) study reports on learners using concordances of a graded corpus to learn vocabulary. However, while this is a useful approach for some uses of concordances in language learning, there are potential drawbacks to using graded corpora as well. First, even without considering the additional cost of producing a graded-English for specific purposes (ESP) corpus, Wible et al. (2002) see the cost of creating and maintaining ungraded-ESP corpora as an argument in favour of developing citation filtration techniques for exploiting large general corpora with learners. This observation suggests that the costs involved in preparing graded texts are likely to conflict with any

affordances of concordances that are based on access to specific varieties of a language (see Chapter 2). Second, the majority of graded texts that are already available are fiction. Although non-fiction graded readers are available (Nation, 2013), non-fiction texts make up a very small proportion of published graded reading materials (Hill, 2008). In other words, although using concordances of graded readers would appear to circumvent some of the linguistic challenges of concordance use, the variety of language they tend to represent contrasts with many of the language varieties commonly addressed in studies of concordance-based learning: general English (Boulton, 2010a; Gordani, 2013; Kaur & Hegelheimer, 2005; Koosha & Jafarpour, 2006; H. Yoon, 2008), newspapers (Allan, 2006; Boulton, 2009b; Chan & Liou, 2005) and academic English (Charles, 2012, 2015; Lee & Swales, 2006).

For the reasons discussed above, it seems that neither filtering citations from authentic corpora (grading outputs) nor using concordances of already graded texts (grading inputs) can easily circumvent the issue of authentication burden while also maintaining the quantity and quality of citations that furnish the language learning affordances of concordances (Chapter 2). Consequently, it is important that research examines the authentication burden of concordances of authentic texts and other potential means of pedagogical mediation. Following the studies discussed above, the study reported in this chapter sees a useful role for measures of word frequency in pursuing such a project. As we will see in the next subsection, there is good reason to treat word frequency measures as useful indices of readability, and therefore also as an important component of authentication burden.

### **Word frequency as a gauge of readability**

It is not surprising that measures of word frequency have featured strongly in initiatives to lower the authentication burden of concordances. First, psycholinguistic studies of reading that track eye movements have found word frequency to be one of the strongest and most robust predictors of word processing time (Rayner, 1998, 2009): all things being equal, the more frequent a word is, the more fluently it is processed. Second, research on L1 and L2 reading development has also found rapid and unconscious word recognition to be a fundamental requirement of fluent reading comprehension (Grabe & Stoller, 2001; Nation, 2013). For instance, studies of second language vocabulary acquisition and reading comprehension have found that knowledge of a higher percentage of words in a text correlates with better comprehension of the text (Grabe &

Stoller, 2001; Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2013; Schmitt et al., 2011). Thus, if the frequency of a word is a significant and robust predictor of word recognition speed, and having sufficient vocabulary to quickly recognise the words in a text is a prerequisite of reading comprehension, then the frequency of words in a text in conjunction with estimates of a reader's vocabulary knowledge provides a useful predictor of reading comprehension. The general consensus appears to be that knowledge of around 95% of the words in a text provides a sufficient basis for gaining some comprehension of a text, but nearer 98% is optimal (Nation, 2013).

Of course, this is not to say that recognising the majority of the words in a text is a sufficient condition for comprehension, there are numerous other factors that can impinge on reading (Grabe & Stoller, 2001; Nation, 2013; Rayner, 1998, 2009); rather, it is to suggest that knowledge of the majority of words in a text is in practice a necessary condition for fluent reading comprehension: necessary, but not necessarily sufficient. Grabe and Stoller (2001, p. 19) illustrate the necessity of word knowledge neatly:

Anyone who has tried to read a text on political policy written in Chinese – without knowing any Chinese characters – will quickly recognise the primacy of linguistic processes for reading comprehension. If we cannot understand any words, we are not going to comprehend the text.

Consequently, research on word frequency, text coverage and reading comprehension has developed alongside research on how large a vocabulary a learner needs to be able to recognise a particular percentage of words in a text (Nation, 2006, 2013): how many of the most frequent words a learner needs to know to be able to recognise 95% of the words in a text (a ratio of 19:1 known to unknown) or 98% (a ratio of 49:1). Whichever percentile figure is used, the number of words required to provide this coverage can be referred to as the texts *vocabulary coverage level*: for instance, if you need to know the 4,000 most frequent words in a language to be able to recognise 95% of the words in a text, the text can be said to have a vocabulary coverage level of 4,000 at the 95% coverage level. The next section discusses existing research using such figures.

### **Research on texts, word frequencies and vocabulary load**

Corpus approaches to text have provided useful estimates of vocabulary coverage figures in relation to word frequencies. If each word in a text is counted as a *token*, but only each

orthographically different word is counted as a *word type*, then the question can be formulated as *how many word types provide coverage of how many word tokens*. In a review of corpus-based descriptions of English, G. Kennedy (1998) summarises results from a number of corpus studies in a table showing text coverage in relation to the minimum number of word types able to provide it (see Table 39 below).

*Table 39 Approximate proportions of different word types at different frequency levels in texts (G. Kennedy, 1998, p. 96)*

% of total word tokens in text	No. of different word types
95	3,000 – 15,000
90	2,000 – 3,000
80	1,000 – 1,500
70	600 – 800
50	50 – 100
25	10 – 15

He also notes that the same number of word types in a spoken corpus can account for a significantly higher coverage figure than an equivalent number of words in a written corpus, and that coverage of larger, broader corpora will have a more consistent range of words than that provided by smaller, more narrowly focused corpora. Nation (2006), approaching the coverage issue from a pedagogical perspective, reformulates the question of vocabulary coverage in two important ways: using a text independent list of words to determine coverage levels, and using *word families* as the unit of counting.

Following Graeme Kennedy's observations that the most frequent words in texts vary depending on the types of text examined but that larger more general corpora provide more robust lists of words, Nation uses frequency lists based on large publically available corpora to assess the vocabulary coverage levels of different texts (see for example Nation, 2006). Vocabulary coverage levels based on such text external word frequency estimates can be considered good estimates of the vocabulary load of text, vocabulary load being an estimate of the amount of vocabulary knowledge that would be required to provide coverage of a text. To exemplify the reasoning, if a text only required 3,000 words to provide a particular percentage of vocabulary coverage, but the most common 8,000 words was required to provide coverage of these particular 3,000 words, then the vocabulary load of the text would be best considered 8,000

words because a learner would likely need to know these 8,000 words to have those 3,000 specific words available to them. As noted above, this notion of vocabulary coverage in relation to vocabulary load can be referred to as *vocabulary coverage level*, the level of vocabulary needed to provide vocabulary coverage of a text.

Nation uses the pedagogically motivated concept of *word families* as his unit of counting. Word families are groups of word types that constitute morphological variations on a shared base word form wherein it can be assumed that recognition of one type will facilitate recognition of another (Bauer & Nation, 1993; Biemiller & Slonim, 2001; Nation, 2006, 2013, 2015). Word families are discussed in more depth in the methodology section below but, thinking for now only of his results, Nation (2006) provides vocabulary coverage levels in word families for a range of different text types including novels, newspapers, graded readers, children's movies and unscripted spoken English. He concludes that a reader would need to know an average of between 8,000 and 9,000 word families to recognise 98% of the words in written texts, and between 6,000 and 7,000 for spoken texts. The discrepancy between his results and those reported in G. Kennedy (1998) is explained by Kennedy's figures being derived from the frequencies of the words as they appear in the particular texts assessed, and Nation's being derived from vocabulary coverage levels, as defined above.

The figures reported in Kennedy and Nation cannot, however, be directly transferred to word frequencies in concordances. Their figures are based on the analysis of whole texts, but concordances are essentially fragmented texts. Consequently, coverage figures based on whole texts may be misleading when applied to concordances. As Tognini-Bonelli (2010) argues, while a typical text is read whole and constitutes a coherent communicative event, a concordance is a fragmentary text and, as such, does not constitute a coherent communicative event. As a consequence, from a concordance user perspective, it is best to evaluate each citation in a concordance as, in effect, a separate text because, unlike continuous text, unknown vocabulary in one citation should have little if any effect on the processing of another. This also means that it is not necessary for a user to be able to meet the aggregate coverage requirements of every citation in a concordance to make use of it. For this reason, while vocabulary coverage level figures for whole texts are quite straightforward, such figures for concordances are better based on averaged coverage figures for the citations they contain. Analysing vocabulary coverage levels for concordances this way better represents their actual use.

## Research questions

This study explores measures of word frequency in citations generated through concordancing. First, it examines word frequency measures in relation to citations from three different authentic corpora, testing the hypothesis that concordances of different corpora present learners with different vocabulary loads. As such, the first research question is:

1. Are there statistically significant differences between scores on measures of word frequency for citations from three different authentic corpora?

The study then compares the data from these three corpora of authentic materials with a corpus of graded readers, estimating the scale of vocabulary load differences between both the authentic corpora themselves as well as between each of the corpora of authentic materials and the graded corpus. That is,

2. What is the magnitude of differences in word frequency measures between citations from the different authentic corpora?
3. What is the magnitude of differences in word frequency measures between citations from the three authentic corpora as compared with word frequency measures of citations from a corpus of graded readers?

## Methodology

This study takes a quantitative approach to evaluating the vocabulary load of concordances, an important component of a concordance's authentication burden. To estimate the vocabulary load of concordances, two measures of vocabulary load based on word frequency level were developed. Both measures of word frequency were based on word frequency lists that provide estimates of how frequently word family members occur in general usage. The first measure operationalised vocabulary load as the number of word families a user would need to know to be



able to recognise 98% of the words in a citation: this measure is termed *vocabulary coverage level* (VCL). The second measure operationalised vocabulary load as the mean average word frequency level of the words in a citation: this measure is termed *mean word frequency* (MWF). To test the hypothesis that concordances generated from different corpora present users with different vocabulary loads, both measures are tested in four corpora of similar size and composition which were developed for this purpose: a corpus of academic prose, a corpus of news, a corpus of fiction and a corpus of graded readers.

### **Developing word frequency lists**

Word frequency lists assign orthographic forms relative frequency ranks on the basis of how frequently they occur in a given language or sub-language: the form that occurs most frequently, second most frequently, 100<sup>th</sup> most frequently, 1000<sup>th</sup> most frequently, and so on. These individual rankings can then be grouped into rank-bands such as the most frequent thousand (1k), the second most frequent thousand (2k), the tenth most frequent thousand (10k), and so forth. Paul Nation has advocated using the *word family* as the unit of ranking if the word frequency list is to be used in relation to receptive vocabulary knowledge (Nation, 2006, 2013, 2015; Nation & Waring, 1997). The idea of a *word family* is presented in Bauer and Nation (1993):

From the point of view of reading, a word family consists of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately. So, watch, watches, watched, and watching may all be members of the same word family for a learner with a command of the inflectional suffixes of English. As a learner's knowledge of affixation develops, the size of the word family increases. The important principle behind the idea of a word family is that once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort.

[253]

Nation has produced a series of word frequency lists based on the word family: for instance, BNC-14, a list of the 14,000 most common word families in British English (Nation, 2006). His latest and most comprehensive list is BNC-COCA-25 (Nation & Webb, 2011). The BNC-COCA-25 lists present 25,000 word families and are based on the two largest publically available non-

web-based corpora, the 100-million-word *British National Corpus* (BNC) and the 450-million-word *Corpus of Contemporary American English* (COCA). It is significant that these two corpora are designed to be as representative of as wide a range of English usage as possible; Davies (2015) has shown that larger web-based corpora do not reflect many varieties of English adequately. While the BNC and COCA attempt to present as representative a balance of text types as possible, web-based mega corpora primarily consist of web pages. As such, the BNC-COCA-25 lists can be seen as both the most comprehensive and most representative lists of word frequency in English in general that are available at this time.

Although the BNC-COCA-25 lists were broadly suitable for use in this study, analysing text at the citation level presented unusual challenges to the effective use of the lists due to the very small size of the texts to be assessed: around 16 words per citation. Piloting indicated that the distribution of orthographic forms that were not included on the lists (off-list word types) was uneven. As text segment length increased, the standard deviation of the occurrence of off-list word types decreased and, conversely, as text segment length decreased, SD increased. This observation is in line with previous research on text length, sample size and text coverage (Chujo & Utiyama, 2005). Hence, while it is probably safe to assume that, for large text samples, the occurrence of off-list words is likely to approach a ratio approaching their distribution in the text category that they are intended to represent, off-list words represented a problematic source of variation in this study because of the extremely short text lengths examined. Consequently, it was necessary to expand and supplement the coverage of the BNC-COCA-25 lists to reduce the number of off-list word types in the data. To this end, this study supplemented the BNC-COCA-25 lists with 26k, 27k and 28k word family lists that are currently under development but were kindly made available by Paul Nation: while these lists have not been fully familiarised, they do list 3,000 more word types as headwords and thereby expand the coverage of the BNC-COCA-25. Nevertheless, from the perspective of assigning a frequency rank to as many word types in a text as possible, there remained two main problems with the BNC-COCA-25 lists: the word family level used to delineate a word family, and four categories of off-list word type that are not integrated with the word frequency lists, word types such as *William*, *weekend*, *urgh* and *NSAIDs*.

### *Expanding word family level*

In relation to the first issue, word family level, this study expanded the BNC-COCA-25 word lists so as to include more word types within the 25,000 word families listed. Word families are intended to reflect the way in which morphological knowledge facilitates comprehension of one word type on the basis of knowledge of another word type that shares a base word form (Bauer & Nation, 1993). In effect, the notion that the frequency of one word type is transferable to the frequency of another recognisably related word type. However, there are a number of different ways in which morphological knowledge can be seen as having the potential to facilitate such recognition, and it appears that the ability to make use of such knowledge develops over time (Bauer & Nation, 1993). Consequently, word families can be defined at different levels, with word family levels progressing additively through successively broader morphological criterion for inclusion of word types within a word family:

Level 1: Each form is a different word

Level 2: Inflectional suffixes

Level 3: The most frequent and regular derivational affixes

Level 4: Frequent, orthographically regular affixes

Level 5: Regular but infrequent affixes

Level 6: Frequent but irregular affixes

Level 7: Classical roots and affixes

Further affixes

(Bauer & Nation, 1993, pp. 257-263)

Nation's lists initially demarcated word families at Level 6. Consequently, word types containing affixes at Level 7 and beyond were not included within the word family of their headword but formed the headword of word families of their own. Furthermore, even though the criterion for word family membership is relaxed for lower frequency word family lists, "on the assumption that learners at higher proficiency levels will know a great number of affixes and will have greater skill at recognizing the same bound root in different words" (Nation & Webb, 2011, p. 137), there are many word forms that are not included on the BNC-COCA-25 lists, even though their corresponding headword does. When trying to assign frequency ranks to as many word

types as possible, such omissions are problematic in relation to the principle of recognisability that underlies the idea of the word family.

The fundamental problem is that the frequency count for a single word type cannot be compared legitimately with that of a word family, making it difficult to assign credible rankings to the unusual derivatives of common headwords that have not been listed within a word family. Ranking such words as less frequent than the lower frequency words of the BNC-COCA-25 is implausible as their degrees of recognisability are often incomparable: for example, *disapplied* seems far more reasonably ranked as part of the higher frequency word family *applied* than alongside lower frequency word families such as *apodictic*, *balletomane* or *chumash*. Consequently, such errant off-list word types were appended to the most appropriate word family available by broadening the criterion for inclusion within a word family as far as possible. In addition, many previously unlisted word family members at Level 6 or below were also added. Examples of word types that were added to existing word families include *greened*, *unloving* and *epeeists*.

The resulting lists do include some complex or unusual word types within high frequency word families, for instance *alcoholised* and *decasualisation* join word families for *alcohol* and *casual* at the 2k level. However, for the purposes of this study, this is preferable to presenting them as novel word families at word frequency levels that grossly misrepresent their recognisability. However, as a consequence, results derived from these adapted lists may not be suitable in relation to language learners who only know one or two thousand word families because not all such learners are likely to have sufficient morphological knowledge to be able to decode all of the word types now scored at these levels; the expanded lists represent word family frequency rankings that assume fairly well-developed morphological knowledge.

### ***Integrating four categories of off-list word type***

The second problem with the coverage of the BNC-COCA-25 was the exclusion of four categories of word type from the frequency lists: transparent compound words (e.g. *fulltime*, *drugstore* or *childbirth*), marginal words (e.g. *ehm*, *brr* or *psst*), proper nouns (e.g. *Dinah*, *John* or *Dan*) and abbreviations (e.g. *adj.*, *lb*, or *BMJ*). As suggested by Nation and Webb (2011, p. 138), transparent compounds were listed under their least frequent component part (so *lifespan* would be assigned the same frequency rank as *span*). On the principle of familial recognisability,

and on the assumption that knowing both component words in a transparent compound is sufficient for recognition of that word type, the frequency level of its lowest ranked component seems to best reflect its frequency level in relation to word family lists. The other three categories of word type were dealt with by assigning them de facto word frequency scores, as discussed under scoring below. Consequently, word types found in the data conforming to any of these three categorises were also recorded.

### ***The expanded lists***

Following the procedures outlined above, approximately thirteen thousand new word types were assigned a frequency ranking. This allowed the BNC-COCA-25 lists to provide maximal coverage of the word types found in the data, resulting in well over 99% of word types being identified as either a member of one of 28,000 word families, or else a marginal word, proper noun, or abbreviation. Of the small number of words that remained unlisted, obvious typographical errors in the texts were corrected where it was obvious what the intended word or words were. The words that then remained unlisted tended to be either non-English words that had not been listed in word families previously or else novel representations of spoken dialect that were not represented in online dictionaries.

### **Developing corpora**

Four corpora of comparable size and composition were developed for this study. Three authentic corpora were developed by creating subcorpora from the British National Corpus (BNC), one of the best known and most popular corpora available (Tribble, 2015). The fourth was compiled from Tom Cobb's graded reader corpus texts (available for research purposes from <http://www.lex tutor.ca/conc/graded/>).

The BNC was used to generate three comparable subcorpora for three main reasons. First, the BNC is well known, publically available, and has been used in many studies of concordance-based language learning (see for instance Boulton, 2010a; Gordani, 2013; Kaur & Hegelheimer, 2005). Second, it is also extremely well-documented and has been very influential on the development of other corpora (Burnard, 2002). Third, Lee (2002) has explored the classification of texts within the BNC and produced *the BNC Indexer* as an advanced tool for defining custom

subgenre within the BNC. For these reasons, the BNC represented a particularly useful and appropriate resource for compiling comparable subcorpora for use in this study.

The three subcorpora extracted from the BNC for this study are each drawn from what Lee designates single genres within single super-genres of the BNC (Lee, 2002). Recognising that terms such as *genre*, *register*, *text type*, *domain* and *style* are notoriously slippery terms, Lee embraces a concept of genre based on culturally recognisable groupings of texts that can be defined at varying degrees of specificity. Thus, Lee takes the existing BNC classification criteria of *domain* and *medium*, but then subcategorises them into 70 genre and 10 super-genre. This study selected a single genre from each of the main written super-genre: *academic prose*, *fiction*, and *news*. Spoken genre were avoided for two main reasons: first, word frequency measures are likely to be misleading when used with spoken corpora because of the preponderance of written materials in the generation of word frequency lists; second, spoken language is considerably more context dependent and phatic than written, which problematizes the relationship between the coverage and informativeness of words in spoken text (M. McCarthy & Carter, 1997).

Each of the three written corpora extracted from the BNC consisted of over two million words in length. The *fiction* and *academic prose* corpora were made by selecting approximately one million words from the largest files within each genre, and approximately one million words from the smallest files in each genre. The *news* genre used to create the news corpus was only slightly larger than these two corpora, and so was taken in its entirety. In addition to these three corpora extracted from the BNC, a graded reader corpus was also compiled. There were 108 texts in the graded reader corpus. The texts spanned six reading levels (300 headwords to 2800 headwords), three publishers (Oxford Bookworms series; Penguin Readers; Cambridge English Readers); and both fiction and non-fiction. This corpus was slightly smaller than the authentic corpora, totalling slightly more than 1.3 million words.

### **Sampling frame**

Developing an appropriate sampling frame for gathering data required balancing a number of competing demands: the quantity of data required for each analysis and the availability of data in relation to data processing times. These issues are discussed below in terms of the power analysis used to determine the appropriate sample size to use and the search terms selected to generate concordances for analysis.

### *Power analysis*

In determining an appropriate sampling frame for this study, it was necessary to ensure the sample size would have sufficient power to detect even rather low effect size figures. As Kilgarriff (2005) argues, statistical significance is not a particularly informative criterion of difference when using corpus data: words are not distributed randomly in texts and so, with a large enough data set, it will always be possible to demonstrate this with a test of statistical significance. As this study would also involve word distributions and the potential to access large quantities of the data, Kilgarriff's point would seem to apply to this study of concordances as well as to studies of corpora. Consequently, the important research question has less to do with whether a difference meets a particular level of statistical significance, and more to do with the magnitude of any differences detected. To this end, this study emphasises measures of effect size. While significance tests tell us the likelihood of an observed difference if there is actually no difference in the parent populations, effect sizes tell us the extent of a difference in distributions (Larson-Hall, 2010; Norris, 2015). However, as there is not an established criterion for judging effect sizes in relation to the measures used in this study, it was desirable to ensure that even numerically small effect sizes were likely to be detected. After all, conventional effect size descriptions are just that, conventions; "whether treatment effects are large or small is a relative matter which should be assessed in terms of some external criterion of practical or theoretical utility" (Winer et al., 1991, p. 121).

Before conducting the power analysis, piloting had shown that the natural distribution of the data was heavily skewed, as might be expected given the data is based on word frequency scores: word frequency follows a Zipfian type distribution wherein the rank order of frequencies follows an approximation of a power law vis-à-vis their absolute frequencies (Baayen, 2001; Sorell, 2015). Thus, the non-parametric Mann-Whitney U was identified as an appropriate statistical test to use, and an a priori power analysis for the Mann-Whitney U test was conducted. Power analysis used an alpha of .05, a power of .95, an effect size of  $d = .2$ , and two tails. Assuming an equal allocation of participants into each group, the minimum sample size required was 1364 in total.

### *Selection of search terms*

In generating a concordance, the two main variables involved are the corpus concordanced and the search term entered; therefore, to control for the effect of search term, it was necessary to ensure that a variety of search terms were used to generate data and that the search terms used to generate data were balanced across conditions as far as possible. In practical terms, this meant generating word lists for each of the corpora and only selecting search terms that occurred at a sufficient frequency rate in each. Based on the power analysis described above and the availability of search terms across corpora, the basic sampling frame to be used was set as 70 citations of 11 search terms per condition (770): a total sample size of 1540 when two of the conditions are compared in a Man-Whitney U test, and hence more than sufficient to achieve the desired power. The search words used in the first sample set are shown in Table 40 below. As it transpires, each word sampled is from the first 1000 word families of the BNC-COCA-25 lists; however, it should be noted that some changes to the sampling scheme had to be made when a sample was sought from a corpus of graded readers because of the limited availability of even these very high frequency words (see notes to Table 40 for details).

To be able to replicate the findings from the first sample set, a second set of search terms was sought. However, there were again limits as to the availability of search terms that occurred in sufficient numbers in all the corpora analysed; the second sample set aimed to sample words that were comparable to those in the first, and so avoided the highest frequency function words like *the* and *on*. As a consequence, the second sample set was only available on the provision that the number of citations per search term varied in the second citation set (see Table 41 below). For the corpus of graded readers, only a single sample set (that is, a sample set corresponding to Citation Set 1), was available.



*Table 40 Search terms selected for Citation Set 1: Academic prose, Fiction, News and Graded readers*

Search Terms	<i>n</i>
Ground	70
Idea	70
Information	70
Interest	70
Mistake	70
Movement	70 <sup>a</sup>
Papers	70
Prisoner	70
Relationship	70 <sup>a</sup>
Teacher	70
Terms / Time <sup>b</sup>	70
TOTAL	770

<sup>a</sup> For the graded reader corpus, it was necessary to supplement 49 citations of *movement* with 14 citations of *movements* and 7 citations of *move*; and 42 citations of *relationship* with 4 citations of *relationships* and 24 citations of *relations*.

<sup>b</sup> For the graded reader corpus, the search term *terms* was substituted for the search term *time*.

*Table 41 Search terms selected for Citation Set 2: Academic prose, Fiction, News*

Search Terms	<i>n</i>
Middle	71
Showed	73
Against	216
Wide	73
Finally	193
However	70
Among	74
TOTAL	770

### **Generating data**

To analyse data in the required quantities it was necessary that concordances could be generated with a high degree of automated processing and analysis. To facilitate this, a programme was developed in Python: Awesome-parser. Awesome-parser was able to run standard concordancing operations integrated with a number of important non-standard functionalities, primary among which are the following. First, the config file for Awesome-parser allowed specification of the number of citations required and random selection of citations from within the text file specified. Second, the number of characters of cotext to extract per citation could also be set in the config file. Third, the programme automatically rounded the number of characters extracted down to the nearest number of whole words contained within the span of cotext set (i.e. whole words being demarcated by white space characters). Fourth, on the basis of editable word frequency lists, the programme automatically scored each word in the cotext of each citation extracted. Fifth, it implemented advanced probabilistic proper noun handling to assign ambiguous words to either proper noun lists or common word lists (a combination of comparison with common word lists and checking for surrounding punctuation; cf. Cobb, 2010). Sixth, citations and scores were exported to .csv files, facilitating inspection of citations, words by score, words categorised as proper nouns and words categorised as off-list. Seventh, the programme automatically calculated vocabulary coverage levels for each citation at a vocabulary coverage percentage set in the config file, which was also exported to the .csv file.

Using Awesome-parser, it was possible for each citation generated in the study to be selected randomly from each corpus, extracted with all complete words within 100 characters of the search term, automatically assigned a vocabulary coverage level score and available for visual inspection along with relevant scoring information. Scores for mean word frequency were then calculated in Excel. Scoring procedures for both vocabulary coverage level and mean word frequency are described below.

### **Scoring**

Each citation sampled was scored on two measures: vocabulary coverage level (VCL) and mean word frequency (MWF). Both measures scored words on the basis of the frequency list they correspond with: a word that was included on the expanded 1k frequency list scored one, a word that was included on the expanded 2k frequency list scored two, and so forth, with words that are

included on the 28k<sup>h</sup> frequency list scoring 28. For words that could not be ranked in the expanded BNC-COCA-25, the following decisions were made. Word types that were identified as proper nouns (*Dan, Nanjing, Microsoft*), abbreviations (*Prof., CIA, ml.*) or marginal words (*uh, oh*) were assumed to be reasonably intelligible as such, and so were scored as being equivalent to 1k words. In contrast, off-list words that could not be placed on the proper noun, abbreviations or marginal words lists were given the score of 29, a notional frequency level above the highest frequency rank of 28 in the expanded BNC-COCA-25 lists.

### ***Vocabulary coverage level***

To determine the vocabulary coverage level score of each citation, each word in a citation other than the search word itself is scored in relation to its frequency level on the expanded BNC-COCA-25 lists. The VCL score for the citation is then the score of a predetermined coverage level, calculated as a percentile. This study calculated the 98<sup>th</sup> percentile for each citation: the frequency level that provides coverage of at least 98% of the words in the citation, search term excluded. For instance, a citation from a concordance of *relationship* might produce the following citation: *home. The girl is said to have a poor relationship with her father and the prospect that he.* Excluding the search word, *relationship*, the citations consists of sixteen 1k words (*home, the, girl, is, said, to, have, a, poor, with, her, father, and, the, that, he*) and a single 3k word (*prospect*). The 98<sup>th</sup> percentile for the scores of these 17 words is the score of the 17<sup>th</sup> word in rank order ( $0.98 \times 17 = 16.66 \approx 17$ ), and so this citation would be given a score of three, the score of the 17<sup>th</sup> word in rank order. Indeed, because of the length of the citations used in this study (100 characters rounded down to the nearest whole words), the score of each citation is in effect always the frequency level of the lowest ranked word in the citation. Not only are citations of this length fairly typical of concordance outputs but using citations of this length also reduced the effect of differences in the number of words per citation, a potentially unhelpful source of variance.

Vocabulary coverage level figures calculated in this manner have been widely cited and correspondingly influential in second language acquisition research (see for instance Hu & Nation, 2000; Schmitt et al., 2011), but vocabulary coverage level is not a particularly sensitive measure of the average vocabulary load of citations. As a measure of the vocabulary load of citations, vocabulary coverage level suffers from two main drawbacks. First, as indicated above,

because of the small number of absolute words in a citation, there is rarely enough data for the small differences in coverage figures discussed in the literature to be operative: if a citation consists of less than 20 words, the 95<sup>th</sup> and 98<sup>th</sup> percentiles both equate to the lowest ranked word in the citation. Second, from a statistical perspective, percentile based measures of citations are relatively wasteful of data. In particular, a citation that contains a single low frequency word is in effect scored the same as a citation that contains multiple words at the same level. For example, consider a second citation from a concordance of the search word *relationship*: *council, spoke of the tremendous working relationship between the parish and district council*. Excluding the search word, this citations consists or seven 1k words (*spoke, of, the, working, between, the, and*) three 2k words (*council, district, council*) and two 3k words (*tremendous, parish*). Again, calculating the 98<sup>th</sup> percentile, we derive a score of three for the citation's vocabulary coverage level. While this does correctly indicate that a concordance user who knows the most frequent three thousand word families of English would be able to recognise 98% of the words in each citation, it ignores the fact the second citation contains a greater density of less frequent words than the first. For this reason, a second, more sensitive word frequency based measure of vocabulary load was also calculated.

### ***Mean word frequency***

In addition to vocabulary coverage level scores for each citation, mean word frequency (MWF) scores were also calculated. To determine mean word frequency scores, each word in a citation other than the search word itself is scored in relation to its frequency level on the expanded BNC-COCA-25 lists. The mean average score is then taken as the score for the citation. This measure has the advantage of using more of the data and, as a consequence, it is a more sensitive measure of the overall difference between the frequencies of the words contained in citations. For instance, if we return to the citation consisting of 16 1k words (*home, the, girl, is, said, to, have, a, poor, with, her, father, and, the, that, he*) and a single 3k word (*prospect*), its mean word frequency score can be calculated as  $\left(\frac{(1 \times 16) + (3 \times 1)}{17}\right) = 1.12$ . The second citation consisted of seven 1k words (*spoke, of, the, working, between, the, and*) three 2k words (*council, district, council*) and two 3k words (*tremendous, parish*), and so its mean word frequency score can be calculated as  $\left(\frac{(7 \times 1) + (2 \times 3) + (3 \times 3)}{12}\right) = 1.58$ . In other words, while vocabulary coverage level is

insensitive to the density of lower frequency words in a citation (both citations scored three on the measure of vocabulary coverage level), mean word frequency scores reflect the average frequency level of all the words in the citation.

Mean word frequency is, however, a rather abstract measure of a citation's vocabulary load. While research has consistently shown word frequency to be a predictor of readability, this does not entitle us to assume that a word at the 2k level will be 'twice' as readable as a word at the 4k level. Consequently, mean word frequency is not only a more sensitive measure of difference, but also a far more abstract measure. That is, although there is good reason to think that a citation containing more low frequency words will be harder to process than one containing less low frequency words and mean word frequency is a measure that can quantify such differences successfully, the unit of measurement used in mean word frequency is far removed from practical application. As such, mean word frequency and vocabulary coverage level are best seen as complementary frequency based measures of vocabulary load. Both measures are reported in the experiments below.

## **Results**

In this section, three experiments are reported. In the first experiment, an initial data set was generated for each of the three test conditions (Citation Set 1): citations from a corpus of academic prose, citations from a corpus of news and citations from a corpus of fiction. To test the hypothesis that there are statistically significant differences between scores on measures of word frequency for citations from the three different corpora, the Kruskal-Wallis test was used. To estimate the magnitude of differences in word frequency measures between the citations from the different corpora, a series of Mann-Whitney U tests were performed. The second experiment sought to replicate the findings of the first in a second data set. Again, the hypothesis was tested via the Kruskal-Wallis test, and effect sizes were examined through a series of Mann-Whitney U tests. Finally, in a third experiment, data from Experiment 1 was compared with data from a corpus of graded readers via Mann-Whitney U tests, using Hodges-Lehmann estimator to derive effect sizes in terms of word frequency levels. This third experiment provides an estimate of the magnitude of differences in word frequency measures between citations from different authentic

corpora as compared with word frequency measures of citations from a corpus of graded readers. In brief, the tests reported in this section assess the extent to which concordances of authentic and graded corpora differ on measures of word frequency, and so are likely to present learners with different levels of authentication burden.

### Experiment 1: Testing corpus concordance as a variable in the vocabulary load of citations

In the first experiment, 70 citations for each of 11 search terms were generated for each of the authentic corpora: *academic prose*, *news* and *fiction*. Inspecting histograms of the data, although the data did not conform to a normal distribution, the shape of distributions and extent of variance indicated that the non-parametric Kruskal-Wallis test could be used to test the hypothesis that the corpora differed on the two measures of vocabulary load: vocabulary coverage level and mean word frequency (see Table 42 below).

Table 42 Citation Set 1 scores on VCL and MWF

Genre	n	Vocabulary Coverage Level			Mean Word Frequency		
		Mean	Mean SD	Median	Mean	Mean SD	Median
Academic prose	770	4.56	3.61	3	1.50	.43	1.41
News	770	4.31	3.00	3	1.41	.37	1.31
Fiction	770	3.93	3.67	3	1.29	.40	1.16

The test results indicated that there were statistically significant differences between scores on vocabulary coverage level ( $H(2) = 61.747$ , 95% CI [ $<.001$ ,  $<.001$ ]), with a mean rank of 1255.28 for academic prose, 1204.56 for news and 1006.66 for fiction. Statistically significant differences were also observed for scores on mean word frequency ( $H(2) = 233.041$ , 95% CI [ $<.001$ ,  $<.001$ ]), with a mean rank of 1381.58 for academic prose, 1203.78 for news, and 1006.66 for fiction. A series of Mann-Whitney U tests were used as post-hoc tests to locate the differences identified in the Kruskal-Wallis test. To estimate effect sizes, the following formula was used:

$$r = \frac{Z}{\sqrt{N}} \text{ (Rosenthal, 1991 cited in Larson-Hall, 2010, p. 378).}$$

A Mann-Whitney U test for *vocabulary coverage level* found no statistically significant difference between *academic prose* ( $Mdn = 786.78$ ) and *news* ( $Mdn = 754.22$ ),  $U=283912.5$ , 95% CI [.136, .150]. The effect size was calculated as:  $r = \frac{-1.473}{\sqrt{1540}} = .03$ . However, a Mann-Whitney U test for *mean word frequency* indicated that *academic prose* ( $Mdn = 832.87$ ) was greater than *news* ( $Mdn = 708.13$ ) on this measure,  $U=248425$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-5.505}{\sqrt{1540}} = .14$ , which would conventionally be termed a small effect size ( $>.1$ ,  $<.3$ ).

A second pair of tests were run to compare results for the citations from the news corpus with those from the fiction corpus. A Mann-Whitney U test for *vocabulary coverage level* indicated that *news* ( $Mdn = 835.85$ ) was greater than *fiction* ( $Mdn = 705.15$ ),  $U=246132.5$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-5.844}{\sqrt{1540}} = .15$ . A Mann-Whitney U test for *mean word frequency* also indicated that *news* ( $Mdn = 881.15$ ) was greater than *fiction* ( $Mdn = 659.85$ ),  $U=211248$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-9.778}{\sqrt{1540}} = .25$ .

Finally, the data from the academic prose citations was tested against that of the fiction corpus. A Mann-Whitney U test for *vocabulary coverage level* indicated that *academic prose* ( $Mdn = 853.99$ ) was greater than *fiction* ( $Mdn = 687.01$ ),  $U=232160.5$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-7.487}{\sqrt{1540}} = .19$ . Again, the Mann-Whitney U test for *mean word frequency* indicated that *academic prose* ( $Mdn = 934.21$ ) was greater than *fiction* ( $Mdn = 606.79$ ),  $U=170394$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-14.459}{\sqrt{1540}} = .36$ , a medium effect size ( $>.3$ ,  $<.5$ ).

In summary, these tests found statistically significant differences between all three corpora on both measures, with the exception of the measure of vocabulary coverage level in relation to the contrast between citations from *academic prose* and citations from *news*. This exception is not terribly surprising given the pattern of results and the observable tendency of the vocabulary coverage threshold measure to provide a less sensitive measure of difference than mean word frequency. The results of the Mann-Whitney U tests are tabularised below.

*Table 43 Summary of effect sizes observed in Mann-Whitney U tests in Experiment 1*

Mean Rank		Measure	
Higher	Lower	Vocabulary Coverage Level	Mean Word Frequency
Academic	Fiction	$r = .19$	$r = .36$
News	Fiction	$r = .15$	$r = .25$
Academic	News	$r = .03$ ( <i>nps</i> )	$r = .14$

It seems clear from the table that the vocabulary load of the citations from the fiction corpus had a lower vocabulary load than those from either the academic prose corpus or the news corpus. However, the relationship between the vocabulary load of citations from the academic corpus and citations from the news corpus is less clear. The test result for vocabulary coverage level is not significant, but both the mean word frequency result and the relative effect sizes reported in relation to the fiction corpus suggest that the news corpus may actually have a slightly lower vocabulary load than the academic corpus. Thus, the non-significant Mann-Whitney U test result for the vocabulary coverage level of academic prose and news may be interpreted in two ways: either as evidence that the citations from these two corpora do not differ significantly, or as an indication of the relative insensitivity of VCL to what is certainly the smallest difference observed.

### **Experiment 2: Retesting corpus concordance as a variable in the vocabulary load of citations**

The second data set was generated with the intention of replicating the results obtained with the first. As in the first experiment, non-parametric tests were appropriate given the distributions of the data, but inspecting basic descriptive data for this data set suggested that the citations from the news corpus represented a considerably higher vocabulary load to the equivalent citations in the first experiment (see Table 44 below). It is also worth noting that, while the distribution of scores on mean word frequency mirror those of Data set 1 reasonably well, the vocabulary coverage figures do not.



Table 44 Citation Set 2 scores on VCL and MWF

Genre	n	Vocabulary Coverage Level			Mean Word Frequency		
		Mean	Mean SD	Median	Mean	Mean SD	Median
Academic prose	770	4.52	3.00	3	1.54	.41	1.44
News	770	4.72	3.21	4	1.51	.44	1.40
Fiction	770	4.54	3.61	4	1.39	.44	1.27

The non-parametric Kruskal-Wallis test was used to test the hypothesis that the corpora differed on the two measures of vocabulary load: vocabulary coverage level and mean word frequency. The test results indicated that there were statistically significant differences between scores on vocabulary coverage level ( $H(2) = 8.304$ , 95% CI [.013, .018]), with a mean rank of 1192.23 for news, 1173.43 for academic prose, and 1100.84 for fiction. Statistically significant differences were also observed for scores on mean word frequency ( $H(2) = 119.499$ , 95% CI [ $<.001$ ,  $<.001$ ]), with a mean rank of 1308.45 for academic prose, 1209.25 for news, and 948.79 for fiction. As in Experiment 1, a series of Mann-Whitney U tests were used as post-hoc tests to locate the differences identified in the Kruskal-Wallis test and to estimate effect sizes.

As in Experiment 1, a Mann-Whitney U test for *vocabulary coverage level* found no statistically significant difference between *academic prose* ( $Mdn = 763.73$ ) and *news* ( $Mdn = 777.27$ ),  $U=291240$ , 95% CI [.533, .553]. The effect size was calculated as:  $r = \frac{-.61}{\sqrt{1540}} = .02$ . A result equivalent to that of Experiment 1 was also obtained in a Mann-Whitney U test for *mean word frequency*: *academic prose* ( $Mdn = 803.82$ ) was greater than *news* ( $Mdn = 737.18$ ) on this measure,  $U=270795$ , 95% CI [.003, .005]. However, the effect size was calculated as:  $r = \frac{-2.941}{\sqrt{1540}} = .07$ , considerably smaller than in Experiment 1 ( $r = .14$ ).

Results for the second set of citations from the news corpus and fiction corpus were tested next. A Mann-Whitney U test for *vocabulary coverage level* indicated that *news* ( $Mdn = 800.47$ ) was greater than *fiction* ( $Mdn = 740.53$ ),  $U=273376.5$ , 95% CI [.005, .009]. The effect size was calculated as:  $r = \frac{-2.672}{\sqrt{1540}} = .07$ . A Mann-Whitney U test for *mean word frequency* also

indicated that *news* ( $Mdn = 857.57$ ) was greater than *fiction* ( $Mdn = 683.43$ ),  $U=229405.5$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-7.687}{\sqrt{1540}} = .20$ .

In the final test, academic prose citations from Citation Set 2 were tested against fiction corpus citations from Citation Set 2. A Mann-Whitney U test for *vocabulary coverage level* indicated that *academic prose* ( $Mdn = 795.2$ ) was greater than *fiction* ( $Mdn = 745.8$ ),  $U=277432$ , 95% CI [ $.025$ ,  $.032$ ]. The effect size was calculated as:  $r = \frac{-2.212}{\sqrt{1540}} = .06$ . The Mann-Whitney U test for *mean word frequency* indicated that *academic prose* ( $Mdn = 890.13$ ) was also greater than *fiction* ( $Mdn = 650.87$ ),  $U=204331.5$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-10.560}{\sqrt{1540}} = .27$ .

The test results from Citation Set 2 were comparable to those of Citation Set 1. Table 45 summarises the results of both. While the mean rank for vocabulary coverage level *academic prose* and *news* was inversed in the second data set, this result was not a statistically significant difference in either sample set. Perhaps more importantly, the effect sizes for the second data set are considerably smaller than they were for the first data set. However, the basic pattern of results is quite consistent: there is a significant difference between the fiction concordances of the fiction corpus and the other two corpora. Finally, it seems clear that mean word frequency provides both a more sensitive measure of vocabulary load and a more consistent measure. This is not very surprising considering mean word frequency makes use of much more data per citation than the vocabulary coverage level measure.

*Table 45 Summary of effect sizes observed in Mann-Whitney U tests in Experiment 1 & 2*

Mean Rank		Vocabulary Coverage Level			Mean Word Frequency		
Higher	Lower	Cit. Set 1	Cit. Set 2	Avg.	Cit. Set 1	Cit. Set 2	Avg.
Academic	Fiction	$r = .19$	$r = .06$	.125	$r = .36$	$r = .27$	.315
News	Fiction	$r = .15$	$r = .07$	.11	$r = .25$	$r = .20$	.225
Academic	News	$r = .03^a$	$r = .02^{a/b}$	.025 <sup>a</sup>	$r = .14$	$r = .07$	.105

<sup>a</sup> Non-significant differences

<sup>b</sup> In this cell, *News* actually had a higher Mean Rank than *Academic*

In summary, the combined results of Experiment 1 and Experiment 2 indicate that there was no significant difference between the citations from the news corpus and the academic corpus on a measure of vocabulary load (*nps*,  $r = .03$  &  $.02$ ), but there was small difference between them on a measure of mean word frequency ( $r = .14$  &  $.07$ ). In contrast, there was a small difference between citations from the fiction corpus and the other two corpora on the measure of vocabulary coverage level ( $r = .19$  &  $.06$ ;  $.15$  &  $.07$ ), and considerably larger difference on the measure of mean word frequency ( $r = .36$  &  $.27$ ;  $.25$  &  $.20$ ). The differences between the *academic prose* and *news* scores on the two measures indicates that the citations from the academic corpus contain a higher density of less frequent words than the news corpus, but require knowledge of a comparable number of word families to achieve coverage; however, citations from both of these corpora required a larger vocabulary size to achieve coverage and presented a higher density of lower frequency vocabulary in comparison to the citations from the fiction corpus.

### **Experiment 3: Comparing citations from genre corpora with citations from a graded reader corpus, estimating practical effect sizes**

Using graded readers has been both suggested and explored as a means of lowering the authentication burden of concordances, and it is reasonable to assume that citations extracted from a corpus of graded readers will have a lower vocabulary load than citations taken from any corpus of authentic materials; however, the magnitude of the difference between graded reader citations and authentic text citations is currently unknown. Furthermore, examining measures of word frequency in relation to a fourth corpus of graded readers not only allows estimation of the magnitude of vocabulary load differences between citations from different authentic corpora as compared with citations from a corpus of graded readers, but it also provides a point of comparison for judging the practical import of differences in vocabulary load between authentic corpora. To this end, a set of citations from a corpus of graded readers was created in parallel to Citation Set 1: that is, drawn from as similar a sampling frame as possible (see notes to Table 40 above, p. 149).

As in the previous experiments, the Mann-Whitney U was used to test whether citations from different corpora present different vocabulary loads. However, in this experiment, Hodges-Lehman estimator was used to determine effect sizes on a scale that can be interpreted in

practical terms: a scale equivalent that is matched to word family lists. Table 46 below shows basic descriptive data for the graded reader corpus below the parallel data set for the three authentic, genre corpora.

*Table 46 Citation Set 1 scores on VCL and MWF compared with Graded Reader scores*

Genre	n	Vocabulary Coverage Level			Mean Word Frequency		
		Mean	Mean SD	Median	Mean	Mean SD	Median
Academic prose	770	4.56	3.61	3	1.50	.43	1.41
News	770	4.31	3.00	3	1.41	.37	1.31
Fiction	770	3.93	3.67	3	1.29	.40	1.16
Graded readers	770	2.44	2.72	2	1.12	.24	1.06

A Mann-Whitney U test for *vocabulary coverage level* found *fiction* ( $Mdn = 905.21$ ) to be greater than *graded readers* ( $Mdn = 635.79$ ),  $U=192723$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-12.27}{\sqrt{1540}} = .31$ . A Mann-Whitney U test for *mean word frequency* found likewise: *fiction* ( $Mdn = 905.37$ ) was greater than *news* ( $Mdn = 635.63$ ),  $U=192602$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-12.094}{\sqrt{1540}} = .31$ .

Unsurprisingly then, a Mann-Whitney U test for *vocabulary coverage level* indicated that *news* ( $Mdn = 989.45$ ) was also greater than *graded readers* ( $Mdn = 551.55$ ),  $U=127860.5$ , 95% CI [.025, .032]. The effect size was calculated as:  $r = \frac{-19.747}{\sqrt{1540}} = .50$ , which would conventionally be deemed a large effect size ( $\geq .5$ ). The Mann-Whitney U test for *mean word frequency* also indicated that *news* ( $Mdn = 1019.88$ ) was greater than *graded readers* ( $Mdn = 521.12$ ),  $U=104430.5$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-22.156}{\sqrt{1540}} = .56$ .

Finally, a Mann-Whitney U test for *vocabulary coverage level* indicated that *academic prose* ( $Mdn = 1016.02$ ) was also greater than *graded readers* ( $Mdn = 524.98$ ),  $U=107400$ , 95% CI [ $<.001$ ,  $<.001$ ],  $r = \frac{-22.129}{\sqrt{1540}} = .56$ . Results for a Mann-Whitney U test for *mean word frequency* were similar: *academic prose* ( $Mdn = 1064.51$ ) was greater than *fiction* ( $Mdn =$

476.49),  $U=70061.5$ , 95% CI [ $<.001$ ,  $<.001$ ]. The effect size was calculated as:  $r = \frac{-7.687}{\sqrt{1540}} = .66$ .

The results of these three tests are presented in Table 47 below.

*Table 47 Summary of effect sizes for Mann-Whitney U tests in Experiment 3*

Mean Rank		Measure	
Higher	Lower	Vocabulary Coverage Level	Mean Word Frequency
Academic	Graded readers	$r = .56$	$r = .66$
News	Graded readers	$r = .50$	$r = .56$
Fiction	Graded readers	$r = .31$	$r = .31$

As might be expected, the effect size of the comparison between the citations from the graded reader corpus and the citations from the fiction corpus was much less than that for the comparison between the citations from the grader reader corpus and the citations from the academic and news corpora. To assess the practical implications of these findings, effect sizes for vocabulary coverage level were recalculated using Hodges-Lehmann estimator.

For the Mann-Whitney U test of *vocabulary coverage level fiction* ( $Mdn = 905.21$ ) greater than *graded readers* ( $Mdn = 635.79$ ),  $U=192723$ , 95% CI [ $<.001$ ,  $<.001$ ], Hodges-Lehmann estimator gives an effect size of 1.00, 95% CI [1.00, 1.00]. That is, to achieve 98% coverage of the fiction corpus, an additional 1000 word families are required, vis-à-vis the graded reader corpus (although, given the original unit of measurement is a word list, an additional 1000-word-family word list might be a more appropriate way of conceptualising this). However, when the Hodges-Lehmann estimator is used to recalculate the effect sizes for *academic prose* and *news* in relation to *graded readers*, it provides an effect size estimate of 2.00, 95% CI [2.00, 2.00], or two 1000-word-family lists.

Returning to data from Experiment 1, if we use Hodges-Lehmann estimator to recalculate the effect sizes for the Mann-Whitney U tests of vocabulary coverage levels, we find the following: for the Mann-Whitney U test using citations from the academic corpus and the fiction corpus, we are provided an effect size estimate of 1.00, 95% CI [1.00, 1.00]. This has interesting implications for interpretations of the effect sizes, as it would appear that, in terms of vocabulary coverage level, there is as much difference in vocabulary load between the data from the graded

reader corpus and authentic fiction corpus as there is between data from the authentic fiction corpus and the academic corpus; furthermore, this difference equates to approximately one 1000-word-family frequency band between each.

## **Discussion**

The results show that there are important differences between citations extracted from the different corpora in terms of their average vocabulary loads. Although tests using a measure of mean vocabulary level proved more sensitive to differences between corpora and provided more stable results, tests using the vocabulary threshold level are easier to interpret in terms of real world applications.

Vocabulary threshold level results support the conclusion that there is a difference of approximately one 1000-word-family vocabulary level band between the average citation drawn from the academic or news corpus and one drawn from the corpus of fiction, and that this difference is of comparable magnitude to the difference between citations drawn from the fiction corpus and those from the corpus of graded readers. In other words, concordances drawn from different corpora have different vocabulary loads and, therefore, are likely to present different authentication burdens. This result has important implications for research on learner use of concordances. For example, studies that have used concordances of news corpora with relatively low level learners (for instance Allan, 2006; Boulton, 2010a) may have presented learners with concordances that represent too high a vocabulary load for fluent reading comprehension and this may have negatively affected the learners' ability to make use of the concordances provided to them. The results reported in this chapter indicate that the vocabulary load of the citations presented to learners in concordances is likely to be significant variable in learner use of concordances and that it should be controlled for and reported on in future studies of concordance-based learning (cf. Chapters 1 & 2). The results also have implications for teachers using concordances with learners; teachers should consider carefully the likely vocabulary load of concordances extracted from different types of source text when selecting which types of texts to concordance.

As suggested above, the results also indicate that concordances of graded materials (Allan, 2009, 2010) are not the only option for reducing the vocabulary load of citations and, therefore, lowering the authentication burden of concordances by lowering their vocabulary loads is not simply a choice between concordancing authentic or graded corpora. Although citations from the graded reader corpus scored approximately two 1000-word-family vocabulary level bands below the concordances of news and academic corpora, citations from the fiction corpus represent something of a midway point between the two (requiring approximately 1000-word families more than citations from the graded readers and 1000 less than citations from the other authentic corpora). For concordances extracted from large general corpora, we can presume that they would approximate an average of the results for the three authentic corpora though, if the distribution of citations is concentrated within a particular type of text within the corpus, then we may expect the coverage figures to represent those of a corresponding text type. Hence, these results suggest that large general corpora may not represent an optimal concordancing resource if learners cannot be expected to have a sufficiently large vocabulary to cope with the vocabulary of the most difficult subcomponent of the corpus; this appears to approximate to knowledge of the first 4k word families of the BNC-COCA-25. The results also suggest that some authentic corpora may be more amenable to successful filtering than others (cf. Chujo et al., 2015; Kilgarriff et al., 2008; Wible et al., 2002). More broadly, corpus composition does indeed seem to have a significant effect on the vocabulary load of the citations extracted and, as such, it may be worth exploring the pedagogical value of concordances of authentic text types that are likely to represent relatively low vocabulary loads, such as writing for young readers, pulp fiction or other such intentionally accessible texts.

We should, however, be cautious of attributing the difference observed to genre. The results indicate that there are meaningful differences between the vocabulary loads of the authentic corpora analysed, and these results were replicable within these particular corpora, suggesting that the differences recorded may indeed be regular properties of concordances taken from the corresponding corpora. However, without replicating the results in new corpora, there is no guarantee that the differences observed are accurately ascribed to the genre of the corpora tested and not some other difference between them, such as style, register or topic, or some combination of such factors. Indeed, it would be rash to assume similar results would be obtained if a fiction corpus composed of notably wordy authors such as James Joyce, Oscar

Wilde and Salman Rushdie were tested. Hence, the results are most reasonably interpreted as indicating that authentic corpora can be expected to vary significantly in terms of the vocabulary load they represent. Whether the vocabulary load of a concordance can be reliably predicated from the composition of the corpus it is drawn from and, if so, exactly what the predictors are, is a matter for future research.

The differences between the results obtained in this study and previous research on whole texts also warrants some explanation. Previous research on whole texts had indicated that 8,000 to 9,000 word families are required to provide 98% coverage of written texts (Nation, 2006). However, the same level of coverage was provided for the citations from authentic corpora in this study by only 3,000 to 4,000 word families, *on average*. This difference is explicable in terms of word frequency distributions and working with the averages of measures. In line with the Zipfian distribution as a rough approximation of true word frequency distributions (Baayen, 2001; Sorell, 2015), we should only expect a small proportion of total tokens in a concordance to be low frequency word types. For instance, Nation (2006) reports only 0.72% of tokens in a corpus being at either the 8,000 or 9,000 word family level. Hence, we should only expect one such word to occur in a set of approximately ten citations, with the other nine citations consisting mainly of much higher frequency words. Thus, once the word frequency scores of these citations are averaged, the score of the citation bearing the low frequency word is mediated by the scores of many citations consisting of relatively high frequency words. It can be thought of in this way: in whole text measures, all of the low frequency items are clustered together in a critical area where the threshold takes effect; in text average measures, the relative weight of such words is mitigated by the presence of high frequency words in the critical areas of other texts. This can also be thought of in terms of the central limit theorem, although, in this case, it is the median value of multiple percentile measures that curves the extreme measures towards those displaying a central tendency.

### **Limitations**

It is perhaps worth reiterating the point that word frequency measures are only one gauge of a text's vocabulary load, and that vocabulary load is only one gauge of a text's readability, and readability may not be the only factor involved in authenticating a concordance. While the next



chapter explores another measure of vocabulary load that may be of use in gauging the utility of concordances as texts for language learners (type token ratio), it should be noted that there are two important ways in which word frequency research may not have the same implications for learner use of concordances as it does for learner use of other texts. First, the fragmentary character of concordances as texts may impose demands on learners quite separate from sufficient word recognition (see Chapter 2) and, as such, these other issues may need to be addressed before learners can make effective use of concordances, regardless of the concordance's vocabulary load. Second, when concordances are used in language learning tasks that are based on the notion of vertical reading, the presence of unknown words may not have the same effect as it does in conventional reading: as vertical reading does not orientate the concordance user to the text as a continuous discourse, skipping over unknown words is unlikely to have the same kind of effect on the user as it would in conventional reading. Consequently, for some affordances of concordances, conventional notions of readability may not be particularly relevant. For these reasons, it would be useful for future research to explore the vocabulary load of concordances in conjunction with measures of learner success in concordance use, which might be conceptualised in terms of successful application of concordance data, if not always straightforward comprehension of citations. These issues are returned to in Chapter 7 of the thesis and considered in relation to the broader findings of the thesis as a whole.



## Chapter 6. Exploring lexical repetition in concordances

### Introduction

Measures of word frequency have important implications for the authentication burden of concordances, and these implications are probably obvious enough to anyone familiar with research on second language vocabulary acquisition; however, measures of lexical repetition also have potentially important implications for learner use of concordances, even if these are less likely to be immediately understood. Chapter 2 explored the multiple affordances of concordances in relation to the various theoretical perspectives on concordance use that have been espoused in the literature, and lexical variability intersects with two important themes in this literature: the affordances of concordances in language for specific purposes (LSP) settings, and the affordances of concordances that are constituted in lexical repetition (most obviously pattern observation, but also to some extent data-driven learning in general). This study engages with these affordances by examining concordances in relation to a measure of lexical variation and repetition, type token ratio, or TTR (see Jarvis, 2013; Malvern, Richards, Chipere, & Durán, 2004; P. McCarthy & Jarvis, 2010). This chapter reports a study that uses a standardised measure of TTR (STTR) as a dependent variable and level of corpus homogeneity/heterogeneity (cf. *corpus preference* in Chapters 3 and 4) as the independent variable. The study tests the somewhat obvious hypothesis that the extent of lexical variation in a concordance is in part a function of the level of homogeneity amongst the texts concordanced, but perhaps more interestingly, the study also provides estimates of the extent to which this relationship can be manipulated.

This chapter begins by discussing TTR and other related measures quite generally, and then argues that the utility of any particular measure is a function of its fitness to a specified purpose and that, therefore, common critiques of TTR are either irrelevant to the use of TTR as a measure of lexical repetition in this context, or else actually an endorsement. Turning to the relationship between measures of lexical repetition and the affordances of concordances, research on TTR in vocabulary studies is briefly surveyed, focusing on TTR's relationship to notions of vocabulary load and authentication burden. It is argued that, in LSP contexts, TTR may actually be a preferable measure of a text's vocabulary load or authentication burden than measures that are based on word frequency (cf. Chapter 5). The relationship between lexical

repetition and opportunities for a learner to observe lexical patterns is then discussed: a potential affordance of concordances that has attracted much research interest (Barlow, 2004; Gilquin & Granger, 2010; Hunston, 2002, 2010; Sinclair, 1991; Tribble, 2010). Indeed, lexical repetition can be thought of as underlying many of the repeated features in concordances that concordance users may be interested in and, as such, can be thought of as underlying the looking for repetition *task preferences* identified in Chapters 3 and 4.

This chapter reports standardised TTR (STTR) data from 3 concordancing conditions that differ in terms of the level of generality of the corpus they are extracted from: a corpus of general written English, a corpus of written academic English, and a corpus of written academic ESP English (extracts from a single academic journal). It uses ANOVA to test for differences between STTR scores in the 3 test conditions and finds significant differences between them. The analysis is then repeated with a second data set, helping to validate the findings.

## Literature Review

### Type token ratio and more advanced measures

Type token ratio (TTR) is a simple and intuitively appealing measure (Jarvis, 2013). First proposed by Johnson (1939, 1944 cited in Jarvis, 2013, p. 16), a text's type token ratio can be determined by dividing the count of the number of word-types in a text by the count of the word-tokens in a text. Higher type token ratio scores represent texts with greater lexical variation than texts with lower type token ratio scores; texts with lower TTR scores contain less lexical variation, a greater number of repeated word tokens. However, in many research contexts, a significant shortcoming of TTR is its sensitivity to text length.

The effect of text length on TTR scores has been a major issue in many fields of research in which texts of varying length are compared (Jarvis, 2013; Malvern et al., 2004; P. McCarthy & Jarvis, 2010). Because a normal text creates coherence between different parts of the text through reference, the ratio of novel word tokens to previously used word types decreases as a function of text length (Baayen, 2001). For this reason, text length is a potential confound when comparing the TTR scores of texts of different lengths; when comparing texts of different lengths, differences observed in TTR scores may in part be the result of differences in text length

and not only the researcher's intended variable of interest. Because many researchers have been interested in comparing texts of different length, there has been a great deal of research that has explored the potential to adapt the simple TTR measure so that it is no longer affected by variations in text length (for a chronology of research in this area, see Jarvis & Daller, 2013). Typically, such measures use sampling techniques, transformations to a linear relationship, or a combination of the two, to counteract the effect of text length (P. McCarthy & Jarvis, 2010; van Hout & Vermeer, 2007). However, as concordances can be generated to a relatively uniform length by specifying the number of citations to be extracted, a simple TTR measure is not vulnerable to confounds of this sort as long as concordances composed of the same number of citations are being compared. In fact, the effect of any remaining minor differences in text length can be further reduced by standardising TTR scores: for instance, by calculating the TTR for every  $n$  number of words in the text, and then averaging the result (Scott, 2016). Such scores are known as a standardised type token ratio (STTR) scores.

An advantage of a simple STTR measure of lexical variation in concordances is that it does not take word order into account (cf. MTLT, P. McCarthy & Jarvis, 2010). Because concordances are fragmentary texts composed of unrelated sentences (Tognini-Bonelli, 2010), it would not be appropriate to assume that a measure that utilises inter-sentential relations, such as word order, can be directly transferred to measuring lexical variation in concordances. Such a measure would not distinguish between inter-sentential relations and the qualitatively different inter-citation relations found in a concordance. Contiguous sentences within a whole text have relationships of coherence and cohesion, citations within a concordance do not. Thus, a simple STTR measure is more appropriate for analysing concordances than existing advanced TTR-based measures, assuming that is that any TTR-based measure adequately addresses the construct under investigation. This is the subject of the next subsection.

### **Type token ratio as a measure of lexical variation and repetition**

The adequacy of a measure is a function of its ability to represent an underlying construct of interest. From this perspective, much of the criticism of TTR as a measure is a consequence of its misapplication. As the best known measure of *lexical diversity*, TTR has been used for a very wide range of purposes (stylistics, neuropathology, first language acquisition, second language acquisition, data mining and textual forensics), and interpreted as an index of a wide range of

underlying constructs (writing quality, vocabulary knowledge, speaker competence, Alzheimer's onset, hearing variation and socioeconomic status) (P. McCarthy & Jarvis, 2010). However, in the context of proposing a measure of language quality, Jarvis (2013) makes a distinction between *lexical diversity* and *lexical variability*, a distinction that undermines many of these intended applications of TTR. Jarvis notes that TTR measures do not reflect reader perceptions of repetitiousness, and hence do not provide an adequate index of lexical diversity. To briefly summarise his argument, although the objective measure of the ratio of types to tokens within a text reveals increasing repetition of word types as a text's length increases, this is unlikely to be perceptible to the reader because it is a natural quality of cohesive text (cf. Baayen, 2001); consequently, while TTR measures lexical variation and repetition objectively, to index the subjective quality of lexical diversity a measure would need to be able to distinguish redundancy from cohesively appropriate repetition. Of course, this is not a criticism of TTR as measure per se, only of interpreting it as a measure of lexical diversity. Indeed, when the construct under investigation is intentionally lexical variation and repetition, Jarvis' critique is actually an argument in favour of using TTR as a measure. The two main reasons for intentionally measuring lexical repetition/variation in the context of learner use of concordances are discussed below.

### **Type token ratio as a measure of authentication burden in LSP settings**

Lexical repetition can be interpreted as presenting learners with a lower vocabulary burden than texts that present learners with a wider variety of vocabulary, and so also as indicative of a lower authentication burden. However, type token ratio has so far received relatively little attention in research on vocabulary load. For instance, while Schmitt (2000, pp. 74-76) does provide a brief account of TTR as a measure of vocabulary load, Nation (2013) makes no direct mention of the term whatsoever. Instead, Nation talks more informally about reducing the vocabulary load of texts by reducing the number of different words in a text through *narrow reading*, citing a study that shows this kind of effect by comparing the TTR of a single, lengthy economics text with a similar quantity of text composed of a range of shorter academic texts (Sutarsyah, Nation, & Kennedy, 1994). Indeed, Nation (2013) goes some way towards explaining the relative neglect of TTR as a measure of vocabulary load when he points out that the effect of narrow reading on TTR is relatively modest because of the natural distribution of low frequency words in text.

From this perspective at least, the main drawback to TTR as a measure of vocabulary load is the absence of a text type in which significant differences in TTR can be observed. If, however we consider concordances as a means of providing exceptionally narrow reading, we might expect to find greater differences in TTR when concordances are examined, and hence find TTR a more useful measure of vocabulary load in concordance use than it is in whole-text-based narrow reading.

Encountering the same word in multiple citations clearly has potential to provide a type of text that can be used in narrow reading, and this possible affordance of concordances is complementary to various other possible affordances of concordances discussed in the literature (Chapter 2). For instance, Cobb (1997, 1999) has argued that concordances can provide learners with the kind of massed exposure to a word in use that is needed for developing word knowledge that extends beyond a form-meaning link between the L2 form and a related L1 meaning: for example, how a word is used in context. This is an approach to narrow reading driven by a vocabulary learning agenda (developing depth of vocabulary knowledge), and although it makes no explicit appeal to greater lexical repetition, these two views of concordances as texts for learners to read would appear complementary. Similarly, Bernardini (2000, 2002, 2004) has also argued for the value of exploring corpora through a concordancer in an open-ended reading procedure she terms *discovery learning*, an approach that emphasises attention to usage, learner autonomy and high levels of task engagement. This perspective on concordances as text for learners to read would also appear compatible with a view of concordances as texts with low TTR scores that are therefore suitable for use in narrow reading. Thus, although concordances may not typically be conceived of as a source of reading material, there are pedagogical motivations for viewing concordances as a source of reading material, even when a reduction in lexical variation is not explicitly referred to. In other words, not only do concordances have the potential to provide a concentrated text for massed encounters (Chapter 2), but they may also represent a text type with a lower vocabulary load as defined in narrow reading: a type of text that can presents learners with significant reductions in TTR. However, before endorsing TTR as a useful measure of vocabulary load, it is also necessary to consider the relationship between TTR scores and vocabulary knowledge.

The relationship between TTR and the notion of vocabulary load may at first appear straightforward. First, less vocabulary variety can be interpreted as indicative of a lower

vocabulary load in the straightforward sense of requiring learners to recognise fewer word types, a type of bottom-up processing advantage (Grabe & Stoller, 2001). Second, Nation (2013, p. 230) suggests that “Narrow reading may be more effective in increasing the amount of background knowledge that a learner brings to a text, thus aiding comprehension”. This perspective is akin to the discussion of lowering authentication burden by using concordances of familiar texts or small, focused corpora (see Chapter 2, pp. 39-47). In fact, lower TTR scores may also be indicative of topic homogeneity between citations, requiring readers to access a correspondingly narrow repertoire of schemata (for discussion of the role of schemata in reading, see Grabe & Stoller, 2001). Thus, from the perspective of reading, low TTR scores may also be considered to represent a top-down processing advantage (Grabe & Stoller, 2001). However, the extent to which low TTR scores represent a lower vocabulary load would appear to be contingent upon the text having already met a more basic criterion of vocabulary load, that of word recognisability (cf. Chapter 5).

As a measure, type token ratio makes no assumptions about which words a reader knows and which words are present in a text. Indeed, this can be seen as the main drawback of TTR as a measure of vocabulary load. Schmitt (2000) gives the example of an academic text that repeats key terms and hence scores higher than a children’s text with a simpler but more diverse vocabulary, arguing that TTR scores may not always reflect our intuitions as to the relative difficulties of texts. If a text has a relatively low type token ratio, it indicates that a reader will have to deal with a relatively narrow range of vocabulary as compared with a text of similar length but a higher type token ratio, but this does not mean that the words are likely to be known to the reader. As discussed in Chapter 5, being able to recognise the majority of words in a text can be considered a necessary, if not sufficient, condition for successful reading comprehension. Thus, the facilitative effect of a narrower range of vocabulary is likely to be contingent upon word recognition. If a learner does not know many of the words in a text, a low TTR score could actually represent a high concentration of unknown words, and so the text may actually have a very high vocabulary load for the learner, and therefore a high authentication burden as well. The utility of TTR as a measure of vocabulary load depends on the assumption that the text consists primarily of words types that the reader is likely to know. This is most likely the rationale for Nation’s (2013) prescription of narrow reading for advanced learners reading in a subject area that they are familiar with.



Because lower type token ratio scores can only be considered indicative of lower vocabulary loads when most of the words in a text are known, TTR as a measure of vocabulary load has limited relevance to vocabulary load in the amorphous *English as foreign language* context, but it may be important in relation to learner use of concordances in language for specific purposes settings. This point can be illustrated by comparing word frequency and TTR as measures of vocabulary load. Word frequency measures are usually based on the frequencies of words in a language in general (Nation, 2006; Schmitt, 2000) and reflect the order in which native speakers typically acquire knowledge of words (Biemiller & Slonim, 2001). Consequently, word frequency provides a useful index of the likelihood that a general reader (defined in terms of their cumulative knowledge of the most frequent words in the language) will be able to recognise a certain proportion of a text (defined in terms of how many of the most frequent words in a language are needed to provide a specified proportion of words in the text). In other words, the relationship between the reader's lexicon and the vocabulary in the text is mediated through description of the general frequency of words in a language. Thus, as a measure of vocabulary load, word frequency is independent of information about the details of any particular reader's lexicon, making it a highly generalisable measure of vocabulary load: if a reader knows all of the words up to a certain word frequency level, they will have sufficient vocabulary to recognise most of the words in a text of an equivalent or lower vocabulary frequency level. This is not the case with TTR. Only when there is a match between the learner's lexicon and the vocabulary in a text can TTR be considered a useful indicator of a text's vocabulary load.

Research has begun to show that second language learners' vocabulary knowledge often deviates from that of native speakers, with learners often having mixed word knowledge profiles: that is, profiles wherein they know words at a variety of frequency levels but do not necessarily possess complete knowledge of even the highest frequency word lists (Cobb, 2010; Milton, 2009; Webb & Chang, 2012). This is not to suggest that word frequency measures are of no use, but only to emphasise the possibility of learners being able to recognise the majority of words in a text that has a higher word frequency level than the word frequency level of their lexicon in general. We can think of this as the difference between being a general reader and a specialist reader (Ballance, 2012). For reading texts that are not a specific match to their lexicon, word frequency is a good indicator of the text's vocabulary load, but where such a convergence is

present, TTR may be a preferable measure. As Nation and Waring (1997, p. 10) point out, “Within narrowly focussed areas of interest, such as an economics text, a much smaller vocabulary is needed than if the reader wishes to read a wide range of texts on a variety of different topics”. Thus, in ESP contexts in which the learner’s vocabulary knowledge can be expected to match the vocabulary presented in a text, for instance when the topic of a text is within the learner’s area of expertise and they may be expected to be familiar with technical terms and other topic related low frequency vocabulary, TTR may be a better measure of authentication burden than word frequency.

### **Type token ratio and lexical patterning affordances**

In addition to providing a potentially useful measure of authentication burden in LSP contexts, type token ratio can also be used to evaluate the extent of lexical patterning in a text.

Concordances have attracted considerable research interest because of their potential to illustrate lexical patterns (Barlow, 2004; Gilquin & Granger, 2010; Hunston, 2002, 2010; Sinclair, 1991; Tribble, 2010), and just as concordances have helped researchers observe the patterns of collocation a search word occurs in, many researchers have suggested that concordances may also be able to help learners acquire such patterns (Basturkmen, 2006; Bernardini, 2002; Cobb & Boulton, 2015; Hunston, 2002; Johns, 1986; O’Keeffe et al., 2007; Sinclair, 2004; Warren, 2011; Willis, 1998). Fundamentally, all such lexical patterns depend on repeated co-occurrence: collocations, ngrams, clusters, skipgrams and congrams are all based on the analysis of recurring orthographic forms. Because concordances present citations, all of the word tokens in a concordance co-occur within the span of the search word specified in extraction of the citation and, consequently, the TTR of a concordance not only reports the extent of lexical repetition in the text (Jarvis, 2013), but also the extent of lexical repetition co-occurring with the search word. Concordances with lower TTR scores contain more collocates than concordances with higher TTR scores, and so afford learners more opportunities to observe lexical patterns of co-occurrence in a text than concordances with higher TTR scores. This function of TTR is independent of which words a learner knows in that, because it is not intended as a measure of vocabulary load, it does not address whether or not a learner knows these words or not. As such, it has application for research in learner use of concordances generally, not only learner use of concordances in LSP settings. Indeed, identifying recurrent lexis in concordance lines is

something of a mainstay of the literature on learner use of concordances (see for instance Barlow & Burdine, 2006; Bennett, 2010; Breyer, 2011; Burdine & Barlow, 2008; Coxhead & Byrd, 2007; Johns, 1986, 1991a, 1991b, 2002; Reppen, 2010b; Thurston & Candlin, 1997, 1998). TTR scores provide a useful measure of the extent to which a concordance affords learners the opportunity to observe repeated lexis, lower scores indicating more opportunity, higher scores, less.

### **Type token ratio and corpus composition**

Corpus composition is a key variable in concordance generation generally, and in relation to LSP in particular, a corpus's level of generality is a particularly important issue (Gavioli, 2005). In fact, Chapter 3 found that the second largest dimension of variance amongst concordance users' preferences was *corpus preference*: whether a concordance is extracted from a large, general corpus or a small, narrow corpus. This study explores differences between the STTR of concordances extracted from corpora that differ in terms of the level of generality instantiated in their composition: whether a corpus is composed of more or less homogenous or heterogeneous collection of texts. This reflects the relationship between topic homogeneity and lexical homogeneity observed in previous studies of TTR and narrow reading (Nation, 2013; Sutarsyah et al., 1994), but also the condition under which concordancing has been operationalised in studies of learner use of concordances. For instance, while some studies report having used concordances of corpora that seek to represent a language in general (Boulton, 2010a; Gordani, 2013; Koosha & Jafarpour, 2006; H. Yoon & Hirvela, 2004), others have presented learners with concordances from a more narrowly defined language variety of one type or another: concordances of only the written section of the BNC (Kaur & Hegelheimer, 2005), concordances of newspapers representing a national variety of English (Boulton, 2009b; Chan & Liou, 2005), or concordances of various collections of academic English (Charles, 2011; Lee & Swales, 2006). At a further level of specificity, we find studies in which learners were able to concordance the reading materials available to them on the course they were studying (Cobb, 1997, 1999), or corpora of writing relevant to a task they were working on (Chambers & O'Sullivan, 2004; O'Sullivan & Chambers, 2006; Park, 2012). Finally, at perhaps the finest level of corpus topic specificity, are studies in which learners compiled their own corpora of texts related to their individual field of language use (Charles, 2015; Lee & Swales, 2006). Because

such a wide range of corpora at different levels of generality have been used, and because TTR scores have implications for the authentication burden of concordances and their capacity to afford learners observation of lexical patterning, it is important to explore level of corpus generality as a potential predictor of the TTR scores of concordances.

### Research questions

1. Do concordances extracted from corpora with differing levels of generality have significantly different standardised type token ratios?
2. What is the magnitude of difference between the type token ratio scores of concordances extracted from corpora of differing levels of generality?

### Methodology

Three corpora representing different levels of generality were constructed. From each corpus tested, seventy citations for each of one hundred search terms were extracted, providing three hundred concordances for analysis. Each concordance was scored for lexical variation and repetition via the standardized type token ratio feature of WordSmith tools (Scott, 2016). The data obtained was split into two halves, each examined through ANOVA. The first half of the data comprised concordances generated with the first fifty search terms, the second half, the last fifty. Running the analysis on two independent data sets confirmed the stability of the results obtained.

### Corpora used

The experimental conditions used in this study were three different corpora used to generate concordances. Each corpus used represented a different level of generality in terms of the texts it was composed of. The three corpora each consisted of different samples of texts from the BNC. At the highest level of generality, texts were selected to represent written language in general: *corpus of written English*. At the middle level of generality, texts were selected to represent written academic language: *corpus of written academic English*. At the lowest level of generality,

all the texts in the corpus were extracts from a single academic journal: *ESP corpus of written academic English*. This corpus represents the kind of learner-compiled, language for specific purposes corpora discussed above. The table below presents a brief summary of each corpus used, and each corpus is described in more detail below that.

*Table 48 Summary of corpora used*

Corpus	Size	Composition
Corpus of written English	~16 mil.	Texts from 54 non-academic subgenre
Corpus of written academic English	~14 mil.	Academic texts
ESP corpus of written academic English	~700,000	Five texts from a single journal

#### ***Corpus of written English (written)***

It was not possible to use the full written sub-corpus of the BNC due to the file size constraints of the software used to extract concordances (Awesome-parser). Consequently, it was necessary to sample texts from within the written section of the BNC. Texts were sampled from a variety of genre categories (as defined in Lee, 2002) from the nine domain categories of the BNC: applied science, arts, belief and thought, commerce and finance, imaginative, leisure, natural and pure science, social science and world affairs (Burnard, 2002). The intention was to form a sub-corpus of the written BNC within which each of the nine domains contributed an approximately equal quantity of data and thereby represented a broad scope of written English. The resulting sub-corpus comprised approximately sixteen million words from 54 non-academic subgenre (Lee, 2002) of written English.

#### ***Corpus of written academic English (academic)***

The corpus of academic English comprised all texts from the written subsection of the BNC identified as academic (Lee, 2002), with the exception of five text files (discussed below). The resulting corpus of written academic English comprised approximately 14.5 million words.

### ***ESP corpus of written academic English (ESP)***

The corpus constructed to represent the lowest level of generality comprised five text files from the BNC: HU2, HU3, HU4, HWS and HWT. They were identified via David Lee and Paul Rayson's BNC Web Indexer (<http://ucrel.lancs.ac.uk/bncindex/>). All five files are extracts from the academic journal *Gut: Journal of Gastroenterology and Hepatology*, and comprised approximately 700,000 words. This corpus represent the kind of self-compiled source texts discussed in Charles (2015) and Lee and Swales (2006).

### **Data generation**

In brief, after compiling the three corpora that formed the independent variable in this study, one hundred search terms were identified as appropriate for generating concordances. Each of these search terms was used to generate a concordance in each of the test conditions, and then scored using a measure of type token ratio standardized to one hundred.

### ***Search terms***

Selection of search terms to use in this study addressed two main data collection concerns: availability and dispersion. On one hand it was important that there were sufficient occurrences of the search words used in each of the three corpora. On the other, it was necessary to ensure that the search terms generated citations that had sufficient dispersion within the more general corpora to meaningfully represent citations drawn from the general corpora. That is, that the citations were not extracted from just one or two texts within the larger, more general corpora. To address the first issue, because the ESP corpus was the smallest corpus used in this study, the availability of words in the ESP corpus was used as the starting point for identifying search terms to use: words available in sufficient quantities in the smallest corpus were likely to be available in sufficient quantities in the two much larger corpora. However, to address the second issue and provide some measure of control over the dispersion of search terms in the two larger corpora, high frequency words were also chosen. Thus, each of the 100 search terms used were word types found in the most frequent 1000 word families in the BNC COCA 25. Each set of search terms is shown in Table 49 below. As hypothesized, they were found to occur sufficiently frequently in the more general corpora, and visual inspection of the citations in the concordances

was used to confirm an appropriate degree of dispersion; more advanced quantitative approaches to evaluating dispersion were not available within the software used to extract concordances.

*Table 49 Search terms used in STTR studies*

Search terms 1-50	Search terms 51-100
able, action, added, age, agreement, air, allow, almost, alone, already, also, amount, animal, apparent, appearance, area, base, become, better, blood, body, bone, breath, carried, case, caused, change, children, class, clear, collection, complete, considered, continuous, control, cost, count, course, cross, daily, day, death, degree, dependent, died, differences, difficult, double, drugs, during	earlier, end, especially, even, except, expected, experience, explanation, expressed, fact, fall, family, fast, fat, fed, final, finding, fixed, follow, food, forms, found, free, full, further, gave, general, good, greater, group, growth, health, high, history, hospital, hour, human, important, indeed, information, informed, interest, involvement, known, large, last, later, lead, least, left

### ***Concordance generation***

The concordancer described in Chapter 5, Awesome-parser, was also used to generate concordances in this study. For each of the one hundred search terms used in this study, a concordance of seventy citations was extracted. Citations were selected randomly, and each citation was extracted with as many complete words of cotext as found within an eighty character span of the search term. Duplicate citations were manually identified, removed and replaced with randomly selected non-duplicates. The data was then split into two halves to provide two sets of concordances consisting of fifty search words each.

### **Scoring**

To generate consistent measures of lexical variation and repetition, each concordance generated was scored via the standardized type token ratio (STTR) feature of WordSmith Tools (Scott, 2016). The scoring proceeded on the basis of word type (not lemma or family), no stop list was used and the program was set to standardize scores to TTR per 100 words. This transformation of the data compensated for any minor difference between concordances in terms of their exact length.

### Data analysis

Data for item sets 1-50 and 51-100 were both examined before conducting statistical analyses. Both data sets satisfied the assumption of normal distribution, but failed the assumption of equal variance. However, because group sizes were equal and  $n = 50$  in each test, the effect of unequal variance is a very minor concern: a distortion of  $\alpha$  by a few hundredths (Sheskin, 2004). More importantly, failure to satisfy the assumption of equal variances indicated that Dunnett's T3 was the appropriate post-hoc test to use: Dunnett's T3 is a conservative test that can be used when all pairwise comparisons can be made but the assumption of equal variance is not met (Larson-Hall, 2010).

### Results

This section reports the analysis of each data set in turn. The results are summarised in Table 50, presented at the end of this section.

#### Analysis of variance for items 1-50

A one-way ANOVA was conducted to compare the effect of *level of corpus generality* on *standardized type token ratio* (STTR). Descriptive statistics for the groups were: *written*,  $X = 70.74$ ,  $SD = 1.97$ ,  $n = 50$ ; *academic*,  $X = 67.78$ ,  $SD = 2.57$ ,  $n = 50$ ; *ESP*,  $X = 64.12$ ,  $SD = 3.57$ ,  $n = 50$ . Since the assumption of equal variances was not met, the test statistic used for the ANOVA was *Welch's* adjusted  $F$  ratio. ANOVA showed that the effect of corpus generality of STTR was significant, *Welch's*  $F(2, 93.267) = 70.667$ ,  $p = <.001$ . Comparisons using Dunnett's T3 with 1000 bootstrap samples found statistically significant differences between all groups: *written* and *academic*,  $MD = 2.96$ , 95% CI [2.04, 3.88],  $p = <.001$ ,  $d = 1.5$ ; *academic* and *ESP*,  $MD = 3.66$ , 95% CI [2.15, 5.17],  $p = <.001$ ,  $d = 1.42$ ; *written* and *ESP*,  $MD = 6.63$ , 95% CI [5.22, 8.03],  $p = <.001$ ,  $d = 3.36$ . This result indicates that concordances extracted from more general corpora have higher STTR levels as compared with concordances extracted from more specialised corpora, supporting the hypothesis that level of corpus generality has a very large effect on the STTR of concordances.



### Analysis of variance for items 51-100

A second one-way ANOVA was conducted to compare the effect of *level of corpus generality* on *standardized type token ratio* (STTR) in a second data set. Descriptive statistics for the groups were: *written*,  $X = 71.07$ ,  $SD = 2.31$ ,  $n = 50$ ; *academic*,  $X = 68.31$ ,  $SD = 2.10$ ,  $n = 50$ ; *ESP*,  $X = 64.14$ ,  $SD = 4.30$ ,  $n = 50$ . Again, since the assumption of equal variances was not met, the test statistic used for the ANOVA was *Welch's* adjusted  $F$  ratio. ANOVA showed that the effect of corpus generality of STTR was significant, *Welch's*  $F(2, 92.786) = 54.254$ ,  $p = <.001$ .

Comparisons using Dunnett's T3 with 1000 bootstrap samples found statistically significant differences between all groups: *written* and *academic*,  $MD = 2.76$ , 95% CI [1.89, 3.63],  $p = <.001$ ,  $d = 1.32$ ; *academic* and *ESP*,  $MD = 4.17$ , 95% CI [2.89, 5.61],  $p = <.001$ ,  $d = 1.99$ ; *written* and *ESP*,  $MD = 6.93$ , 95% CI [5.24, 8.61],  $p = <.001$ ,  $d = 3.01$ . The analysis of the second data set confirmed the results of analysis of the first: concordances extracted from more general corpora had higher STTR levels as compared with concordances extracted from more specialised corpora, supporting the hypothesis that level of corpus generality has a very large effect on the STTR of concordances. The results of both data sets are presented in Table 50.

Table 50 Summary of STTR comparisons for Level of corpus generality

		MD	$d$	95% CI
Written cf. academic	Items 1-50	2.96	1.5	2.04, 3.88
	Items 51-100	2.76	1.32	1.89, 3.63
	AVG	2.86		
	SD	0.1		
Academic cf. ESP	Items 1-50	3.66	1.42	2.15, 5.17
	Items 51-100	4.17	1.99	2.89, 5.61
	AVG	3.92		
	SD	0.4		
Written cf. ESP	Items 1-50	6.63	3.36	5.22, 8.03
	Items 51-100	6.93	3.01	5.24, 8.61
	AVG	6.78		
	SD	0.2		

## Discussion

The results clearly indicate that the concordances of high frequency words extracted from the three different corpora differed significantly on a standardized measure of type token ratio. Not only are all of the effect sizes reported very large (anything over .08 is typically considered large), but they represent significant differences in real terms. Because ratios were standardized to 100 word tokens, average score differences between conditions are equivalent to differences in the average number of word types per 100 words. Hence, taking the average of the mean difference between scores for both data sets, we can say that concordances from the corpus of written English contained an average of 2.86 [SD 0.1] more word types per hundred word tokens than concordances from a corpus of written academic English. This figure approximates to around 3 more word types per 8 citations. Comparing concordances from a corpus of written academic English with concordances from an ESP corpus composed of extracts from a single academic journal, we observe an average difference of 3.92 [SD 0.4]: that is, in the more general corpus, almost 4 more word types are present per 8 citations. Unsurprisingly then, when concordances from the general written corpus are compared with those from the ESP corpus, the average is 6.78 [SD 0.2]: nearly 7 more word types per 8 citations in the general concordances than the ESP concordances. To put this another way, learners would be confronted with nearly one fewer word type per citation in the ESP concordance. This finding has important implications for both the authentication burden of concordances and their capacity to illustrate patterns of lexical co-occurrence.

In regard to authentication burden, these findings show that concordancing may be an effective way of realizing the goal of a reduced vocabulary load of a text. While TTR score reductions may be difficult to operationalize through narrow reading of whole texts (Nation, 2013), these results show significant reductions in lexical variability can be achieved by concordancing corpora composed of closely related texts. Unfortunately, because TTR measures are sensitive to text length (Jarvis, 2013; Malvern et al., 2004; P. McCarthy & Jarvis, 2010), it is difficult to identify a whole text TTR baseline to compare these results with. To make such a comparison, we would need to compare the STTR of equivalent quantities of whole texts and concordances, but such a comparison might be considered somewhat dubious given the very different characteristics of whole texts and concordances. It is not really a legitimate comparison.

However, in relation to Tom Cobb's perspective on concordances as providing massed exposure to a word in context (Cobb, 1997, 1999), it would be interesting to operationalize *level of corpus generality* as a variable in a study that operationalised concordancing in this way. On one hand, the results reported in this chapter indicate that extracting concordances from corpora with increasingly low levels of generality produces concordances with increasingly low STTR scores, and such concordances may also represent a lower vocabulary threshold if the vocabulary contained is familiar to the learner. This suggests that learners may therefore find them easier to authenticate. However, such a concordance is also likely to show a correspondingly restricted range of usage as compared with a concordance extracted from a corpus at a higher level of generality, and this is likely to interact with the learning outcomes of concordance use, although whether this interaction would have a primarily positive or negative effect on learning outcomes is hard to predict. The increased uniformity may make learning attested language features easier, but it may also limit the scope of the learning outcomes. The implications of TTR scores for learning outcomes is an important area of research on learner use of concordances that would benefit from further research.

It is also important to consider these findings in relation to other measures of readability; after all TTR vocabulary load is only one measure of readability (for discussion of other factors see Grabe & Stoller, 2001; Nation, 2013). Chapters 3 and 4 suggest it might also be worth considering concordances and readability in relation to citation format. It would probably make most sense to use sentence format citations if readability is likely to be a significant factor in the type of learner use of concordances envisioned because sentence format citations present the reader with grammatically complete stretches of text which are therefore likely to be easier to read. However, this has the potential to interact with the findings from this study because this study used fixed-length-format citations, and sentence-format citations typically contain more words per citation than fixed-length-format citations. Thus, we might expect the STTR scores of concordances composed of longer citations to have slightly more homogenous STTR scores because they will typically include more word tokens at greater distances from the search word, reducing the overall influence of any collocational pull the search term may possess. This can also be thought of as presenting a less fragmentary text. However, the effect of level of corpus generality on STTR scores that is reported in this study would still remain relevant to the central

portion of longer citations even if longer, sentence-format citations with more homogenous STTR scores were used.

It should also be remembered that the potential for a reduction in lexical variability to represent a lower authentication burden is dependent on the vocabulary contained being known: reduced TTR vocabulary load is probably irrelevant to readability when many words are unknown (cf. Chapter 5). Hence, readability gains reflected in reduced STTR scores for concordances extracted from corpora at lower levels of generality are likely to depend on a match between learners' internal lexicons (Cobb, 2010; Milton, 2009) and the particular words contained in a given concordance. In other words, from the perspective of readability, these results are most pertinent in relation to the use of concordances in LSP settings in which it is feasible to narrow the composition of a corpus in such a way that it reflects the learners' areas of familiarity or expertise (cf. Chambers & O'Sullivan, 2004; Charles, 2015; Lee & Swales, 2006; O'Sullivan & Chambers, 2006; Park, 2012). Hence, evaluating the extent to which concordances with lower STTR scores provide concordances with lower authentication burdens in LSP settings is also an important research goal that needs further attention.

In contrast to the authentication burden perspective on these results, if these results are viewed from the perspective of the potential for concordances to afford learners exposure to patterns of lexical co-occurrence, these results have wide application beyond LSP settings. Lexical repetition underlies the concept of one word being a common collocate of another. If a certain word only co-occurs with the search word a single time in a concordance, the concordance user can observe that these words *can* be used together, but if a concordance shows the word co-occurring with the search word in multiple citations, then the concordance user has a basis on which to estimate the strength of the association between these two words as compared with other words that occur in the same syntactic relation to the search word a different number of times. Of course, the reliability and generalizability of this observation depends upon the composition of the concordance and the selection of the texts examined (see Chapter 2), but the point stands: if there is no repetition, one can only infer that a certain pair of words can co-occur with the search word, not that any particular co-occurrence is more frequent than another. Hence, lexical repetition provides the underlying basis for one type of generalisation – generalisations as to which words more frequently occur with the search term, an important aspect of many DDL type exercises (see for instance Barlow & Burdine, 2006; Bennett, 2010; Breyer, 2011; Burdine

& Barlow, 2008; Coxhead & Byrd, 2007; Johns, 1986, 1991a, 1991b, 2002; Reppen, 2010b; Thurston & Candlin, 1997, 1998). This study has demonstrated that concordances extracted from corpora at lower levels of generality can provide concordances that offer more opportunity to observe such lexical repetitions: lower STTR scores are an indication of increased opportunity to observe words collocating with the search term. Consequently, more homogenous corpora represent better resources when these types of affordance are sought. For this reason, it is strongly recommended that STTR scores are provided for the concordances used in research on learner use of concordances when concordances are conceptualised as affording learners the opportunity to observe patterns of co-occurrence: it is a potentially significant variable.

### Limitations

Although STTR scores do underlie patterns of lexical repetition, STTR is a somewhat crude indicator of linguistic patterning in some important respects. Lower type token ratios indicate less lexical variation and more lexical repetition, and so do indicate an increased opportunity to observe a word co-occurring with the search term in a concordance, but several points need to be remembered about exactly what it is that STTR measures. First, it is important to remember that STTR does not distinguish between word types that are repeated within a citation and word types that are repeated across citations. Some proportion of repeated word types may be within a single citation, and hence not representative of the type of collocation observed when a word type is repeated across citations. However, such words are likely to be function words (it is relatively unlikely to find lexical words repeated within a very short span or words, though phrases such as *more and more* and *better and better* are an obvious exception); there is no strong reason to think that the extent of this phenomenon should vary much between conditions. Hence, repeated lexis within citations probably has a very marginal effect on the results obtained as compared with words repeated across citations. Second, STTR does not take co-occurrence of lexical sets (e.g.: *oil, water, liquid* and *fluid*) nor grammatical categories (e.g.: *adjectives, nouns, verbs*) into account, only orthographic forms. It is only sensitive to lexical repetitions: in this study, lexical repetition at the level of word types. It cannot effectively assess the extent of semantic or grammatical repetition in a concordance. Third, STTR makes no warrant as to the value or significance of co-occurrence, even when a word co-occurs with the search term in multiple citations. There may be a sentence boundary such as a full-stop between the repeated co-

occurring word and the search word, or the word may occur at a greater distance from the search word than is typically considered in studies of collocation (for instance, 6 or 7 words away from the search term). From this perspective, it is interesting to note that there is increasing interest in collocates that occur at greater removes from the search term than the traditional 5-word span (Hoey, 2015), and so this may not be as significant a disadvantage of TTR as a measure of lexical patterning as it may at first appear; it would, however, still be interesting to explore this issue further, perhaps by measuring at the citation level the frequency of concgrams (Cheng, Greaves, & Warren, 2006) occurring in each concordance. An analysis based on concgrams would not only provide a much finer grained quantitative account of the types of lexical repetition contained in the concordances, but would also provide a starting point for a more qualitative analysis of the types of co-occurrences prevalent in concordances.

It should also be remembered that this study examined concordances of high frequency words. This was done to ensure that the search words used to generate concordances occurred in sufficient quantities in the ESP corpus examined but also had a reasonable range of distribution within the general corpora. Consequently, it is not clear how generalisable these results are to less frequent search terms. For instance, if we search the whole BNC for an extremely technical term such as *endoprosthesis*, we find that all forty-three occurrences are in two texts, HU4 and HWS. That is, 2 of the 5 texts that formed the ESP corpus. This suggests that concordances of very low frequency words in large general corpora are likely to form their own *de facto* specialized concordance in virtue of their limited distribution. Hence, it seems likely that the effect of level of corpus generality will interact with the dispersion of a search word within a corpus. As less frequent words are generally more narrowly dispersed, we may expect the magnitude of the differences in STTR scores of concordances extracted from corpora at different levels of generality to diminish when less frequent, less dispersed words are concordanced. A better understanding of this phenomenon would help to clarify the circumstances in which level of corpus generality can be expected to affect a concordance's STTR.

Finally, although the experimental results obtained were replicated in two sets of data derived from two different sets of search terms, the corpora examined remained constant. Therefore, we cannot be sure that similar results would be obtained if level of generalizability were operationalized differently. From this perspective, a cross-validation study using different corpora might be useful. But perhaps more importantly, it is not completely obvious which

factors in compiling a corpus at a lower level of generality have an effect on the STTR scores of concordances. For instance, the BNC was originally divided into 9 domain categories (Burnard, 2002) without a distinction being drawn between academic and non-academic texts (Lee, 2002). Lee's BNC Indexer provides a plethora of options for compiling narrower sub-corpora of texts from the BNC: medium, domain, genre, audience age, audience sex, audience level, author age, author sex, author type, text sampling type, circulation status, interaction type, time period and mode, as well as customizable criterion for identifying texts on the basis of library catalogue keyword fields (COPAC keywords), an indicator of a text's topic. It is not clear which ways of narrowing a corpus can be expected to affect the STTR of concordances extracted from them, or the relative magnitude of any affect that each may have. Consequently, exploring the effects of different ways of narrowing the focus of a corpus is an interesting avenue for future research.

In sum, this study has shown that corpora representing different levels of generality can consistently produce concordances with markedly lower STTR scores, and this has important implications for pedagogy. However, there are many more issues that remain to be explored: first, the extent to which lower STTR scores affect what can be learned from a concordance; second, the extent to which these STTR differences can represent more readable concordances; third, a finer grained analysis of the types of lexical repetition that low STTR scores represent; fourth, the extent to which there is an inter-action effect between level of corpus generality, the distribution of a search term within a corpus and STTR scores; and fifth, which factors within level of corpus generality are operative upon the STTR scores of concordances extracted from said corpora.





## **Chapter 7. Discussion and conclusion**

This chapter discusses the meaning and significance of the findings reported in Chapters, 3, 4, 5 and 6. It begins with short summaries of the main findings for each of the four studies reported in these chapters. It then looks at the implications of these findings for theoretical perspectives on learner uses of concordances (see Chapter 2), synthesising the findings to identify two prototypical models of concordance use. Each is discussed in relation to the affordances it promotes and its authentication burden. The chapter then turns to the implications of the findings for research on learner use of concordances, interpreting the findings in terms of the models of experiment design discussed in Chapter 1. This subsection is then followed by reflections on the methodologies chosen in Chapter 3, 4, 5 and 6, their strengths and their weaknesses. This leads to a discussion of key directions for future research. In the final subsection, the results are interpreted from the perspective of ground level teaching practice and implications are reinterpreted as tentative, but empirically informed, recommendations for practice.

### **Summary of findings**

The four studies reported in this thesis have wide ranging implications for learner use of concordances. Each is briefly summarised below.

#### **Study 1. Analysing concordancing: a simple or multifaceted construct?**

The factor analyses reported in Chapter 3 showed that concordancing should not be thought of as a simple, unidimensional construct but a complex composite of factors, different permutations of which may offer distinct, sometimes mutually exclusive, affordances. Although not exhaustive of possible factors within concordancing, the study found 4 dimensions of variance that underlay the concordancing preferences of the concordance users surveyed. First, a preference for the citations that form a concordance to be presented in the key word in context (KWIC) format, or else an alternative, fundamentally incompatible format, in which each citation is likely to span more than a single line of text, and citations are aligned to the left-hand margin without consideration of the position of the search term — a non-KWIC citation format that we might think of in terms of sentence format citations. Thus, even if KWIC formatting can be considered

the default format for concordances in many linguistic settings (Sinclair, 1991), this study shows that it is not the preferred formatting style for many concordance users. The study also found respondents were distinguishable on the basis of three more preferences: preferences for large, general corpora or small, focused corpora; preferences for concordances that present citations in cotext-order or in text-order order; and preferences for looking for information repeated in multiple citations, or not. Hence, we should indeed be wary of transferring the paradigm of concordance use in research to the paradigm of concordance use in learning (Pérez-Paredes, 2010). A proportion of the respondents had preferences for each level of each factor identified, and so each can be assumed to afford a proportion of users some form of benefit as compared with its alternative. From this perspective, the problem may not be the transfer of techniques from research to learning, but the selection of the most appropriate techniques to transfer from amidst the full ambit of concordance use. In terms of experiment design, as a treatment condition, concordancing does not appear to be a fixed constant with a single level; the results reported in Chapter 3 clearly indicate that concordance use should be conceptualised as either a fixed constant with multiple levels or a variable component. As such, we need to think carefully about how we report on and generalise from studies of learners using concordances (see *Implications for research* below). Each factor identified is a potentially significant variable in learner use of concordances.

### **Study 2. Correlations between concordance users and concordancing preferences**

The second study, reported in Chapter 4, analysed more data from the survey and found correlations between several groupings of concordance users and their concordancing preferences. This study moves beyond Study 1 in that it identifies types of concordance preference with types of concordance user. It found that users of bilingual concordances had extremely pronounced preferences for citations to be presented in non-KWIC format and for citations to be ordered by the texts they have been extracted from (not their cotext). It found that how frequently respondents use concordances is predictive of both their *formatting preferences* and their *task preferences*: the more frequently a respondent uses concordances, the more likely they are to be viewing concordances in KWIC format and looking for content that is repeated in multiple citations, the less frequently, the less likely. It also found that the field of concordance use that the respondent thought of when responding to the survey questions was also predictive

of these same preferences, and also *corpus type* preference. Using concordances *for translation* was marked by a very strong preference for reading concordances for non-repeated content, using non-KWIC format concordances, and using concordances of *small, focused corpora*. A pronounced preference for these concordancing operationalisations was also observed for respondents' thinking of using concordances *for language learning* and *as a reference resource*. In contrast, the strongest preference for using concordances in looking for repeated information, using KWIC format citations and large, general corpus was observed from respondents thinking of using concordances *in teaching: directly*, followed by respondents who were thinking of using concordances *in linguistic research*. Once again, this strongly echoes Pascal Pérez-Parades' observation that there has been a tendency for the research paradigm of concordance use to be transferred to pedagogical contexts (Pérez-Paredes, 2010), perhaps at the expense of considering other models of concordance use that may have more relevance to learner use, such as concordance use in translation or editing. It also suggests that KWIC format concordances may not be the most natural choice of formatting when learners are set tasks that draw on reference type engagement with concordances, such as editing tasks, error correction tasks or translation tasks.

### **Study 3. Exploring the authentication burden of concordances through word frequency measures**

The third study examined concordances as texts, exploring word frequency measures of citations as a dependent variable and corpus concordanced as an independent variable (Chapter 5). It included data collected from a corpus of graded readers that facilitated comparison of authentic text with pedagogically mediated text. The study found that concordances had surprisingly low vocabulary loads *on average*. The average citation presented in concordances of medium-sized authentic corpora required a vocabulary size of approximately 4000 words for a learner to be able to recognise the majority of words presented in it. This figure departs considerably from those reported for whole texts (8-9000 words; Nation, 2006), and this difference was attributed to the study exploring vocabulary coverage averages at the level of citations, as compared with vocabulary coverage figures for whole texts. Concordances are fragmented texts composed of separate citations, and therefore it makes most sense to measure vocabulary coverage on a citation-by-citation basis, assuming that learners would be able to ignore a proportion of citations

in a concordance without it affecting their comprehension of other citations containing more recognisable vocabulary.

This study also found significant word frequency differences between concordances from different authentic corpora, and found the magnitude of these differences to be comparable to the magnitude of difference observed between concordances of a corpus of graded readers and concordances of the easiest authentic corpus, the corpus of fiction. The results indicate that careful selection of corpus materials can have a considerable effect on the vocabulary load of concordances extracted from them and, as such, represents an alternative approach to pedagogical mediation of concordance texts through concordances of graded materials (Allan, 2009, 2010) or filtering of data from large general corpora (Chujo et al., 2015; Kilgarrieff et al., 2008; Kilgarrieff et al., 2015; Wible et al., 2002). The study also reported figures based on mean average word frequency per citation, as opposed to vocabulary coverage level per citations. Mean average word frequency figures were comparable to vocabulary coverage figures, and perhaps also a little more reliable as they are more resistant to outliers than vocabulary coverage figures (mean averages as compared with percentiles).

The results of this study highlight the importance of reporting on the vocabulary load of concordance-based materials used in research on learner use of concordances and of ensuring that the vocabulary load of concordances is appropriate to the proficiency level of the learners (cf. suggestions that learner vocabulary knowledge should be measured and reported in the context of research on learner use of concordances: Franken, 2014). The study shows that the vocabulary load of concordances can vary significantly, and that a considerable proportion of this variance is a function of corpus concordanced, as research on whole texts would suggest (Nation, 2006). However, without cross-validation, we cannot be sure that the operational factor responsible for the differences observed was genre, and not some other difference between the corpora tested. Nevertheless, with more research in this area, it might be possible to provide reliable guidelines as to the effect of selecting particular types of text on the vocabulary load of the concordances extracted from them. The findings from this study are very much preliminary findings in this respect, but they do suggest that popular fiction texts present less of a vocabulary burden than concordances from newspapers or academic texts and so, presumably, will they will also present learners in general language learning contexts with a lower authentication burden than concordances of news or academic texts. It appears that concordances of authentic, popular

fiction texts produce concordances with a vocabulary load midway between the vocabulary load of concordances of graded readers and concordances of news or academic texts.

#### **Study 4. Exploring lexical repetition in concordances**

The final study used type token ratio (TTR) to measure the extent of lexical variation and repetition in concordances (see Chapter 6). In this study, the independent variable was the level of generality of the corpus that concordances were extracted from, operationalised as a corpus of general written English, a corpus of academic written English, and an English for specific purposes (ESP) corpus of written academic English. Again, large differences were found between test conditions. The results indicate that extracting concordances from corpora at a lower level of generality can significantly narrow the range of vocabulary presented in a concordance. The results can be interpreted in relation to both vocabulary load (and so also authentication burden) and opportunities for observing patterns of co-occurrence. The lower the TTR score, the fewer word types contained in the text (lexical repetition); the higher the TTR score, the more word types (lexical variation). Thus, in language for specific purposes (LSP) settings in which there is a match between the learner's lexicon and the vocabulary contained in the texts, lower TTR scores represent lower vocabulary loads as learners are presented with a narrower range of vocabulary. When the affordance sought from a concordance is the opportunity to observe patterns of co-occurrence, low TTR scores indicate more frequent repetitions of orthographic forms, and so also increased rates of co-occurrence at the level of orthographic forms. This presumably makes observation of patterns considerably easier, because the visibility of any particular pattern of use will depend on the proportion of citations that display that pattern and how dispersed they are (Frankenberg-Garcia, 2014b; Gavioli, 2001; Sinclair, 2004). In brief, the findings from Study 4 suggest that concordances extracted from more homogenous collections of text may not only represent a lower authentication burden in LSP settings, but also greater opportunity to observe patterns of co-occurrence and increased opportunity for making pedagogical generalisations in data-driven learning tasks. In the next subsection, these findings, along with the findings from the other three studies, are discussed in relation to theoretical perspectives on learner use of concordances.

## Implications for theory

Each study reported here makes an individual contribution to research on learner use of concordances, but the results are probably most interesting when considered holistically, and from an explicitly theoretical perspective. In the first year of my PhD, I was fortunate to have the opportunity to discuss my research with Professor Birgit Henriksen, who strongly encouraged me to develop a theoretical model of concordance use to locate an empirical study of learner concordance use within. Birgit considered the grounding of an empirical study within an explicit theoretical model a prerequisite of any meaningful empirical investigation. Needless to say, I found her perspective insightful and provoking, and it radically changed the direction of my PhD studies. As discussed in Chapter 2, treating concordances as input texts in SLA from a theoretical perspective, there would appear to be two main patterns of engagement with a concordance: engaging with the relationships between citations, a pattern of engagement I referred to as *analysis*, or engaging with the relationships between individual citations and an external context of use, a pattern of engagement I referred to as *reference*. The results from the studies reported in this thesis both lend a degree of support to such a model of learner use of concordances, but also shed light on how the operationalisation of concordancing may interact with the affordances of each.

The two main patterns of engagement with a concordance advanced here reflect the findings presented in Chapters 3 and 4. In Chapter 4, we saw that three of the four factors identified in Chapter 3 correlated with *field of use* (*task preference*, *format preference* and *corpus preference*) and two of these also correlated with *frequency of use* (*task preference* and *format preference*). Very interestingly, examining the signs of these correlations in relation to the theoretical perspectives on learner use of concordances discussed in Chapter 2 reveals that operationalisations of concordancing that are hypothesised to make concordancing harder coalesce in the first pattern of engagement, *analysis*, while operationalisations of concordancing that are hypothesised to make concordancing easier coalesce in the second pattern of engagement, *reference*. As the table below illustrates, aspects of how concordancing is operationalised that are associated with less cognitive complexity or lower authentication burden coalesced, as did aspects associated with more cognitive complexity or greater authentication burden.

*Table 51 Analysis and reference as composites of concordancing preferences in relation to aspects of concordancing hypothesised as harder or easier*

Factor	Analysis	Reference
Task	Looking for repetition	Looking for relevant examples
<i>cognitive complexity</i>	<i>harder</i>	<i>easier</i>
Format	KWIC	Non-KWIC
<i>authentication burden</i>	<i>harder</i>	<i>easier</i>
Corpus	Large, broad	Small, narrow
<i>authentication burden</i>	<i>harder</i>	<i>easier</i>

In Chapter 2, consideration of theoretical perspectives on learner concordance use suggested that looking for repetition under the conditions for making robust generalisations is a cognitively complex task and is likely to involve quantities and qualities of data that may be inappropriate for learners (Ädel, 2010; Boulton, 2010a; Braun, 2007; Coxhead, 2014; Flowerdew, 2012a, 2012b; Franken, 2014; Gavioli, 2001; Hunston, 2002; C. Kennedy & Miceli, 2001; Levy, 1990; O’Sullivan & Chambers, 2006; Partington, 2001; Pérez-Paredes, 2010; Pérez-Paredes et al., 2012; Stevens, 1991a; Sun, 2003; Varley, 2009; H. Yoon & Hirvela, 2004); that short, truncated KWIC formatting may exacerbate the already heavy authentication burden of concordances as fragmented texts (Aston, 2001; Boulton, 2009b; Cobb, 1999; Johns, 1986; Levy, 1990; Mishan, 2004; Stevens, 1991a; H. Yoon & Hirvela, 2004); and that large heterogeneous corpora are likely to pose a heavier authentication burden than smaller, more narrowly focused collections of text (Aston, 2001; Braun, 2005; Gavioli, 2005; Gavioli & Aston, 2001). This coalescence of harder and easier aspects of concordancing suggests that there may indeed be both a prototypically harder and a prototypically easier operationalisation of concordancing available to learners (cf. Gabreilatos, 2005), and that we should indeed be wary of assuming that the paradigm of research-like concordance use is the most appropriate model of concordance use for language learners (Alcaraz Calero & Pérez-Paredes, 2008; Pérez-Paredes, 2010). Learners might be able to benefit more easily from the *reference* type exploitation of concordances that seems to be typical of infrequent concordance users, people using concordances for translation purposes, people

using concordances as a reference resource and, small in number as they may be, people who do actually use concordances for language learning (see Chapter 4, and Figure 1 below).

Figure 5 Two models of concordance use with prototypical aspects of concordancing in relation to selected categories of concordance user with means of group centroids

		Frequency of use preferences	
		Very rarely (-.73) Occasionally (-.39)	Very often (+.33) Quite often (+.27)
Field of use preferences	For translation (-.83) To learn languages (-.39) As a reference resource (-.33)	<u>Reference</u> Relevant examples Non-KWIC format Small, narrow corpora	
	For linguistic research (+.28) For teaching directly (+.41)	<u>Analysis</u> Repetition KWIC format Large, broad corpora	

These results support a model of concordance use in which there are two prototypical types of concordance use, *analysis* and *reference*. *Analysis* type use is likely to be more difficult, *reference* type use is likely to be easier. Each prototype is discussed in more detail below.

### Prototype 1. Analysis

To suggest that it may be easier for learners to use concordances in a *reference* type way is not to suggest that using concordances in an *analysis* type way is without value. As suggested in Chapter 2, looking for patterns of co-occurrence, the process of analysing concordances and making generalisations, potentially making locally appropriate generalisations, or even potentially making scientifically defensible generalisations, can all be seen as potential affordances of concordances.

It is interesting that this pattern of results is associated most strongly with the preferences of teachers using concordances directly with learners, more so than for any other field of use, including *for linguistic research*. The explanation may be that language workers such as



linguistic researchers are almost certainly using concordances hands-on and so the strength of their preferences may be tempered by a pattern of use in which multiple affordances are drawn upon: recall also that using concordances within multiple fields was the norm amongst respondents, not the exception (Chapter 3), and so we might expect that many concordance users will be predisposed to use concordances flexibly. It may be that much professional use is characterised by a flexible approach in which the operationalisation of the concordance is continually adapted to the users' needs at any given time: the preferences of linguistic researchers may reflect the dominant trend in their field of use, but this dominant trend may well be underpinned by flexibility of approach in which concordances are used in alternative ways as and when needed. This may account for the relatively low score of respondents answering from the perspective of *for linguistic research*.

The stronger results for teachers using concordances directly with learners certainly seems to indicate that teachers have seen great value in this prototypical type of engagement with concordances in which learners are looking for repeated information in a concordance extracted from a large general corpus and presented in a KWIC format. There are a number of possible reasons that this may be so. One possible view is that it may simply be an uncritical and unreflective transfer of an assumed paradigm of researcher-like use to pedagogy. A second view would be that teachers seeking to provide their students with particular affordances adopt this model in hands-off use in which learners are not provided the opportunity to switch between different formats and corpora: this may be because teachers wish to simplify the task of concordance use by *taking the computer out of the equation* (cf. Boulton, 2010a), or simply because of limited access to appropriate resources (cf. Breyer, 2011). For instance, in contexts of hands-off use in which a single formatting choice has to be made, teachers may be electing to present their learners with KWIC format concordances because of the purported affinity between pattern observation affordances and this formatting choice. Alternatively, teachers may not have access to smaller, more focused corpora; they may be dependent on freely available corpora learners can self-access online. The relatively strong scores of teachers may well represent conditions of using concordances directly with learners that preclude a more flexible approach to how concordances are used. However, whatever the reason for the relative strength of scores, although teachers using concordances with learners directly do seem to endorse this model of concordance use more strongly than any other category of user reported (.41), it is fairly close to

the pattern of use reported by respondents answering from the perspective of linguistic research (.28). There are, however, good reasons to think this model of use may benefit from some revision in the context of learner use.

This model of engagement with concordances involves three aspects: task ( $r = .93$ ), format ( $r = .54$ ) and corpus ( $r = .30$ ) but, while the first two are the most strongly correlated and have a theoretically supportable relationship with the affordances of *analysis* type use of concordances described above, it is less clear why using large general corpora should feature in this model. KWIC formatting is often associated with the observability of patterns of co-occurrence (Aston, 2001; Boulton, 2009b; Braun, 2005, 2006; Charles, 2012; Daskalovska, 2013; Flowerdew, 2012b; Gavioli, 2001; Götz & Mukherjee, 2006; Johns, 1991b; Kaur & Hegelheimer, 2005; Levy, 1990; McKay, 1980; Mishan, 2004; Römer, 2010, 2011; Sinclair, 2004; Sun & Wang, 2003; Tono et al., 2014; H. Yoon & Hirvela, 2004), and patterns of co-occurrence are logically related to repetition. However, in stark contrast, a large heterogeneous corpus would *a priori* appear likely to reduce frequency of repetition. Chapter 6 tested this hypothesis, and found clear evidence of corpus heterogeneity as a predictor of type token ratio, a reliable measure of lexical repetition and variation (Jarvis, 2013). The more heterogeneous the corpus concordanced, the less repetition and the more variation; the more homogenous the corpus, the more repetition and the less variation. Hence, from the perspective of observability and the making of generalisations, smaller, more narrowly focused corpora would appear to have better affordances than larger, more general corpora.

It is unclear exactly why using a large general corpus should appear as part of the *analysis* prototype of concordance use; however, if it is because of the greater generalisability of observations made in regards to such corpora, it should be recalled that the conditions of such robust generalisation are not only the representativeness of the citations presented, but also the selection of the citations to be analysed (see discussion of exhaustive analysis and iterative random sampling in Chapter 2). Given apparent consensus that such intensive analysis of concordances is likely to present too great a burden for language learners (Ädel, 2010; Boulton, 2010a; Braun, 2007; Coxhead, 2014; Flowerdew, 2012a, 2012b; Franken, 2014; Gavioli, 2001; Hunston, 2002; C. Kennedy & Miceli, 2001; Levy, 1990; O'Sullivan & Chambers, 2006; Partington, 2001; Pérez-Paredes, 2010; Pérez-Paredes et al., 2012; Stevens, 1991a; Sun, 2003; Varley, 2009; H. Yoon & Hirvela, 2004), the results from Chapter 6 provides strong support for

using less generalisable but more homogenous text collections with learners when the affordances sought are related to pattern observation: concordances of more homogenous collections of text present more lexical repetition (Chapter 6), which equates to more observable patterns of co-occurrence (Frankenberg-Garcia, 2014b; Gavioli, 2001; Sinclair, 2004) and so also concordances that may be expected to be more supportive of learners making locally appropriate generalisations (Gavioli, 2005).

### **Prototype 2. Reference**

Prototypical reference type engagement with a concordance is characterised by operationalisations of concordancing suggested in the literature to reduce the difficulty of concordance use. At the simplest level, the *task preference* of looking for relevant examples in a concordance does not require advanced inductive reasoning skills such as degeneralisation (Sinclair, 2003), does not require any particular quantity of citations to be engaged with and would seem to allow of picking and choosing which citations are attended to (see Chapter 2). From this perspective, a *reference* type orientation to concordances would appear much simpler than an *analysis* type orientation. Interestingly, reference type *task preferences* coalesced with preferences for using non-KWIC format concordances and smaller, narrower corpora: both of which have been suggested to be easier for learners to authenticate (see above). While it is not one-hundred percent clear why these preferences should coalesce, a number of points can be made.

One reason may be that concordances with lower authentication burdens are simply preferred by default; if you are not seeking an affordance advanced by KWIC formatting (for instance, increased observability of lexical patterning) or using large general corpora (the potential for making scientifically defensible generalisations about languages in general), then there is no obvious reason not to operationalise concordancing in ways that are cognitively simpler and more amenable to authentication.

A second reason may be that, if the affordance sought in *reference* type use is relevance (see Chapter 2), a concordance needs to be processed by the user in a manner more congruent with traditional reading skills: *reference* type use may be drawing on traditional reading skills more heavily than *analysis* type use. There is, as far as I am aware, no research on the processes that are involved in vertical reading, but it is not entirely clear that vertical reading is actually a

form of *reading* in the ordinary sense. Certainly, a degree of disfluency in the processing of words would not seem to preclude successful completion of a typical data-driven learning task of the identify-classify-generalise type described by Johns (1991b): when identifying and classifying words in each citation's cotext, there is no obvious reason why looking words up in a dictionary would constitute a major obstacle. Words are being read, but it is not clear that the text has to be processed at any levels of processing above basic word recognition, or that word recognition must be rapid enough to facilitate fluency (cf. Grabe & Stoller, 2001). However, if the affordance sought is relevance, it seems likely that citations do need to be read in much the same way that ordinary non-fragmented texts are, and that the reader will need to draw on both bottom-up processes of decoding as well as top-down processes, such as schemata access; it seems very likely that both these types of process would be required to decide whether a particular citation meets your criterion of relevance or irrelevance. Thus, it may be that *reference* type use encourages concordancing operationalisations that lower the authentication burden of concordances and thereby facilitates judgements about the relevance of the citations accessed.

Chapter 5 picked up this important issue and examined concordances from the perspective of authentication burden. Focusing on an important and readily quantifiable index of readability, word frequency, it found that concordances extracted from different corpora impose significantly different vocabulary loads, and vocabulary load is almost certainly one component of a concordance's broader authentication burden. It demonstrated that the magnitude of difference in vocabulary load within concordances of authentic texts is comparable to the magnitude of difference between concordances of graded readers and concordances of some particular types of authentic text. These results underline the importance of evaluating the vocabulary load of concordances, and ensuring that the vocabulary load of a concordance is appropriate for the proficiency level of learners (cf. Franken, 2014), especially if learners are using concordances in a *reference* type manner and attempting to identify citations that are relevant to an immediate context of use, such as when using concordances in editing (cf. Charles, 2012; Gavioli, 2005; Lee & Swales, 2006). Thus, the reference model of concordance use may furnish a particularly useful set of default concordancing operationalisations in the context of using concordances for productive purposes (cf. Charles, 2012; Franken, 2014; Frankenberg-Garcia, 2014a, 2014b; Johns, 1986; Lai & Chen, 2013).

It might also be added that the results of the study reported in Chapter 6 can be interpreted from an authentication burden perspective within the context of LSP. While narrow reading may generally be unable to provide significant reductions in vocabulary load, the results reported in Chapter 6 demonstrate that concordances can facilitate what we might call *exceptionally narrow reading* when combined with exploitation of narrowly defined collections of text. In combination with the purported benefit of using concordances of familiar texts (Aston, 2001; Braun, 2005; Gavioli, 1997) or text types (Aston, 2001; Gavioli & Aston, 2001), using concordances of narrow corpora may also represent a major reduction in authentication burden that may greatly enhance the utility of concordances in *reference* type use in particular.

In sum, casual use of concordances for *reference* type affordances may simply be taking advantage of operationalisations of concordancing that are conducive to lower authentication burdens, by default as it were. There are however also positive reasons to think that when the affordances sought are specifically relevance affordances, lowering the authentication burden of concordances may be the primary concern and so a strong motivation to operationalise concordancing in the prototypical *reference* type operationalisation advanced here. Chapters 5 and 6 have complemented the reference model identified in Chapters 3 and 4 with preliminary studies of variables in concordance generation that require no manual mediation between learner and concordance (each study examined randomly selected samples) but produced significant changes in potentially important indicators of authentication burden (word frequency measures and type token ratio). The results reported indicate that the corpus used to generate concordances should be considered a major variable, particularly in *reference* type use.

## **Implications for research**

The studies reported in this thesis also have important practical implications for research on learner use of concordances. Most obviously, the results clearly demonstrate that concordances can be used in multiple ways and that the linguistic qualities of concordances can vary systematically. In the terms discussed in Chapter 1, the results show that the number of possible treatments that could be termed concordance treatments is greater than one ( $K > 1$ ). This means that research on using concordances needs to consider the experiment design model employed

more carefully, operationalising treatments more fully (increasing  $k$  so that  $k = K$ ), defining the experimental treatment condition more narrowly (reducing  $K$  so that  $k = K$ ), or taking a random sample of treatments and adopting an appropriate statistical model for treating  $k$  as a variable component (Model II). Chapters 3 and 4 identified variables in the construct of concordancing pertaining to operationalisation decisions. Chapters 5 and 6 identified variables in concordancing resulting from one of those operationalisation decisions in particular: corpus concordanced. Each is discussed below.

The study reported in Chapter 3 identified four theoretically interesting variables that concordance users have distinguishable preferences for. Thus, at a minimum, the format of citations, the corpus concordanced, the order citations are presented in and the task engaged in are all variables within the broad construct of concordance use that need to be treated as such in experiment design. Furthermore, Chapter 4 showed that there are complex relationships between different types of users and various combinations of the variables identified; thus, researchers should ideally also be alert to the possibility of inter-action effects amongst these variables. For instance, as discussed above, it seems that reading for repetitions and using KWIC concordances may be mutually supportive of each other in the context of frequent use and linguistic research. It may be that it is only in combination with each other that these factors lead to reliable results. Chapter 4 showed that variables within concordancing can coalesce and, therefore, experiments that investigate a single variable without controlling for other variables or looking for interaction effects are going to be far less informative than those that do.

Chapter 5 showed that the texts concordanced can systematically affect concordances in terms of the word frequency of the cotext presented. Word frequency is held to be an important component of a text's vocabulary load, especially in terms of their being a threshold of vocabulary knowledge required for fluent reading and adequate comprehension (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). As such, the word frequency rating of concordances is very likely to be a component of the authentication burden of the citations within a concordance. Therefore, it is essential that researchers take account of this variable in research on learner concordance use. On one hand, it is important that learner proficiency level and concordance vocabulary load are considered in relation to each other: it does not seem at all unlikely that some null results in the field are the result of showing learners concordances that contained too many unknown words. It also seems perfectly possible that

using concordances with a lower vocabulary burden with language learners might lead to better results in concordancing treatments. In an ideal world, researchers would be able to assess the vocabulary level of concordance treatments concordance-by-concordance. It would allow researchers to control, manipulate and report on word frequency levels with a high degree of precision. In combination with reporting of learners' vocabulary levels (Franken, 2014), this would allow us to explore word frequency level as a variable very effectively, and perhaps even to ascertain whether word frequency level is more or less important in concordance use than it is in the use of other text types (cf. discussion of the role of intuition in reading concordances in Chapter 2), and whether it has comparable effects on effective concordance use in both *analysis* type use and *reference* type use. As suggested above, there is every reason to think that *reference* type use of concordances involves traditional reading skills and that, as a consequence, word frequency level will be an important variable in *reference* type learner use of concordances. How important traditional reading skills are in *analysis* type use is considerably less clear.

Chapter 6 investigated concordances using a measure of type token ration (TTR) and demonstrated that level of corpus homogeneity-heterogeneity has a significant and systematic effect on the TTR score of the concordances extracted. This finding has two main theoretical implications. First, it has implications that are similar to those for word frequency as an indicator of authentication burden, especially in LSP contexts. Second, it has implications for any learner use of concordances where lexical repetition underlies the affordance sought (most DDL tasks and learner use involving observation of patterns): it underlies the strength and observability of any orthographic patterning in a concordance. Because of these theoretical implications, in both LSP contexts of learner use and learner use involving attending to patterns of lexical co-occurrence, TTR is a variable that should be assessed, controlled or manipulated, and reported.

### **Methodological implications, limitations and reflections**

The research reported in this thesis was innovative in a number of ways, but there are many methodological limitations to the research reported as well. Chapter 3, 4, 5 and 6 have each discussed direct and detailed limitations to the studies reported. In this section, a broad perspective is taken on the research methods adopted, the limitations of the approaches taken are

discussed, and reflections on the research process overall are given. It first discusses the survey research reported in Chapter 3 and 4, and then the lexical analysis research reported in Chapters 5 and 6.

### **Survey research**

In terms of the survey reported in Chapter 3 and 4, I found great value in gathering rich descriptive data alongside the narrowly defined survey items that were used to answer the research questions posed. This was achieved in a number of ways, incorporating varying degrees of innovation. Most innovatively, using survey flow options to link response categories of later questions to earlier questions allowed me to collect some very fine grained descriptive information without overwhelming respondents with multiple questions that presented a large number of response categories. For instance, the survey used tiered questions to gather rich descriptive data for *field of use* before finally assigning respondents to a single fixed response category for the purposes of quantitative analysis: that is, a multiple answer, multiple choice question containing *yes*, *occasionally*, *no*; followed by a constant sum question asking respondents to distribute 100% of their time over categories they had answered *yes* or *sometimes* to; followed by the final single response, multiple choice question used in the analysis. The finer grained picture available through these tiers of question items provided important insights into the general picture of concordance use amongst respondents, insights into the respondents' concordance use that would not have been available otherwise. For instance, it had not occurred to me that the majority of respondents would be concordancing dilettantes, perhaps primarily using concordances in one or two fields, but also dipping into concordances for many other purposes on occasion. It was also interesting to see just how ambiguous multiple choice data such as *yes* and *occasionally* can be, especially as compared with the comparatively nuanced data collected in terms of percentages. It is clear that words like *yes* and *sometimes*, and by extension also *no*, cover a very wide range of figures and that the range of proportions that each choice category refers to varies by individual. For this reason, I think gathering rich descriptive data is very important, especially when the constructs concerned are treated as simple variables within the survey instrument and the items tapping them are not designed from a psychometric standpoint.



In a similar vein, although not particularly innovative, I found providing open response, text comment boxes throughout the survey very helpful as well. The comments helped a great deal with data cleaning and getting to know the data: unusual patterns of response were often accompanied by text entry boxes in which respondents explained their interpretation of a question or gave other pertinent details about their answers. It also appeared to let frustrated respondents explain their concerns or, on occasion, let off some steam while, thankfully, continuing to complete the survey in an apparently satisfactory manner: in other words, it appears to have had a beneficial effect on the respondent's overall attitude to the survey, quality of response and completion rates: very important considerations in overall survey design (Punch, 2003).

Finally, in regards to the analyses of open text responses gathered for biographical data, I followed a similar approach to Tribble (2015) and used corpus tools to identify recurrent themes in the data. It underlined to me how unsatisfactory a tick box system would have been, especially if it was operationalised as single answer, multiple choice questions. Respondents defined themselves in multiple capacities, or with nuances that were informative but would have been lost if a closed response format had been used. Closed response categories might have been easier to interpret and summarise, but they would have been a very poor reflection of the nuanced responses that were collected through text response boxes.

More generally, while it is well-recognised that successfully analysing quantitative data involves getting to know your data well, it is perhaps less commonly acknowledged that there is a qualitative dimension to getting to know your data, even in quantitative research. As Scholfield (1995) points out, in the final analysis, reliability and validity are intertwined. Without an assurance of validity, reliability measures are meaningless because we cannot be confident about what it is that we are measuring, no matter how reliably we are measuring it. While assessment of reliability is fundamentally quantitative, assessment of validity is essentially qualitative. Punch (2003) is correct to point out that a well-designed, quasi-experimental survey instrument is a research instrument for detecting variation and so the research instrument will almost inevitably use fixed response items, but interpreting that variation is greatly enhanced by access to rich descriptive data because it is the rich descriptive data that helps you to evaluate the validity of your survey instrument *a posteriori*, not the tests of reliability that are used to assess the psychometric properties of the survey as a research instrument.

On this note, it is worth mentioning some of the shortcomings of the research items used in the survey instrument. As suggested above, closed response items are pretty much a necessity in the design of a quasi-experimental survey. Developing classification schemes from open responses would be likely to result in as many categories of response as respondents, numerous subjective grouping decisions, or large quantities of unclassifiable data. This is not a serious problem in terms of descriptive data, but it is a serious problem with any data that is to be used in an inferential statistical procedure. As such, developing appropriate fixed response categories is an important task for the designer of a survey instrument. While the survey instrument used in Chapters 3 and 4 was broadly successful in detecting variation, the results would probably have been much clearer and easier to interpret if it had been possible to develop the response categories used further. In particular, it now seems obvious how inadequate a response category *in linguistic research* was. Looking at the accompanying open text responses for this response category, it appears that further subdivision would have been desirable: *linguistic research* is a very broad church, and in retrospect it seems that different paradigms of linguistic research are likely to involve different prototypical exploitations of concordances. Without the possibility of an infinitely extendable piloting period, there will inevitably have to come a time at which a research instrument is deemed fit for purpose and brought into use, but if the survey was to be used again, I would certainly look to improve upon the response categories provided for *field of use*, and the response category *for linguistic research* in particular.

On reflection, it still seems valuable to have adopted a survey approach to the question of the construct of concordancing. The findings reported in Chapters 3 and 4 have made contributions to the field in that they provide empirical evidence that supports a view of concordancing as a complex construct, provides indications of how different operationalisations of concordancing may reflect diverse and divergent affordances, and supports a degree of theory building. An alternative approach would have been to use richer qualitative methods such as observation and interview.

Quantification is the process of reducing the infinitely rich experiential world to numerical abstractions. This process of turning observations into numbers allows observations to be interpreted within statistical models which, in turn, facilitate the making of defensible, generalisable statements about the world. Hence, quantification necessarily involves a loss of granularity and selective attention. Numerous subtle differences are ignored when observations

are coded or measurements are rounded, and the selection of which facets to code or measure also implies the rejection of other potentially important facets of the phenomenon under investigation. These inherent drawbacks of quantitative methods can be mitigated to some extent by the use of qualitative methods. Qualitative methods can provide a check upon this process of simplification because they do not necessitate the same degree of observational reduction. While qualitative approaches would have had comparatively limited generalisability, they might have provided a richer account, potentially identifying aspects of concordance use not addressed by the survey research reported here, or have identified more nuanced differences between the observations made. They might also have provided a more direct window onto the motivations behind concordance users' concordancing preferences. Qualitative methods might have allowed the research to move beyond perceptions and orientations towards behaviours and their interactions with perceptions and orientations. But this does not entirely undermine the value of the quantitative survey research reported here. Personally, I find Punch (2003, p. 22) insightful on this topic:

Research-based knowledge in any field builds and accumulates as much through the weight of evidence, where the results of numerous smaller studies are accumulated, as through the findings of a single definitive study. Thus well-conducted small-scale surveys have a valuable contribution to make.

The survey results presented here, as with so much research in our field, represent preliminary findings, but the results obtained can also be seen as providing an indication of where more in-depth qualitative approaches might begin, as providing an empirical grounds for theory building, and as helping to identify some largely neglected variables of interest within quantitative research on learner use of concordances.

### **Lexical analysis research**

Turning to the studies reported in Chapter 5 and 6, data collection and processing turned out to be much more complex and time consuming than envisaged. To provide as much comparability across treatments as possible, it was desirable to have as much control as possible over as many variables as possible, and this necessitated the use of tailor made software. To give a simple

example, extracting individual citations and entering them into a frequency counting program would have multiplied data processing times by a factor of thousands, equating to literally years of work. It would have precluded pilot studies that helped to estimate optimal conditions for obtaining samples with relatively low standard deviations and investigation of the likely distributions of the formal samples eventually obtained, both considerations having important implications for determining sample sizes, and so also sampling schemes. In relation to Study 5 in particular, the necessity of manually expanding upon the BNC-COCA-25+ word lists placed real limits on how many samples could be analysed, even with the capacity to extract and process sets of citations en masse and the capacity to rank words by vocabulary threshold automatically. In this respect, I am grateful beyond words to my brother, Daniel Ballance, who not only developed the software used in the first instance, but also implemented multiple layers of expanded functionality as piloting revealed new and unforeseen needs. What was achieved in these studies was only achieved through his generosity and kindness. Nevertheless, the final scale of the project ended up far less than as first conceived.

In terms of experiment design, the studies reported were only able to operationalise the treatments a limited number of times vis-à-vis the actual text collections that concordances were extracted from. With more data processing capacity, it would have been possible to cross-validate results in concordances extracted from multiple corpora of the same genre or level of corpus homogeneity, and to do so would have allowed greater confidence in the characterisation of exactly what the active variable in each case was. While I am quite confident that the active variable in Chapter 6 was indeed corpus homogeneity, more data would help us ascertain whether particular ways of narrowing corpus composition are more effective than others. In regards to Chapter 5, I rather suspect we might find considerable variability between concordances drawn from multiple corpora that are defined in terms of a single super genre, depending on the exact composition of the corpora tested and finer grained accounts of the types of texts they are composed of. The studies reported in Chapters 5 and 6 successfully demonstrated that corpus composition can have a significant and reliable effect of considerable magnitude on a concordance's word frequencies and type token ratio, but we are far from being able to predict such effects in relation to concordances extracted from corpora that were not examined in this study.

Again, Punch's (2003) remarks on the value of small-scale studies not only seem apt, but also transferable to other paradigms of research: the lexical analysis studies reported in this thesis are best seen as a first step in contributing to the weight of research knowledge available. They show that the language presented in concordances both varies, and is in principle predictable, underling the importance and feasibility of examining the lexical qualities of the language presented in concordances as a variable in learner use of concordances. However, what research such as this fails to do is empirically investigate the link between observations of the lexical properties of concordances and actual learner use. The results have implications for learner use, but how variables such as vocabulary load and lexical repetition actually interact with learner use is fundamentally speculative. This is a topic that will be picked up in the section below, *Future directions for research*.

### **Future directions for research**

As indicated earlier in the thesis, there are numerous avenues for further research on both the construct of learner use of concordances and individual variables. This subsection will discuss three particularly significant avenues of future research: research on prototypical models of *analysis* and *reference* type use; cross-validation studies of the research reported in Chapters 5 and 6; and research that examines the significance of the variables identified in these studies through direct observation of use.

This thesis has argued on both a theoretical and an empirical basis that two basic models of concordance use can be identified, *analysis* type use and *reference* type use. It would be useful to investigate such a proposition further, as it might help us to simplify the construct of concordancing considerably, thereby facilitating the design of studies and the generalisation of findings. At present, the model is just that: an empirically informed model. Triangulation of the results reported in this thesis with observational studies or interview data would be one way in which we could provide stronger evidence in favour of such a model, help to refine the model, or suggest alternative interpretations of the data presented here.

As suggested above, the studies reported in Chapters 5 and 6 would both benefit from cross-validation studies that repeated the experiments reported in those chapters but substituted

the corpora used in these studies with alternative collections of texts compiled to equivalent specifications. Realistically, such studies would probably identify a degree of divergence in the results obtained, and identifying the source of such divergence would be a helpful step towards specifying the operative factors in corpus composition and thereby pushing the research agenda towards predictive specification of factors in corpus composition vis-à-vis the scores of concordances on theoretically important measures of lexical properties.

More generally, the research reported in this thesis would be complemented by research that operationalised concordancing preferences and the lexical properties of concordances in studies that involved actual observation of people using concordances. With so many variables that could affect learning outcomes having been identified, studies that emphasized observation of performance over observation of learning outcomes might be particularly useful in helping to ascertain the relative importance of these variables. Within the paradigm of eye movement research (Rayner, 1998, 2009), eye movements underlie online processes of reading and attention. It would be very worthwhile for research to explore the extent to which factors such as concordance format, concordance task, type-token ratio and word frequency interact with each other in both learners' and proficient users' online processing of concordances as texts. Such research would give us greater confidence in such factors as they pertain to the affordances and authentication burdens of different operationalisations of concordancing, underpinning the importance of the research reported here but also helping to unpack the relative importance of each variable in particular contexts of use.

### **Recommendations for pedagogical practice**

Although the findings and implications reported so far in this thesis would benefit from further exploration and confirmation through further research, the current findings have interesting implications for pedagogical practice. The recommendations made here should of course be treated cautiously, but they also represent the discussion of findings that are probably most interesting and accessible to the general applied linguistics reader.

One major theme of this thesis has been the importance of specifying the language learning affordance sought and matching this to an operationalisation of concordancing that

maximises provision of the specified affordance while minimising potential barriers to learner use, such as authentication burden. The research reported here can be interpreted as suggesting that, by default, learners should be presented with sentence format citations, use concordances as reference resources, use concordances of texts with relatively low vocabulary burdens, and use concordances of relatively homogenous collections of texts. All of these factors within the operationalisation of concordancing are likely to lower the authentication burden of concordances and so make concordances easier for learners to use. If concordances are simply envisaged as a source of massed encounters with target vocabulary or an opportunity for discovery learning, without good reason to operationalise concordancing differently, these are probably the optimal conditions for learner use of concordances. We might also note that, if learners are instructed to ignore citations containing many unknown words, learners may be able to make use of concordances of authentic texts at surprisingly low proficiency levels; a site vocabulary of the 4000 most common words would appear to be sufficient for the average citation in a concordance, even when news or academic texts have been concordanced.

If the affordances sought are based on pattern observation or generalisation, as in the case of many typical data-driven learning type tasks, then teachers might then consider using KWIC format concordances. It has also been suggested that, in relation to these types of affordances, corpus homogeneity might be a particularly important variable. In hands-on use, using more homogenous concordances can significantly increase the amount of lexical repetition the learner has to work with. If teachers are taking a more active mediation role in learner concordance use, concordancing more homogenous corpora should make the selection of citations to include in a pattern or generalisation focused exercise much less time consuming. The gains in lexical repetition available would appear to operate at multiple levels: delimiting corpora on the basis of genre or domain would appear to make some gains, but locally compiled corpora focusing on particular topics would appear to have the most potential.

A second important theme identified in this thesis is the need for critical evaluation of whether or not concordancing is the most appropriate means of availing a learner of a particular affordance. Stepping back somewhat, teachers should critically consider whether or not learner use of concordances really is the best means of affording learners a particular type of input text. Can concordance data be used by the teacher in production of alternative teaching materials? Are there commercial resources available that might afford equivalent language learning

opportunities but more efficiently? These are important questions to ask. Graeme Kennedy put the case well (1998, pp. 293-294):

Excessive claims should not be made about concordancing. It is not a language teaching methodology nor a panacea but one among many techniques or aids which may be used to facilitate learning for some learners.

It is important that teachers considering using concordances with learners consider the issues involved critically. However, that being said, one affordance of concordances that appears to offer fairly unique language learning opportunities is using concordances to access varieties of language poorly attested in commercially available or teacher prepared resources. Hence, the third theme in this thesis is the potential for concordances to furnish learners with a self-access language learning resources for use in editing, composition or translation.

As well as the default *reference* type operationalisation of concordancing advocated above as having the potential to lower the authentication burden of concordances, there would also seem to be very high value in providing learners with either tailor-made text collections for use in autonomous concordancing, or else providing learners with the skills to compile such text collections for themselves (cf. research by Maggie Charles such as Charles, 2012, 2015). Fundamentally, this thesis finds reason to endorse learner use of concordances in LSP contexts in which learners possess a sight vocabulary of at least the most frequent 4000 word families, are using small text collections as closely matched to their specific field of target use as possible, are using concordances as a source of examples that are relevant to an immediate context of productive use, and are, by default, engaging with citations presented in sentence format. There may of course still be many challenges for such learners to overcome, and it would still be recommended that learners are using familiar texts, that they have access to contextualising information about citations and, preferably, convenient access to the whole texts citations are drawn from; however, the aforementioned operationalisation of concordancing would appear to offer the optimal balance of affordances that are unique to learner use of concordances and legitimate concern over the authentication burden inherent in concordance use. To reiterate, this does not imply that concordances cannot be used by teachers to provide learners with input texts that possess particular qualities, only that concordances may have a special place in the hands of



learners in language for specific purposes contexts, and that an appropriate model for such use may not be the *analysis* type use of concordances associated with linguistic research, but a *reference* type use indicative of the use of concordances in other contexts. Indeed, it is tempting to add my own metaphor to the existing theoretical perspectives available: perhaps it is time we started to think in terms of *learner as editor*.



## Appendices

### Appendix 1: Screenshots of the questionnaire

Please note: the survey used ‘*display logic*’ to determine exactly which questions would be displayed. Information from answers at the beginning of the survey determined exactly which choices and questions were displayed at later points of the survey. Consequently, the screenshots presented here are only indicative of what was displayed to each individual respondent; how the exact options presented to each respondent were determined is indicated in Table 8 (pp. 91-92).

### Survey of Concordance Use: Information Sheet

**Researcher:** Oliver James Ballance

**Supervisors:** Dr. Averil Coxhead & Dr. Anna Siyanova-Chanturia

**Affiliation:** School of Linguistics and Applied Language Studies, Victoria University of Wellington, New Zealand

#### Background to the survey

Concordances can be made in many different ways, and used for many different purposes. This survey is designed to identify the ways in which concordances are used by different types of users. It is part of a broader research project about how concordances may be able to support language learning. It is not about other corpus tools, such as word lists or key words. It aims to collect data from as wide a range of concordance users as possible, including (but not only) inexperienced users, infrequent users and 2nd language users.

#### About the survey

The survey consists of 6 sets of short answer questions, and piloting suggests that it will take around 10-15 minutes to complete. The questions are about how you use concordances, why you use concordances, and your attitude to different features of concordances.

#### This project has been approved by the Victoria University of Wellington Human Ethics Committee.

1. Informed consent and voluntary participation in filling out the survey is implied by ticking the box next to “I consent to taking this survey” below. If you wish to withdraw from this survey while taking it, please close the relevant window in your browser. If you wish your data to be withdrawn from the survey at a later date, please contact me before the 1st of April 2015 and I will see to it that your data is deleted and not used.
2. Data from this study will be kept in password protected files for a maximum of 8 years, after which time it will be destroyed. The survey data collected will only be used for research, publications and presentations based on this research. No one other than me or my supervisors will have access to this information.
3. If you would like to receive an email with a summary of the results of the research, the final item in the survey offers you this option.
4. If you have any questions, would like to receive further information about the project, or would like me to send you a summary of the information you provide, please contact me at:  
Oliver.Ballance@vuw.ac.nz

---

☐ I consent to taking this survey.



### The Questions

1. The questions aim to collect data on your general preferences and typical patterns of use; they are not concerned with rare or exceptional circumstances.
2. There are no right or wrong answers.
3. Do not worry if you cannot see the difference between two questions: parts of this survey use a multiple-questions research method in which several related questions will be asked about a single broader topic.



[Back](#) [Next](#)

**Use** (Section 1 of 6)

**How long have you been using concordances for?**

*Please tick one*

- ☐ Less than a year
- ☐ 1 to 3 years
- ☐ 3 to 10 years
- ☐ 10 to 20 years
- ☐ More than 20 years
- ☐ Please feel free to comment if you wish

**How often do you use concordances?**

*Please tick one*

- ☐ Very rarely (I have only used concordances a few times)
- ☐ Occasionally (I use concordances every now and again, but are not a regular part of my work or study)
- ☐ Quite often (It's not unusual for me to be using concordances for one reason or another)
- ☐ Very often (Concordances are key component of my work or study)
- ☐ Please feel free to comment if you wish

☐



### Which languages do you use concordances of?

Please tick as many as apply

	Yes	Occasionally	No
I use concordances of my 1st language	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I use concordances of a 2nd language	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I use bilingual concordances	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

### What do you use concordances for?

Please tick as many as apply

*I use concordances ...*

	Yes	Occasionally	No
for translation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
as a reference resource	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
to learn languages	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
in teaching	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
for general interest	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
for lexicography	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
for linguistic research (please specify)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
for socio-cultural research (please specify)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
other (please specify)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Back Next

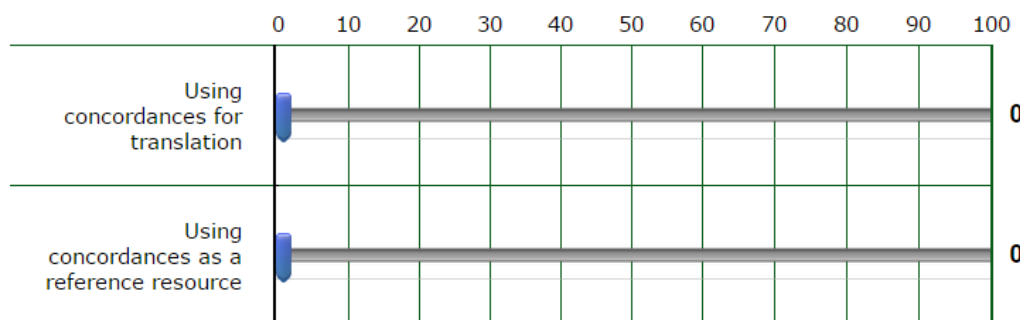


**Please estimate what percentage of your time using concordances is focused on each.**

*Your answers should sum to 100%!*

[Example answers](#)

If you cannot answer, or do not wish to answer, this question, please leave this question blank and click next.



Please feel free to leave a comment if you wish:



[Back](#) [Next](#)

**Important**

In the rest of this survey I would like you to answer the questions with just one type of concordance use in mind.

Please choose which use you will think of while answering the rest of the survey:

- ☐ for translation
- ☐ as a reference resource
- ☐ other (please specify)

Please choose the type of concordance the rest of your answers will refer to:

- ☐ Concordances of my 1st language
- ☐ Concordances of a 2nd language

Please feel free to comment if you wish:

[Back](#) [Next](#)





**When you use a concordance, which of the following are you interested in?**

*Please tick as many as apply*

- ☐ Meaning/Semantics/Usage
- ☐ Grammar/Syntax
- ☐ Collocation/Colligation
- ☐ Register/Genre
- ☐ Pragmatics
- ☐ Orthography
- ☐ Phonology
- ☐ Social/Cultural
- ☐ Literary use/Stylistics/Creativity
- ☐ other

Please feel free to comment if you wish:

☐

[Back](#) [Next](#)

**Please indicate the extent to which you agree with the following statements.**

*please tick one box for each statement*

***When I use a concordance...***

	Not at all	Not much	I neither agree nor disagree	Quite a lot	Greatly	I do not understand the question
I read through a concordance until I find a useful citation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I read through a concordance looking at which things are repeated	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I focus on relationships between citations	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I focus on particular citations	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I am searching for general rules or descriptions	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I am searching for particularly clear examples	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I read whole citations	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I examine the words around the node	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I am looking for patterns	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I am trying to find particular citations with relevance to a particular situation of use outside of the concordance (such as a sentence I am working on)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please feel free to comment if you wish:



Back Next



### Corpus used (Section 2 of 6)

Concordances are made by searching collections of texts (a corpus). One major type of corpus is a large general corpus, such as the BNC, or COCA, that tries to represent a wide variety of language use. Another major type of corpus tries to focus on a comparatively narrow range of language use, such as a particular genre, domain, mode or location of language use: for example, a corpus of British newspaper articles from the 1990s.

### To what extent do you agree with the following statements?

*please tick one box for each statement*

	Not at all	Not much	I neither agree nor disagree	Quite a lot	Greatly	I do not understand the question
I find it most useful when the citations in the concordance are from a narrow, focused corpus	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I find it most useful when the citations in the concordance are from a large general corpus	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
In my experience, concordances are easiest to use when the citations in the concordance are from a narrow, focused corpus	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
In my experience, concordances are easiest to use when the citations in the concordance are from a large general corpus	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I like to use concordances of a large general corpus	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I like to use concordances of narrow, focused corpora	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I think that a corpus that I have made myself will suit my purposes best	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please feel free to comment if you wish:

**While answering the questions on the next pages, please follow these definitions:**

**Concordance** = a text that presents citations of a node (e.g.  , 1, 2)

**Node** = the word or phrase used to make a concordance (e.g. word)

**Citation** = each example of a word or phrase in use (e.g. 1, 2, 3)

For example, in concordance **1** below, the node is **centrally aligned**, and 3 citations are presented on a **single line** each:

**1**

- 1 know if babies experience pain as the word is understood by adults. However,
- 2 or form, but we are anxious about the word "acceptability" later in the Clause'.
- 3 . I am glad Tom Regan mentioned this word, because it is often used in the

But in concordance **2**, the citations are **aligned to the left margin**, and each of the 3 citations is presented on **multiple lines**:

**2**

- 1 In summary, we do not know and perhaps we can never know if babies experience pain as the word is understood by adults.
- 2 We do not wish to change that in any way, shape or form, but we are anxious about the word "acceptability" later in the Clause'.
- 3 I am glad Tom Regan mentioned this word, because it is often used in the wrong sense, as he implied, simply for anyone whose ideals are rather extreme.

Please feel free to comment if you wish:



Back Next



### Alignment (Section 3 of 6)

There are two common ways for concordances to present citations: with each citation aligned to the left margin of the page, or with the node aligned to the centre of the page.

#### To what extent do you agree with the following statements?

*please tick one box for each statement*

	Not at all	Not much	I neither agree nor disagree	Quite a lot	Greatly	I do not understand the question
In my experience, concordances are easiest to use when each citation is aligned to the left margin of the page	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
In my experience, concordances are easiest to use when the node of each citation is aligned to the centre of the page	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I find it most useful when each citation is aligned to the left margin of the page	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I find it most useful when the node of each citation is aligned to the centre of the page	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
It is important to me that the citations in a concordance are presented like normal sentences	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
It is important to me that the node is presented in a single column in the middle of the concordance	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please feel free to comment if you wish:



[Back](#) [Next](#)

### Size (Section 4 of 6)

Concordances also differ in terms of the size of the citations they present: in particular, whether each citation is presented on a single line, or on multiple lines.

### To what extent do you agree with the following statements?

*please tick one box for each statement*

	Not at all	Not much	I neither agree nor disagree	Quite a lot	Greatly	I do not understand the question
In my experience, concordances are easiest to use when each citation is presented on a single line of text	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
In my experience, concordances are easiest to use when each citation presents multiple lines of text	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I find concordances most useful when citations present complete sentences	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I find concordances most useful when sentences that are longer than the screen width are cut short	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
For me, the sentences around the node are very important.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
For me, the words closest to the node are very important.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The more context I can see, the more useful I find the citations	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please feel free to comment if you wish:

Back Next

### Citation ordering (Section 5 of 6)

The citations that are presented in a concordance can be ordered in two main ways: in relation to the text each citation comes from (for instance, by file name or by corpus-subsection), or in relation to the words contained in each citation (for instance, one word to the left or right of the node).

### To what extent do you agree with the following statements?

*please tick one box for each statement*

	Not at all	Not much	I neither agree nor disagree	Quite a lot	Greatly	I do not understand the question
In my experience, concordances are easiest to use when the order of citations is determined by the text each citation comes from	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
In my experience, concordances are easiest to use when the order of citations is determined by the words around the node	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I find it most useful when concordances are organised in relation to the texts that citations come from	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
I find it most useful when concordances are organised by the words around the node	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Concordances that present citations from related texts together suit my purposes best.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Concordances that present citations with similar wording together suit my purposes best.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please feel free to comment if you wish:



**Biographical information** (Section 6 of 6)

**How would you describe yourself in terms of nationality or ethnic group?**

**What do you consider to be your first language or languages?**

**How would you describe your profession, field of research, or study?**

*For example: a language teacher doing a masters degree in SLA, a researcher working in translation theory, or an undergraduate student studying law.*

Please feel free to leave a comment if you wish:



[Back](#) [Next](#)





If you would like to receive a short report on the findings of this survey, please leave your email address in the box below:

If you would consider discussing issues arising from this survey via a short Skype interview at some point in the future, please leave your email address in the box below:

Please feel free to leave a comment if you wish:



[Back](#) [Next](#)

Survey Powered By [Qualtrics](#)



Thank you very much for completing the survey: your time and effort are really appreciated.

If you happen to know of anyone else who may like to take this survey, please feel free to forward the link to this survey to them, or to copy-and-paste it elsewhere.



Survey Powered By [Qualtrics](#)



### List of references

- Ädel, A. (2010). Using corpora to teach academic writing: Challenges for the direct approach. In M. C. Campoy-Cubillo, B. Belles-Fortuño, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to ELT*. (pp. 39-55). London, UK: Continuum.
- Adolphs, S. (2006). *Introducing electronic text analysis: A practical guide for language and literary studies*. London, UK; New York, NY: Routledge.
- Alcaraz Calero, J., M., & Pérez-Paredes, P. (2008). What do annotators annotate? An analysis of language teachers' corpus pedagogical annotation. In A. Frankenberg-Garcia (Ed.), *Proceedings of the 8th Teaching and Language Corpora Conference* (pp. 27-37). Lisbon, Portugal: Associação de Estudos e de Investigação Científica do ISLA-Lisboa.
- Allan, R. (2006). Data-driven learning and vocabulary: Investigating the use of concordances with advanced learners of English, . *Centre for Language and Communication Studies*(66), 1-52.
- Allan, R. (2009). Can a graded reader corpus provide 'authentic' input? *ELT Journal*, 63(1), 23-32. doi:10.1093/elt/ccn011
- Allan, R. (2010). Concordances versus dictionaries: Evaluating approaches to word learning in ESOL. In R. Chacón-Beltrán, C. Abello-Contesse, & M. d. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 112-125). Bristol, UK; Buffalo, NY; Toronto, Canada: Multilingual Matters.
- Aston, G. (2001). Learning with corpora: an overview. In G. Aston (Ed.), *Learning with corpora* (pp. 7-45). Houston, TX: Athelstan.
- Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht, The Netherlands; Boston, MA: Kluwer Academic.
- Ballance, O. J. (2012). Corpus technology and vocabulary testing in EAP. *Journal of Second Language Teaching and Research*, 1(2), 143-154.
- Barlow, M. (2004). Software for corpus access and analysis. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 205-211). Amsterdam, Netherlands: John Benjamins Publishing.
- Barlow, M., & Burdine, S. (2006). *Phrasal verbs*. Houston, TX: Athelstan Publications.
- Basturkmen, H. (2006). *Ideas and options in English for specific purposes*. Mahwah, N.J.; London, UK: Routledge.
- Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253-279.
- Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor, MI: University of Michigan Press.
- Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 225-235). Frankfurt am Main, Germany: Peter Lang.
- Bernardini, S. (2002). Exploring new directions for discovery learning. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 165-182). Amsterdam, Netherlands; New York, NY: Rodopi.
- Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J. Sinclair (Ed.), *How to use corpora in language teaching*. Amsterdam, The Netherlands; Philadelphia, PA John Benjamins.
- Bernstein, I. H., Garbin, C., & Teng, G. (1988). *Applied multivariate analysis*. New York. NY: Springer-Verlag.

- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Biber, D., & Reppen, R. (2015). Introduction. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 1-8). Cambridge, UK: Cambridge University Press.
- Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, 93(3), 498-520. doi:10.1037/0022-0663.93.3.498
- Bitchner, J. (2012). *Writing an applied linguistics thesis or dissertation: A guide to presenting empirical research*. Beijing, China: Foreign Language Teaching and Research Press.
- Boulton, A. (2008). DDL: Reaching the parts other teaching can't reach? In A. Frankenberg-Garcia (Ed.), *Proceedings of the 8th Teaching and Language Corpora Conference* (pp. 38-44.). Lisbon, Portugal: Associação Científica do ISAL-Lisboa.
- Boulton, A. (2009a). Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics*., 35(1), 81-106.
- Boulton, A. (2009b). Testing the limits of data-driven learning: language proficiency and training. *ReCALL*, 21(01), 37-54.
- Boulton, A. (2010a). Data-driven learning: Taking the computer out of the equation. *Language Learning*, 60(3), 534-572. doi:10.1111/j.1467-9922.2010.00566.x
- Boulton, A. (2010b). Learning outcomes from corpus consultation. In M. Moreno Jaén, F. S. Valverde, & M. C. Pérez (Eds.), *Exploring new paths in language pedagogy: Lexis and corpus-based language teaching* (pp. 129-144). London, UK: Equinox.
- Boulton, A. (2011). Data-driven learning: The perpetual enigma. In S. Goźdz-Roszkowski (Ed.), *Explorations across languages and corpora: PALC 2009* (pp. 563-579). Frankfurt, Germany: Peter Lang.
- Boulton, A., & Pérez-Paredes, P. (2014). ReCALL special issue: Researching uses of corpora for language teaching and learning Editorial Researching uses of corpora for language teaching and learning. *ReCALL : the Journal of EUROCALL*, 26(2), 121-127. doi:10.1017/S0958344014000068
- Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, 17(1), 47-64.
- Braun, S. (2006). ELISA: A pedagogically enriched corpus for language learning purposes. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy* (pp. 25-47). Frankfurt am Main, Germany: Peter Lang.
- Braun, S. (2007). Integrating corpus work into secondary education: from data-driven learning to needs-driven corpora. *ReCALL*, 19(3), 307-328. doi:10.1017/S0958344007000535
- Breyer, Y. A. (2009). Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning*, 22(2), 153-172. doi:10.1080/09588220902778328
- Breyer, Y. A. (2011). *Corpora in language teaching and learning: Potential, evaluation, challenges*. Frankfurt am Main, Germany: Peter Lang.
- Burdine, S., & Barlow, M. (2008). *Business phrasal verbs*. Houston, TX: Athelstan Publications.
- Burnard, L. (2002). The BNC: Where did we go wrong? In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis: Proceedings of the fourth international conference on teaching and language corpora, Graz 19-24 July, 2000*. (pp. 51-70). Amsterdam, Netherlands; New York, NY: Rodopi.

- Chambers, A. (2010). What is data-driven learning? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 345-358). London, UK; New York, NY: Routledge.
- Chambers, A., & O'Sullivan, I. (2004). Corpus consultation and advanced learners: writing skills in French. *ReCALL*, 16(01), 158-172. doi:10.1017/S0958344004001211
- Chan, T.-P., & Liou, H.-C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. *Computer Assisted Language Learning*, 18(3), 231-250. doi:10.1080/09588220500185769
- Charles, M. (2011). Using hands-on concordancing to teach rhetorical functions: Evaluation and implications for EAP writing classes. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning* (pp. 26-43). London, UK; New York, NY: Continuum.
- Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, 31(2), 93-102. doi:10.1016/j.esp.2011.12.003
- Charles, M. (2015). Same task, different corpus: The role of personal corpora in EAP classes. In A. Lenko & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 131-154). Amsterdam, The Netherlands; Philadelphia, PA: John Benjamins.
- Cheng, W. (2012). *Exploring corpus linguistics: language in action*. London, UK; New York, NY: Routledge.
- Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, 11(4), 411-433.
- Chujo, K., Oghigian, K., & Akasegawa, S. (2015). A corpus and grammatical browsing system for remedial EFL learners. In A. Leńko-Szymańsка & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 109-128). Amsterdam, The Netherlands; Philadelphia, PA: John Benjamins.
- Chujo, K., & Utiyama, M. (2005). Understanding the role of text length, sample size and vocabulary size in determining text coverage. *Reading in a foreign language*, 17(1), 1-22.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4), 335-359. doi:10.1016/S0022-5371(73)80014-3
- Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System*, 25(3), 301-315.
- Cobb, T. (1999). Breadth and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning*, 12(4), 345-360.
- Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a foreign language*, 22(1), 181-200.
- Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 478-497). New York, NY: Cambridge University Press.
- Cobb, T., & Horst, M. (2001). *Growing Academic Vocabulary with a Collaborative On-Line Database*.
- Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26(2), 202-224. doi:10.1017/S0958344014000019
- Coxhead, A. (2014). Corpus linguistics and vocabulary teaching: Perspectives from English for specific purposes. In M. Gotti (Ed.), *Corpus analysis for descriptive and pedagogical*

- purposes* (pp. 289-301). Bern, Switzerland: Peter Lang.
- Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, 16(3), 129-147. doi:10.1016/j.jslw.2007.07.002
- Craik, F. I. M. (2002). Levels of processing: Past, present... and future? *Memory*, 10(5-6), 305-318. doi:10.1080/09658210244000135
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671-684.
- Daskalovska, N. (2013). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, 28(2), 130-144. doi:10.1080/09588221.2013.803982
- Davies, M. (2015). Corpora: An introduction. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 11-31). Cambridge, UK: Cambridge University Press.
- Dörnyei, Z., & Taguchi, T. (2010). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). Beijing, China: Foreign Language teaching and Research Press.
- Ellis, N. C. (1995). Consciousness in second language acquisition: A review of field studies and laboratory experiments. *Language Awareness*, 4(3), 123-146. doi:10.1080/09658416.1995.9959876
- Ellis, N. C. (2008). The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The Modern Language Journal*, 92(2), 232-249.
- Ellis, N. C. (2015). Implicit and explicit language learning: Their dynamic interface and complexity. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 3-23). Amsterdam, The Netherlands: John Benjamins.
- Flowerdew, L. (2012a). *Corpora and language education*: Palgrave Macmillan.
- Flowerdew, L. (2012b). Corpora in the classroom: An applied linguistic perspective. In K. Hyland, M. H. Chau, & M. Handford (Eds.), *Corpus applications in applied linguistics* (pp. 208-224). London, UK: Continuum.
- Flowerdew, L. (2015). Corpus-driven learning and language learning theories. In A. Lenko & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15-36). Amsterdam, The Netherlands; Philadelphia, PA: John Benjamins.
- Franken, M. (2014). The nature and scope of student search strategies in using a web derived corpus for writing. *The Language Learning Journal*, 42(1), 85-102. doi:10.1080/09571736.2012.678013
- Frankenberg-Garcia, A. (2014a). How language learners can benefit from corpora, or not. *Recherches en didactique des langues et des cultures: les cahiers de l'acedle*, 11(1), 93-110.
- Frankenberg-Garcia, A. (2014b). The use of corpus examples for language comprehension and production. *ReCALL*, 26, 128-146. doi:10.1017/S0958344014000093
- Gabreilatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? . *TESL-EJ*, 8(4), 1-37.
- Gavioli, L. (1997). Exploring texts through a concordancer: Guiding the learner. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 83-99). London, UK; New York, NY: Longman.
- Gavioli, L. (2001). The learner as researcher: Introducing corpus concordancing in the

- classroom. In G. Aston (Ed.), *Learning with corpora* (pp. 108-137). Houston, TX: Athelstan.
- Gavioli, L. (2005). *Exploring corpora for ESP learning*. Amsterdam, Netherlands: John Benjamins.
- Gavioli, L., & Aston, G. (2001). Enriching reality: language corpora in language pedagogy. *ELT Journal*, 55(3), 238-246.
- Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 359-370). London, UK; New York, NY: Routledge.
- Gordani, Y. (2013). The effect of the integration of corpora in reading comprehension classrooms on English as a Foreign Language learners' vocabulary development. *Computer Assisted Language Learning*, 26(5), 430-445.
- Götz, S., & Mukherjee, J. (2006). Evaluation of data-driven learning in university teaching: a project report. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy* (pp. 49-67). Frankfurt am Main, Germany: Peter Lang.
- Grabe, W., & Stoller, F. (2001). *Teaching and researching reading* Beijing, China: Foreign Language Teaching and Research Press.
- Hanks, P. (2008). The lexicographical legacy of John Sinclair. *International Journal of Lexicography*, 21(3), 219-229.
- Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Cambridge, MA: MIT Press.
- Hill, D. (2008). Graded readers in English. *ELT Journal*, 62(2), 184-204. doi:10.1093/elt/ccn006
- Hoey, M. (2005). *Lexical priming*. London, UK; New York, N.Y.: Routledge.
- Hoey, M. (2015). Words and their neighbours. In J. R. Taylor (Ed.), *The Oxford handbook of the word* (pp. 141-153). Oxford, UK: Oxford University Press.
- Honeyfield, J. (1989). A typology of exercises based on computer-generated concordance material. *Guidelines: A Periodical for Classroom Language Teachers*, 11(1), 42-50.
- Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading. *Reading in a foreign language*, 13(1), 403-430.  
doi:<http://helicon.vuw.ac.nz/login?url=http://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehc hao.pdf>
- Hudson, T., & Llosa, L. (2015). Design issues and inference in experimental L2 research. *Language Learning*, 65, 76-96. doi:10.1111/lang.12113
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, UK: Cambridge University Press.
- Hunston, S. (2010). How can a corpus be used to explore patterns? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. London, UK; New York, NY: Routledge.
- Hyland, K. (2015). Corpora and written academic English. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 292-308). Cambridge, UK: Cambridge University Press.
- Jarvis, S. (2013). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam, The Netherlands; Philadelphia, PA: John Benjamins.
- Jarvis, S., & Daller, M. (2013). Introduction. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam, The Netherlands; Philadelphia, PA: John Benjamins.

- Johns, T. (1986). Micro-concord: A language learner's research tool. *System*, 14(2), 151-162.
- Johns, T. (1991a). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns & P. King (Eds.), *Classroom Concordancing*. (pp. 27-45). Birmingham, UK: University of Birmingham. .
- Johns, T. (1991b). Should you be persuaded: Two samples of data-driven learning materials. *English language research journal*, 4, 1-16.
- Johns, T. (2002). Data-driven learning: The perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis: Proceedings of the fourth international conference on teaching and language corpora, Graz 19-24 July, 2000*. (pp. 107-117). Amsterdam, the Netherlands; New York, NY: Rodopi.
- Johns, T., Hsingchin, L., & Lixun, W. (2008). Integrating corpus-based CALL programs in teaching English through children's literature. *Computer Assisted Language Learning*, 21(5), 483-506.
- Johnson, K. (2008). *Quantitative methods in linguistics*. Malden, MA: Blackwell.
- Kaur, J., & Hegelheimer, V. (2005). ESL students' use of concordance in the transfer of academic word knowledge: An exploratory study. *Computer Assisted Language Learning*, 18(4), 287-310. doi:10.1080/09588220500280412
- Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning and Technology*, 5(3), 77-90.
- Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. . *Language Learning & Technology*, 14(1), 28-44.
- Kennedy, G. (1998). *An introduction to corpus linguistics*. Beijing, China: Foreign Language Teaching and Research Press (2000).
- Kenning, M.-M. (2010). What are parallel and comparable corpora and how can we use them? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 487-500). London, UK; New York, NY: Routledge.
- Kilgarrieff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2), 263-276. doi:10.1515/cllt.2005.1.2.263
- Kilgarrieff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008, 15-19 July). *GDEX: Automatically finding good dictionary examples in a corpus*. Paper presented at the EURALEX, Barcelona, Spain.
- Kilgarrieff, A., Marcowitz, F., Smith, S., & Thomas, J. (2015). Corpora and language learning with the Sketch Engine and SKELL. *Revue française de linguistique appliquée*, 20(1), 61-80.
- Koosha, M., & Jafarpour, A. A. (2006). Data-driven learning and teaching collocation of prepositions: The case of Iranian EFL adult learners. *Asian EFL Journal*, 8(4), 192-209.
- Kübler, N. (2011). Working with corpora for translation teaching in a French-speaking setting. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning* (pp. 62-80). London, UK; New York, NY: Continuum.
- Kübler, N., & Aston, G. (2010). Using corpora in translation. In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 501-515). London, UK; New York, NY: Routledge.
- Lai, S.-L., & Chen, H.-J. H. (2013). Dictionaries vs concordancers: Actual practice of the two different tools in EFL writing. *Computer Assisted Language Learning*, 28(4), 341-363. doi:10.1080/09588221.2013.839567



- Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York, NY; London, UK: Routledge.
- Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a foreign language*, 22(1), 15-30.
- Laviosa, S. (2012). Corpora and translation studies. In K. Hyland, M. H. Chau, & M. Handford (Eds.), *Corpus applications in applied linguistics* (pp. 67-83). London, UK: Continuum.
- Lee, D. (2002). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis: Proceedings of the fourth international conference on teaching and language corpora, Graz 19-24 July, 2000*. (pp. 247-292). Amsterdam, Netherlands; New York, NY: Rodopi.
- Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25(1), 56-75.
- Leech, G. (1992). Corpora and theories of linguistic performance. *Directions in corpus linguistics*, 105-122.
- Leech, G., & Candlin, C. (1986). Introduction. In G. Leech & C. Candlin (Eds.), *Computers in English language teaching and research: Selected papers from the 1984 Lancaster symposium 'Computers in English language education and research'*. London, UK; New York, NY: Longman.
- Leńko-Szymańska, A. (2014). Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education. *ReCALL*, 26(Special Issue 02), 260-278. doi:10.1017/S095834401400010X
- Leńko-Szymańska, A., & Boulton, A. (Eds.). (2015). *Multiple affordances of language corpora for data-driven learning*. Amsterdam, The Netherlands; Philadelphia, PA: John Benjamins.
- Levy, M. (1990). Concordances and their integration into a word-processing environment for language learners. *System*, 18(2), 177-188. doi:10.1016/0346-251X(90)90052-7
- Levy, M., & Stockwell, G. (2006). *CALL dimensions: Options and issues in computer-assisted language learning*. London, UK; Mahwah, N.J.: Lawrence Erlbaum Associates.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84-99. doi:10.1037/1082-989X.4.1.84
- Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndsmills, NH: Palgrave Macmillan.
- Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research*, 3(1), 97-110.
- McCarthy, M., & Carter, R. (1997). Written and spoken vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy*. Cambridge, UK: Cambridge.
- McCarthy, M., & McCarten, J. (2012). Corpora and materials design. In K. Hyland, M. H. Chau, & M. Handford (Eds.), *Corpus applications in applied linguistics* (pp. 225-241). London, UK: Continuum.
- McCarthy, M., & O'Keeffe, A. (2010). Historical perspective: What are corpora and how have they evolved? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 3-13). London, UK; New York, NY: Routledge.

- McCarthy, P., & Jarvis, S. (2010). MTL, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381-392. doi:10.3758/BRM.42.2.381
- McKay, S. (1980). Teaching the syntactic, semantic and pragmatic dimensions of verbs. *TESOL Quarterly*, 14(1), 17-26. doi:10.2307/3586805
- Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, England: Multilingual Matters.
- Mishan, F. (2004). Authenticating corpora for language learning: a problem and its resolution. *ELT Journal*, 58(3), 219-227.
- Mukherjee, J. (2004). Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In U. Connor & T. A. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 239-250). Amsterdam, Netherlands; New York, NY: Rodopi.
- Mukherjee, J. (2006). Corpus linguistics and language pedagogy: the state of the art - and beyond. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy* (pp. 5-24). Frankfurt am Main, Germany: Peter Lang.
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, 63(1), 59-81. doi:10.1353/cml.2006.0049
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.
- Nation, I. S. P. (2015). Which words do you need? In J. R. Taylor (Ed.), *The Oxford handbook of the word* (pp. 568-581). Oxford, UK: Oxford University Press.
- Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy*. Cambridge, UK: Cambridge University Press.
- Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.
- Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, 65(1), 97-126. doi:10.1111/lang.12114
- Norris, J. M., Ross, S. J., & Schoonen, R. (2015). Improving second language quantitative research. *Language Learning*, 65(S1), 1-8. doi:10.1111/lang.12110
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers*, 32(3), 396-402. doi:10.3758/BF03200807
- O'Keeffe, A., & McCarthy, M. (2010). *The Routledge handbook of corpus linguistics*: Routledge.
- O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge, UK: Cambridge University Press.
- 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.
- Osbourne, J. (2004). Top-down and bottom-up approaches to corpora in language teaching. In U. Connor & T. A. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 251-265). Amsterdam, the Netherlands; New York, NY: Rodopi.
- Park, K. (2012). Learner-Corpus Interaction: A Locus of Microgenesis in Corpus-Assisted L2 Writing. *Applied Linguistics*, 33(4), 361-385.
- Partington, A. (2001). Corpus-based description in teaching and learning. In G. Aston (Ed.), *Learning with corpora* (pp. 63-84). Houston, TX: Athelstan.

- Pérez-Paredes, P. (2010). Corpus linguistics and language education in perspective: Appropriation and the possibilities scenario. In T. Harris & M. Moreno Jaén (Eds.), *Corpus linguistics in language teaching* (pp. 53-73).
- Pérez-Paredes, P., Sánchez-Tornel, M., Alcaraz Calero, J., & Jiménez, P. (2011). Tracking learners' actual uses of corpora: Guided vs non-guided corpus consultation. *Computer Assisted Language Learning*, 24(3), 233-253.
- Pérez-Paredes, P., Sánchez-Tornel, M., & Alcaraz Calero, J. M. (2012). Learners' search patterns during corpus-based focus-on-form activities: A study on hands-on concordancing. *International Journal of Corpus Linguistics*, 17(4), 482-515.
- Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning*, 65(1 [Supplement]), 9-36. doi:10.1111/lang.12111
- Punch, K. F. (2003). *Survey research: The basics*. London, England: SAGE.
- Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the Validity of Quantitative Measures in Applied Linguistics Research. *Language Learning*, 65, 37-75. doi:10.1111/lang.12112
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372-422. doi:http://dx.doi.org/10.1037/0033-2909.124.3.372
- Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8), 1457-1506. doi:10.1080/17470210902816461
- Rayson, P. (2015). Computational tools and methods for corpus compilation and analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge, UK: Cambridge University Press.
- Renouf, A. (2003). WebCorp: Providing a renewable data source for corpus linguistics. In S. Granger & S. Petch-Tyson (Eds.), *Extending the scope of corpus-based research: New applications, new challenges* (pp. 39-58). Amsterdam, The Netherlands: Rodopi.
- Reppen, R. (2010a). Building a corpus. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 31-37). London, UK; New York, NY: Routledge.
- Reppen, R. (2010b). *Using corpora in the language classroom*. Cambridge, UK; New York, NY: Cambridge University Press.
- Romer, U. (2008). Identification impossible? A corpus approach to realisations of evaluative meaning in academic writing. *Functions of language*, 15(1), 115-130.
- Römer, U. (2010). Using general and specialized corpora in English language teaching: past, present and future. In M. C. Campoy-Cubillo, B. Belles-Fortuño, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to ELT* (pp. 18-35). London, UK: Continuum.
- Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, 31, 205-225. doi:10.1017/S0267190511000055
- Ross, S. J., & Mackey, B. (2015). Bayesian approaches to imputation, hypothesis testing, and parameter estimation. *Language Learning*, 65, 208-227. doi:10.1111/lang.12118
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129-158.
- Schmidt, R. (1992). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, 13, 206-226. doi:10.1017/S0267190500002476
- Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

- Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, 12(3), 329-363. doi:10.1177/1362168808089921
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal*, 95(1), 26-43. doi:10.1111/j.1540-4781.2011.01146.x
- Scholfield, P. (1995). *Quantifying language: A researcher's and teacher's guide to gathering data and reducing it to figures*. Clevedon, UK; Bristol, PA: Multilingual Matters.
- Scott, M. (2016). WordSmith Tools version 7. Stroud, UK: Lexical Analysis Software.
- Seidlhofer, B. (Ed.) (2003). *Controversies in applied linguistics*. Oxford, UK; New York, NY: Oxford University Press.
- Sheskin, D. J. (2004). *Handbook of parametric and nonparametric statistical procedures* (3rd ed.). Boca Raton, FL; London, UK; New York, NY; Washington, DC: Chapman & Hall/CRC.
- Sinclair, J. (1991). *Corpus, concordance, collocation*: Oxford University Press.
- Sinclair, J. (2003). *Reading concordances: An introduction*. London, UK: Pearson.
- Sinclair, J. (2004). New evidence, new priorities, new attitudes. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 271-300). Amsterdam, Netherlands: John Benjamins.
- Sorell, J. (2015). Word frequencies. In J. R. Taylor (Ed.), *The Oxford handbook of the word* (First edition.. ed., pp. 68-88). Oxford, UK: Oxford University Press.
- Sripicharn, P. (2004). Examining native speakers' and learners' investigation of the same concordance data and its implications for classroom concordancing with EFL learners. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 233-244). Amsterdam, Netherlands: John Benjamins.
- Sripicharn, P. (2010). Preparing learners for language corpora. In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 371-384). London, UK; New York, NY: Routledge.
- Stevens, V. (1991a). Classroom concordancing: Vocabulary materials derived from relevant, authentic text. *English for Specific Purposes*, 10(1), 35-46.
- Stevens, V. (1991b). Concordance-based vocabulary exercises: a viable alternative to gap-fillers. In T. Johns & P. King (Eds.), *Classroom Concordancing: English Language Research Journal 4* (pp. 47-63). Birmingham, UK: Centre for English Language Studies, University of Birmingham.
- Stevens, V. (1995). Concordancing with language learners: Why? When? What? *CAELL*, 6(2), 2-9.
- Sun, Y.-C. (2003). Learning process, strategies and web-based concordancers: a case study. *British Journal of Educational Technology*, 34(5), 601-613. doi:10.1046/j.0007-1013.2003.00353.x
- Sun, Y.-C., & Wang, L.-Y. (2003). Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer Assisted Language Learning*, 16(1), 83-94.
- Sutarsyah, C., Nation, I. S. P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based study. *RELC Journal*, 25(2), 34-50.
- Teubert, W., & Čermáková, A. (2007). *Corpus linguistics: A short introduction*. London, UK; New York, NY: Continuum.
- Thurston, J., & Candlin, C. N. (1997). *Exploring academic English: A workbook for student essay writing*. Sydney, Australia: NCELTR.

- Thurston, J., & Candlin, C. N. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes*, 17(3), 267-280. doi:10.1016/S0889-4906(97)00013-6
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam, The Netherlands; Philadelphia, PA: John Benjamins.
- Tognini-Bonelli, E. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 14-27). London, UK; New York, NY: Routledge.
- Tono, Y., Satake, Y., & Miura, A. (2014). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL*, 26, 147-162. doi:10.1017/S095834401400007X
- Tribble, C. (1997). *Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching*. Paper presented at the PALC, Lodz.
- Tribble, C. (2010). What are concordances and how are they used? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 167-183). London, UK; New York, NY: Routledge.
- Tribble, C. (2013). Corpora in the language-teaching classroom. In C. A. Chapelle (Ed.), *The Encyclopaedia of Applied Linguistics*. Chichester, UK: Blackwell.
- Tribble, C. (2015). Teaching and language corpora: Perspectives from a personal journey. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 37-62). Amsterdam, Netherlands: John Benjamins.
- Tribble, C., & Jones, G. (1990). *Concordances in the classroom: A resource guide for teachers*. Houston, TX: Athelstan.
- Tribble, C., & Jones, G. (1997). *Concordances in the classroom: A resource guide for teachers*. Houston, TX: Athelstan.
- van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93-115). Cambridge, UK; New York, NY: Cambridge University Press.
- 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. doi:10.1080/09588220902778294
- Wallace, M., & Wray, A. (2011). *Critical reading and writing for postgraduates* (2nd ed.). Los Angeles, LA; London, UK: Sage.
- Warren, M. (2011). Using corpora in the learning and teaching of phraseological variation. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning* (pp. 153-166). London, UK; New York, NY: Continuum.
- Webb, S. A., & Chang, A. C.-S. (2012). Second language vocabulary growth. *RELC Journal*, 43(1), 113-126. doi:10.1177/0033688212439367
- Wible, D., Chien, F.-Y., Kuo, C.-H., & Wang, C. C. (2002). A lexical difficulty filter for language learners. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis: Proceedings of the fourth international conference on teaching and language corpora, Graz 19-24 July, 2000*. (pp. 147-154). Amsterdam, The Netherlands; New York, NY: Rodopi.
- Widdowson, H. G. (1998). Context, community, and authentic language. *TESOL Quarterly*, 32(4), 705-716. doi:10.2307/3588001
- Widdowson, H. G. (2000a). Object language and the language subject: On the mediating role of

- applied linguistics. *Annual Review of Applied Linguistics*, 20, 21-33.  
doi:10.1017/S0267190500200020
- Widdowson, H. G. (2000b). On the limitations of linguistics applied. *Applied Linguistics*, 21(1), 3-25.
- Widdowson, H. G. (2003). *Defining issues in English language teaching*. Oxford: Oxford University Press.
- Widdowson, H. G. (2004). *Text, context, and pretext: Critical issues in discourse analysis*. Malden, MA: Malden, MA : Blackwell Pub.
- Willis, J. (1998). Concordances in the classroom without a computer: Assembling and exploiting concordances of common words. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 44-66). Cambridge, UK: Cambridge University Press.
- Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.. ed.). Boston, MA: McGraw-Hill.
- Woods, A., Fletcher, P., & Hughes, A. (1986). *Statistics in language studies*. Cambridge, UK: Cambridge University Press.
- Yoon, C. (2011). Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes*, 10(3), 130-139.  
doi:http://dx.doi.org/10.1016/j.jeap.2011.03.003
- Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology*, 12(2), 31-48.
- Yoon, H., & Hirvela, A. (2004). ESL student attitudes towards corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257-283. doi:10.1016/j.jslw.2004.06.002