The pedagogical benefits of a lexical database (*SciE-Lex*) to assist the production of publishable biomedical texts by EAL writers

Natalia J. Laso and Suganthi John
University of Barcelona (Spain) and University of Birmingham (UK)
njlaso@ub.edu & s.p.john@bham.ac.uk

Abstract

Research has demonstrated that it is challenging for English as an Additional Language (EAL) writers to acquire phraseological competence in academic English and develop a good working knowledge of discipline-specific formulaic language. This paper aims to explore if *SciE-Lex*, a powerful lexical database of biomedical research articles, can be exploited by EAL writers to enhance their command of formulaic language in biomedical English published writing. Our paper builds on the challenges associated with formulaic language (namely collocations) for EAL writers, it reflects on the benefits of using a lexical database and it evaluates a pedagogical approach to helping EAL writers produce publishable texts. It specifically highlights results from two writing workshops conducted for EAL writers (medical researchers in the present study). The workshops involved medical researchers working on drafts of their writing using *SciE-Lex*. Our paper reports on the specific benefits of using *SciE-Lex* as demonstrated by revisions in the writing produced by the EAL medical researchers. This paper aims to contribute to current discussion on English for Research Publication Purposes (ERPP) for the EAL community who now form the main contributors to research knowledge dissemination.

Keywords: EAL writers, biomedical discourse, English for research publication purposes, lexical database, pedagogical benefits.

Resumen

Los beneficios pedagógicos de una base de datos de léxico (ScieE-Lex) como apoyo a la producción de textos biomédicos publicables por escritores que utilizan el inglés como lengua adicional

La investigación ha demostrado que el uso del inglés como lengua adicional (English as an Additional Language, EAL por sus siglas en inglés) en la escritura académica representa un reto para los investigadores no nativos de dicha lengua, en tanto que estos deben adquirir competencia fraseológica en inglés académico y desarrollar un conocimiento del lenguaje formulaico propio de la disciplina. Este artículo busca explorar si la base de datos de artículos de investigación biomédicos *SciE-Lex* puede utilizarse por estos escritores para mejorar su dominio del lenguaje formulaico en la escritura biomédica en lengua inglesa. En este trabajo se describen los retos asociados al lenguaje formulaico (las llamadas combinaciones de palabras) a los que se enfrentan los escritores de EAL y valora los beneficios pedagógicos de la utilización de esta base de datos léxica como apoyo a la producción de textos publicables. En concreto, se describen dos talleres de escritura diseñados para escritores de EAL (investigadores del ámbito de la medicina en el presente estudio). En los talleres estos investigadores trabajaron sobre varios borradores de textos utilizando *SciE-Lex*. Se describen los beneficios de su uso a través de las revisiones de los textos que llevaron a cabo los investigadores. El presente trabajo busca contribuir al debate actual sobre el llamado English for Research Publication Purposes (ERPP, por sus siglas en inglés) en la comunidad de escritores de EAL, comunidad que juega un papel primordial en la difusión del conocimiento científico.

Palabras clave: uso del inglés como lengua adicional en la escritura académica, discurso biomédico, Inglés para fines de investigación, base de datos léxica, beneficios pedagógicos.

1. Introduction

English for research publication purposes (ERPP) is now a well-established field of research in EAP. It is defined as "a branch of EAP addressing concerns of professional researchers and post-graduate students who need to publish in peer-reviewed international journals" (Cargill & Burgess, 2008: 75). English is the dominant language for research publication and there is strong evidence to suggest that the largest contributors to research publications are writers who use English as an additional language (EAL) (Hyland, 2016: 64). This paper recognises the importance of ERPP for a group of EAL Spanish medical researchers and reports on a study which uses a corpus-based lexical database in two workshops to help the users produce academic language typical of publications in their various fields of research. The paper also highlights the need for more concrete evidence from empirical studies of the impact of corpus-informed pedagogy.

There is wide acknowledgement of the usefulness of corpora for language teaching, for example, through the use of corpus-informed teaching materials such as the COBUILD project (Sinclair, 1987) and the contributions made to pedagogy by work such as the *Longman Grammar of Spoken and Written English* (Biber et al., 1999). Since 2000, there have been a number of influential textbooks in the field of language teaching and learning using corpora (Bennet, 2010; Flowerdew, 2012; to name a few). More recently, there have also been lexicographic developments such as the *Louvain English for Academic Purposes Dictionary* (LEAD) which incorporates a corpus tool with a specialised dictionary of general academic English (Paquot, 2012; Granger & Paquot, 2015). These contributions emphasise the relevance of corpus-informed pedagogy.

There is, however, growing concern that there is insufficient focused research matching the "'hype' given to corpora and/or corpus tools for pedagogical purposes" (Reppen, 2011 cited in Friginal, 2013: 210). Efforts to address this issue exist (Friginal, 2013), but "the evidence for the successful use of corpus resources… remains slight" (Tribble, 2013: 1). This paper contributes to the discussion of ERPP by investigating the use of a lexical database with a group of Spanish medical researchers to assist their production of discourse in their disciplinary area, viz. biomedical science. This study moves current research on corpus-informed pedagogy a step beyond awareness-raising, which is typically the focus of classroom-based research using corpora, to investigating actual language production, in this case, the written drafts of sections of biomedical research articles.

1.1. The role of corpora in the teaching of formulaic language

Large-scale general English corpora (such as the Bank of English), general academic English corpora (such as MICASE) to more specific genre corpora (such as BAWE focusing on the academic essay) and discipline-specific corpora (such as the one used in this study, the Health Science Corpus - HSC) have inspired research studies on real language use. Many of these corpora have been used in classroom situations inspired by Tim Johns' seminal data-driven learning (DDL) approach (1990). DDL is an approach in which learners become "language detectives" by discovering facts about the language they are learning for themselves and drawing conclusions from their exposure to authentic examples.

One of the key contributions of corpora to language teaching and learning has been the recognition of language as being formulaic in nature (Wray, 1999; Gledhill, 2000; Wray, 2002; Flowerdew, 2003; Simpson, 2004; Hyland, 2008, to name a few). This was brought to the fore by the neo-Firthian's pioneering work of Sinclair and Halliday. Following Sinclair's idiom principle, which states that writers can use "a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair, 1991: 110), it must be noted that evidence for formulaic expressions is highly significant in language.

Meunier (2012: 112) argues that "if teaching is meant to help learners improve their proficiency levels, it should then - at least in part - be devoted to improving learners' knowledge and use of formulas". However, it has been recently noted that "research into the teaching and learning of multi-word units is still scarce" (Pellicer-Sánchez, 2015: 1). In fact, Meunier (2012: 116) goes so far as to say that "very few studies provide results of experiments carried out to foster formulaicity *within* a pedagogical task".

To this end, the pedagogical use of databases as collections of information, specifically designed to facilitate language learning, seems very pertinent. Most examples of lexical databases in electronic form are focused on general English, such as, WordNet (Miller et al., 1990), which organises lexical information in terms of word meanings; EuroWordNet (Vossen, 2004), which provides a semantic analysis of semantic relations between synsets; SIMuLLDA (Janssen, 2004), a multilingual lexical database which uses structured interlingua; and Frame-based multilingual databases, which provide a semantic account of lexical units based on semantic frames (for example, Boas, 2005).

Despite the growing development of lexical databases that provide lexicogrammatical and discourse features of languages, more lexical resources are required to suit the needs of specialised discourse communities. As pointed out by Kennedy (2014), lexical databases must provide not only semantic information about the various sense(s) of each lexical unit, but also on how each sense may be realised by a different grammatical patterning, which contributes a great deal to characterising the prototypical environment of occurrence of formulaic expressions in a given discourse.

1.2. The formulaic nature of scientific English and its challenges for EAL writers

Research has demonstrated that it is particularly challenging for EAL writers to acquire phraseological competence in academic English and develop a good working knowledge of formulaic language (Howarth, 1996, 1998; Wray, 1999; Oakey, 2002; Williams, 2005; Granger & Meunier, 2008; Ferguson et al., 2011; Pérez-Llantada, 2014). This fact becomes especially apparent in scientific research articles which must show that the hypotheses have been tested appropriately and that the results reported accurately reflect the materials and methods used (Cargill & O'Connor, 2013).

The skills required for successful scientific writing entail both the accurate selection of correct terms and grammatical constructions as well as a good command of appropriate lexical combinations and phraseological expressions. Phraseological empirical studies have confirmed the important role of formulaic language in the textual development of meaning (Gledhill, 2000; Kaszubski, 2000; Flowerdew, 2003; Hyland, 2008) and have also highlighted the need for further research on the phraseological conventions characteristic of specialist genres. As Kaszubski (2000: 2) points out:

> Word combinations are inextricably related to the layer of style - the appropriateness and/or naturalness of selection and co-occurrence of items, subject to genre-sensitive restrictions and conventions. Thus, in order to compare aspects of lexical use, one is bound to focus attention on phraseology.

The current treatment of phraseology in specialised registers acknowledges the need for corpus-based studies of the prototypical lexicogrammatical patternings and discourse functions of formulaic language across disciplines (Oakey, 2002; Biber, 2006; Hyland, 2008; Laso, 2009; Laso & John, 2013a/b; Verdaguer et al., 2013). As asserted by Hyland (2008: 5), "[g]aining control of a new language or register requires a sensitivity to expert users' preferences for certain sequences of words over others". Thus, it seems that being familiar with the specific phraseology of a discourse community will bring about not only a better knowledge of the genre but also an enhanced competence in the process of reading and writing in specialised registers.

Due to the fact that discipline-specific phrases make up a very important part of the writing, it seems of paramount importance that professionals involved with the practice of research article writing become acquainted with

the formulaic language of their research field, since conforming to those conventions considered to be "good style" will maximise their chances of publishing in international scientific journals.

Bearing in mind that scientific discourse is "highly stereotypical in nature" (Gledhill, 2000: 116), it therefore presents a challenge for EAL writers. Spanish biomedical researchers (our targeted community in this study) must be aware of what Etherington (2008) calls the "game strategies": that is, the formulaicity that characterises scientific writing. Without some understanding and, most importantly, control over the rules of the game that operate across text types, structure, organisation and lexicogrammatical features, EAL writers will find it difficult to successfully publish in international journals in their subject areas (Pérez-Llantada, 2014).

As discussed in the literature (Cohen et al., 1988; Laso & John, 2013a/b), knowing the technical terms of a discipline is not a sufficient condition to write effective scientific papers in an efficient way. It is, in fact, the non-technical words – "terms that have a specialized meaning in a particular field and are used consistently in that field" (Cohen et al., 1988: 162) – which are more problematic to those EAL writers. In this regard, the use of lexical databases that give account of the formulaic language associated with non-technical terms in a given discipline seems a useful writing resource to assist the efficient production of published biomedical discourse.

1.3. Overview of study

With the aim of creating a lexical database to meet the growing demand for pedagogical resources assisting EAL teaching and learning, the GReLiC[1] research group at the University of Barcelona developed *SciE-Lex*, a lexical resource organised around highly prototypical non-specialised terms in biomedical discourse. *SciE-Lex* provides an exhaustive account of the combinatorial possibilities of general lexical units as well as their rhetorical features.

This paper explores if *SciE-Lex* can be exploited by EAL writers to enhance their knowledge of formulaic language, in particular the use of collocations, in biomedical English published writing. In addition, this study highlights the challenges associated with formulaic language for EAL writers, reflects on the benefits of a lexical database and evaluates a pedagogical approach to helping EAL writers produce publishable texts.

In order to provide Spanish biomedical researchers with the necessary skills to produce an academic research article using appropriate academic English and style, two writing workshops were conducted for a group of these biomedical researchers at the University of Barcelona. Workshop 1 aimed at helping our participants recognise the formulaic nature of biomedical discourse and to familiarise them with *SciE-Lex* through a series of exercises which could help them navigate through the database. Workshop 2 intended to provide support for these writers in their production of a publishable research article through consulting *SciE-Lex*.

2. Data and method

2.1. Corpus and the lexical database used in the study

This study is based on corpus evidence, since all formulaic language discussed has been extracted from the Health Science Corpus (HSC), which consists of a 4-million word collection of health science texts from the fields of medicine, biomedicine, biology and biochemistry.

SciE-Lex, which is based on the HSC, provides lexicogrammatical information about the most common collocations of general terms frequently used in the biomedical register as well as information on lexical bundles[2] associated with some of its headwords. This information relates not only to the lexicogrammatical variants of the lexical bundles, but also to the rhetorical functions (moves) performed by these units as well as their most prototypical distribution across the article. *SciE-Lex* can be found at www.ub.edu/grelic/eng/index.php.

2.2. Method

Emails were sent to three leading research institutions for participants to attend two "Writing for Publication" workshops in Barcelona: CRESA-Centre de Recerca en Salut Animal (UAB-IRTA), a public foundation created in 1999 for conducting research in animal health; the Institute for Research in Biomedicine (IRB-UB), a world-class research centre devoted to understanding fundamental questions about human health and disease; and the Institute for Bioengineering of Catalonia (IBEC-UB), a research centre whose purpose is to carry out interdisciplinary research at the highest international quality level which helps to improve health and quality of life and generate wealth.

While we targeted both doctoral and postdoctoral researchers, all our participants were doctoral researchers who were aiming to publish their current research in top international journals in their fields. When the participants registered for the workshops, they were asked to submit an 800-word draft of their writing. Fifteen participants (of at least C1 proficiency using the CEFR system) responded to our email and the final number attending was ten biomedical doctoral researchers mainly from the fields of Life Sciences and Psychological Sciences.

The submitted drafts were carefully read through and some non-prototypical collocations from three word classes (nouns, verbs and adjectives), i.e. collocates not found in the HSC corpus (see Section 2.1), were highlighted. These collocations then formed the basis of the activities developed during Workshop 1.

2.2.1. Workshop 1

The first part of the workshop opened with a discussion on the nature of scientific discourse and the unique characteristics of a journal article (how it is different from other types of research writing, such as thesis writing, which, as doctoral researchers, the participants were familiar with). The participants then completed Worksheet 1 (Appendix 1), which had two aims. Firstly, we hoped to familiarise them with the notion of prototypicality in biomedical discourse, and secondly, we hoped to encourage them to view language as occurring in chunks rather than as individual elements.

The prototypical nature of biomedical English was introduced through exercises using three academic journal articles and asking them to notice similarities in the ways in which these articles were structured and how language was used in general terms. Encouraging them to view language as occurring in chunks was achieved by using exercises with concordance lines which required them to think about context; in this case, collocations before and after a keyword. By the time they reached the end of this workshop, they were also familiar with the interface of *SciE-Lex*.

2.2.2. Workshop 2

The second workshop introduced them to the potential this lexical database had to assist their written production for publication.

The second Worksheet (Appendix 2) was then introduced. This worksheet was designed based on the drafts submitted as the pre-work for participation in the workshop. We had read and identified collocations in their drafts which were not prototypical of biomedical English as demonstrated in the HSC. We created a few collocation activities for them to complete and asked them to extend the observations they made about these collocations to their drafts. We then moved around the room and provided each participant with individual feedback on their drafts. We helped them with their searches in *SciE-Lex* and also made some general comments about their drafts. Then, they were asked to redraft their work and to save their writing.

At the end of Workshop 2, we asked participants to complete a questionnaire about their impressions of *SciE-Lex* and their experience of using it (Appendix 3). Our intention was to be able to correlate the questionnaire findings to revisions in their writing. In other words, we sought to find out if participants felt that *SciE-Lex* was a useful tool for them to improve their writing, then whether this would be demonstrated in the revisions implemented into their writing.

The observations we make in this paper come from 8 participants as 2 of them did not submit a second draft of their writing and thus were not considered in the present study. All participants signed a consent form and were assured of anonymity. The observations are based on a very small set of data, but the contexts of the workshops and the discussions we had with writers as we moved around the room is revealing of the potential for a corpus-based lexical database to be used as a pedagogical writing resource.

3. Observations from the workshops

The observations in this section will be presented in the same order as in the worksheets – noun, adjective and verb collocations. Each example will appear with the participants' first draft, revised draft and a screenshot of what motivated the revisions, not necessarily in this order. A discussion will follow in section 4 after the observations. It is important to note that we are not presenting all the occurrences of each of the nouns, verbs and adjectives we identified as appearing in lexical bundles presented in this study, but our primary focus here is to illustrate and highlight the influence of the use of *SciE-Lex* on the improvement of writing quality.

As already mentioned, we devised a worksheet of exercises consisting of verbs, nouns and adjectives (Table 1) featured in the participants' drafts. These nouns, adjectives and verbs were used by the participants in ways which were not typically found in the HSC and therefore deemed not prototypical of biomedical discourse. All these nouns, adjectives and verbs were highlighted in the participants' drafts as items for them to consider for revision during the workshops. Our aim was to see if the lexical database would enable them to make independent revisions to their drafts both in terms of lexical bundles and text distribution in their writing:

Nouns	Adjectives	Verbs
advance, procedure, resistance, growth, study, finding, purpose, result, research	capable, responsible, related	appear, assess, consist, develop, seem

Table 1. Nouns, adjectives and verbs used in the worksheets.

SciE-Lex provides information on lexical bundles associated with some of the headwords. As mentioned earlier, this information relates not only to the lexicogrammatical variants of the lexical bundles, but also to the rhetorical functions (moves) performed by these units as well as their most prototypical distribution across the article. In other cases, *SciE-Lex* only presents the lexicogrammatical information of the headwords.

3.1. Observation 1: The noun *study*

The abstract noun *study* occurs 6,618 times in the HSC, out of which 3,028 tokens are instances of the inflected form and the remaining 3,590 are base forms. A closer look at corpus data reveals that formulaic expressions of the type *in* + *the/this* + adjective + *study* stand out as recurrent chunks, as illustrated in the figures provided by *AntConc 3.4.4w* (Anthony, 2014) and shown in Figure 1.

Regarding the variability of the formulaic expression *in* + *the/this* + adjective + *study*, the following lexicogrammatical variants were most frequently found: *in this study* (647 occurrences), *in the* (adjective) *study* (301 occurrences), *in * study* (890 occurrences), and *used in this study* (167 occurrences).

Data from participants in the workshops shows some variability in the use of the lexical bundle *in the present study*. Example 1 illustrates, for instance, participant P1B's use of this lexical bundle before revision:

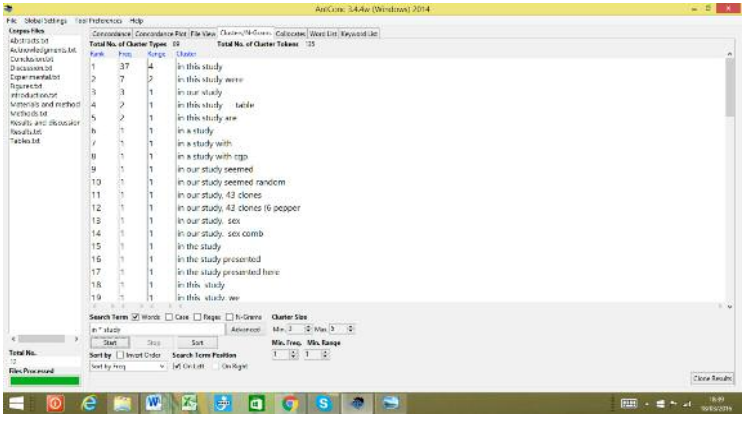


Figure 1. Information on lexical bundles of the noun *study* from *AntConc 3.4.4w*.

(1) **For the present study**, two isolates of Influenza A virus were used: an avian-origin LPAIV H5N2 subtype (A/*Anas platyrhynchos*/2420/2010) (H5N2) and a human-origin H1N1 subtype (A/Catalonia/63/2009) (pH1N1). (P1B)

There are no instances of the lexical bundle *for the present study* in the HSC. During the workshop, this participant was asked to search for the headword *study* (Figure 2):

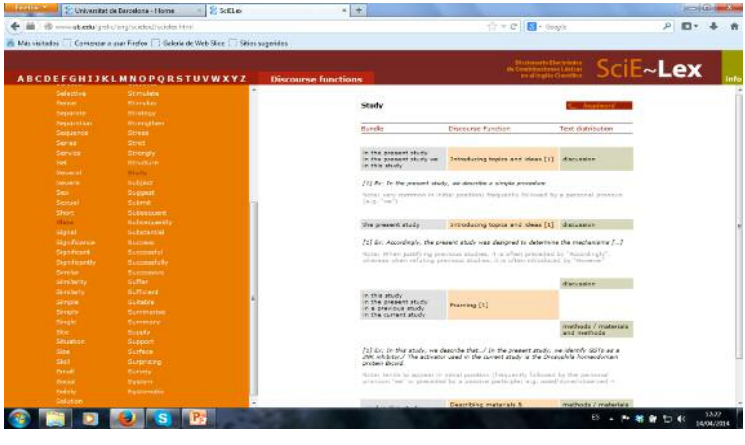


Figure 2. Screenshot of the lexical bundles associated with the noun *study* in *SciE-Lex*.

This motivated the following independent revision by participant P1B (Example 2).

(2) **In the present study**, two isolates of Influenza A virus were used: an avian-origin LPAIV H5N2 subtype (A/*Anas platyrhynchos*/2420/2010) (H5N2) and a human-origin H1N1 subtype (A/Catalonia/63/2009) (pH1N1). (P1B)

There was evidence of other uses of the noun *study* in lexical bundles used by the writers in the workshop. For example:

(3) Eighty-seven ml of OF collected from experimental PRRSV-negative piglets were pooled and **used for the study**. (P5E)

This lexical bundle *used for the study* occurs only once in the HSC, but the bundle *used in this study* occurs 167 times in the corpus data. While this participant did not revise her writing, the tendency to produce bundles which are not prototypical in the corpus should be noted. This was not an isolated example as there were similar examples such as the following one:

(4) **For this study**, only the GMV maps were used for statistical analyses. (P6R)

This lends additional evidence to observations already made by other researchers (Pérez-Llantada, 2014) about the challenges formulaic language poses for EAL writers. The implications of a workshop such as this one provides some indication of the benefits of EAL writers being able to consult corpora to aid their writing of formulaic language typical of the discourse communities they are writing for.

3.2. Observation 2: The noun *attention*

The noun *attention* has 151 occurrences in the HSC. The prepositions it collocates with are dependent on the verb preceding the noun. In the HSC, the two most common verbs which collocate with *attention* are *pay* (12 occurrences) and *focus* (35 occurrences). There is a wide range of other verbs which occur with *attention*, but with fewer occurrences for each verb: *receive* (8), *attract* (7), *bring* (4), *deserve* (4), *require* (3), etc. When the verb *pay* is used with *attention* the preposition it collocates with is *to*, whereas when the verb

focus is used with *attention,* the preposition it collocates with is *on*. The following was noted as occurring in one of our participant's writing:

(5) We **paid attention on** two cell based binding affinity assays: "MHC reconstitution assay" and "MHC-epitope stabilization assay". (P2M)

During Workshop 2, participant P2M consulted *SciE-Lex* with the following information about the noun *attention* (Figure 3):

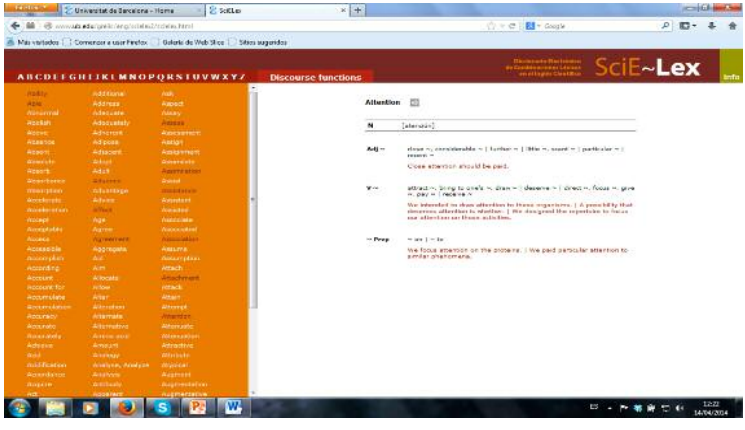


Figure 3. Screenshot of the lexicogrammatical patterning of the noun *attention* in *SciE-Lex*.

As a result of consulting *SciE-Lex*, the participant revised his first draft and changed the preposition used with the noun *attention* to produce the prototypical bundle *paid attention to* as found in the HSC (Example 6).

(6) We **paid attention to** two cell based binding affinity assays "MHC reconstitution assay" and "MHC-epitope stabilization assay". (P2M)

3.3. Observation 3: The adjective *responsible*

The adjective *responsible* occurs 526 times in the HSC. The preposition it collates with is always *for*.

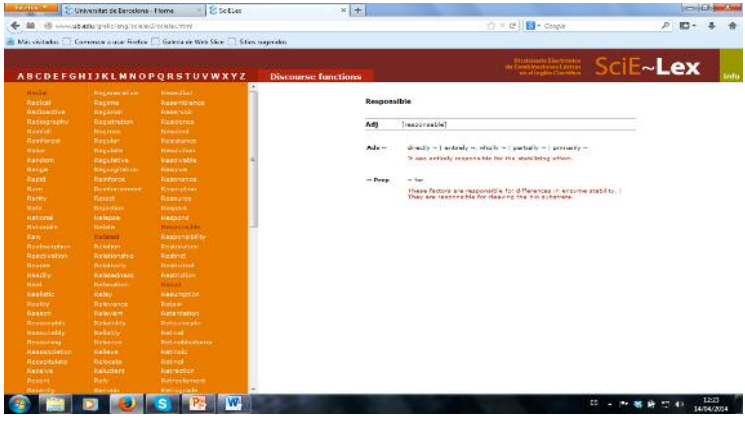


Figure 4. Screenshot of the lexicogrammatical patterning of the adjective *responsible* in *SciE-Lex*.

Example 7 demonstrates the use of this bundle in participant P3J's writing:

(7) Some of these outbreaks were **responsible of** avian-to-mammals transmissions, affecting also humans; thus, representing a threat to public health [2-4]. (P3J)

During Workshop 2, the revision to this bundle was motivated by consultation of *SciE-Lex* (Figure 4) and Example 8 is the revised version, thus reiterating the benefits of the use of the lexical database:

(8) Some of these outbreaks were **responsible for** avian-to-mammals transmissions, affecting also humans; thus, representing a threat to public health [2-4]. (P3J)

3.4. Observation 4: The adjective *capable*

The adjective *capable* occurs 343 times in the HSC. Of these 343 times, it occurs with the preposition *to* only once in the corpus, but 336 times in combination with the preposition *of*. Therefore the prototypical occurrence of this adjective is in combination with the preposition *of*. When participants used the adjective in their writing, we found that two of them (Example 9 and Example 10) used it with the preposition *to*.

(9) Previously, our group identified the peptide VIN1, located in conserved regions of the influenza A virus hemagglutinin subunit 1, as **capable to generate** cross-reactive antibodies (abs) in pigs. (P3J)

(10) The re-introduction of genes **capable to activate** cell death in tumoral cells or genes that can modulate intrinsic cellular factors and eliminate cancer cells are among the most common approaches. (P4L)

The information contained in *SciE-Lex* is illustrated in Figure 5:

Figure 5. Screenshot of the lexicogrammatical patterning of the adjective *capable* in *SciE-Lex*.

After consulting *SciE-Lex* (Figure 5), they revised their writing to produce the more prototypical form with the preposition *of* and concurrently revised the form of the verb to a gerund as demonstrated in Example 11 and Example 12 which follow:

(11) Previously, we identified the peptide VIN1, located in conserved regions of the influenza A virus hemagglutinin subunit 1, as **capable of generating** cross-reactive antibodies (abs) in pigs. (P3J)

(12) The re-introduction of genes **capable of activating** cell death in tumoral cells or genes that can modulate intrinsic cellular factors and eliminate cancer cells are among the most common approaches. (P4L)

3.5. Observation 5: The verb *consist*

The verb *consist* occurs 606 times in the HSC in all its forms: *consist* (73), *consists* (152), *consisted* (199), and *consisting* (182). When it occurs in any of its

forms in a bundle, it is followed by the preposition *of,* as demonstrated in *SciE-Lex* (Figure 6).

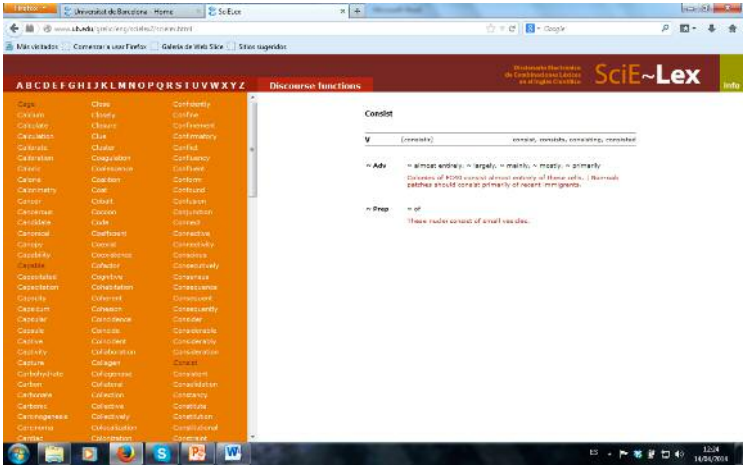


Figure 6. Screenshot of the lexicogrammatical patterning of the verb *consist* in *SciE-Lex*.

In his draft, participant P2M used *consist* in the following way (Example 13):

(13) The last step **consists to complete** the staining and test sample on a Flow cytometer. (P2M)

During Workshop 2, this participant was able to consult *SciE-Lex* (Figure 6) and revise his writing accordingly:

(14) The last step **consists of completing** the staining and test sample on a Flow cytometer. (P2M)

The observations above provide some evidence of the ways in which the medical researchers who participated in both workshops engaged with *SciE-Lex* and revised their writing based on the information they obtained from their searches of these lexical bundles in a lexical database. Our study demonstrates the potential for lexical resources of this kind to have an impact on writing quality.

4. Discussion

Beyond the revisions that the writers made to their drafts after their consultation of *SciE-Lex*, other interesting observations are worth mentioning here. During Workshop 2, we worked individually with participants. As each participant recognised and revised the headwords we had highlighted in their drafts, we noticed that they were also focusing their attention on where these bundles were occurring and concurrently revising other parts of their paper which were not presented to us as part of the pre-work for the workshops. This particular aspect of their revision behaviour was interesting as other studies have noted the benefits of highlighting (to students) language which occurs in particular moves. Bianchi and Pazzaglia (2007), for instance, asked their students to subdivide a research article and to then examine concordances of research-associated keywords (like *study/studies*, *experiment/experiments*, *research/researches,* etc.) in the different sections of the article to sensitise them to different uses of these keywords depending on the moves they occurred in.

Other studies, too, have highlighted the benefits of familiarising students with lexical bundles occurring in particular moves. Bhatia, Langton and Lung (2004, cited in Connor & Upton, 2005) have stressed the need to make law students aware of their lexical choices in moves through a study of the synonymous words *dismiss* and *reject* which have very clear preferences for different moves in law cases. Flowerdew (2015), for instance, reports on an exercise she developed with students in which her aim was to familiarise them with lexical phrases associated with commenting on results. She used the keyword *surprising* and found that using this type of empirical data alerts students to the fact that language has certain phraseological tendencies, depending on the genre under investigation. Similarly, *SciE-Lex* already contains discourse information for a number of words and this can be readily accessed and used by the writers, as was observed in our study (Figure 2).

There are numerous studies which confirm the benefits of awareness-raising activities which have "led to a considerable improvement of the recognition of formulaic language" (Meunier, 2012: 120). Of note are also studies influenced by pedagogy and some key examples are those by Charles (2007, 2011) and Bloch (2008, 2009, 2010), who have both designed and developed innovative hands-on corpus activities for their students. Their papers report on the benefits of students searching specialised corpora for typical lexicogrammatical functions in classroom contexts. What is difficult to

notice from these studies, and which we have attempted to do in the current study, is to try and gather information on the "actual uptake of formulaic sequences", which Meunier has noted as "not always easy to assess" (2012: 120). By recognising the revisions our writers have made to their first drafts, we have attempted to provide some evidence of the influence of corpus consultation on the written production (not only awareness-raising) of the participants in our workshops. This, we feel, goes some way towards addressing Reppen's (2011) suggestion for "heightened classroom research looking at the effects of corpus-informed materials on writing quality", although with several limitations as discussed next.

5. Limitations

One limitation we observed was that the participants in our workshops needed us to identify non-prototypical collocations for them (Worksheet 2). To this respect, and bearing in mind that participants were unfamiliar with the database, the role of the facilitators during the workshops was extremely important in making participants' lexicogrammatical searches in *SciE-Lex* more successful. Our expectation, however, is that as they get acquainted with the use of this pedagogical resource, they are likely to be able to use it more independently so as to improve their writing. One possible suggestion may be to devise some activities which use the most frequent research-oriented headwords (e.g. *study*, *experiment*, *research*, *results*, *limitations*, *discussion*) to sit alongside the database to sensitise them to the way in which *SciE-Lex* works and how it might assist them with their written production. Another possible suggestion could be to improve *SciE-Lex* with a tool to automatically highlight EAP words which are challenging for learners.

Another limitation was the small number of participants. There certainly was the possibility that more participants might have attended the workshops, had we not asked for a piece of writing to be submitted ahead of Workshop 1. In fact, two participants attended the workshops without providing us with writing beforehand. There is also a general reluctance to submit redrafted work. Research involving revision data tends to be small due to the challenge around collecting revised writing (Laso, 2009).

Despite the fact that the observations from this study cannot be generalised because of the small size of our dataset, they can serve to raise awareness about the growing need for corpus-informed materials across disciplines that

contribute to EAL scholarly writers getting familiar with the formulaic language of published research articles in their field of expertise (Friginal, 2013: 216).

6. Conclusion

The first edition of the workshops on "Writing for Publication" proved to be successful on various fronts since it introduced medical researchers to the language and style characteristic of biomedical English by means of *SciE-Lex*. It also showed them how to use a lexical database to eventually be able to consult it independently.

Overall, the experience has been very positive. The use of *SciE-Lex* has contributed to participants improving their drafts from a lexicogrammatical point of view, for example, collocational patterns of non-technical terms in biomedical research articles. Users also considered other factors beyond the actual lexicogrammar. Also, the facilitators' interventions during the workshops helped participants improve their drafts on issues such as paragraph distribution, thesis development, organisation of topic sentences, and punctuation.

One outcome which we did not predict was the way in which participants reacted to *SciE-Lex*. They were very engaged during the workshops and they were comfortable with the terminology we had introduced - collocations and lexical bundles. We feel this is a good way forward for writers to improve their production of publishable articles, as they become familiar with how to recognise and produce effective research articles. Finally, a satisfaction questionnaire was distributed among participants (Appendix 3), all of whom pointed out that they found *SciE-Lex* an extremely useful resource to help them produce phraseologically competent texts in biomedical English.

These workshops have also stressed the fact that further corpus-informed studies on the pedagogical applications of lexical resources are needed so as to contribute to a thorough understanding of the challenges faced by EAL writers of specialised discourses. We opened this article with a quote from Reppen (2011) about the need for more focused research to match the "hype" given to the use of corpora for teaching purposes. This study, while limited in size and scope, provides some evidence towards this. What it has certainly achieved is evidence of the potential for a database to help this particular group of writers produce phraseologically competent texts,

contributing towards the need for more evidence of "the influence of corpora in developing writing skills" (Friginal, 2013: 220).

Acknowledgements

The authors acknowledge the support of the Grup de Recerca en Lexicologia i Lingüística de Corpus group (GReLiC) at the University of Barcelona (2014SGR1374). We would also like to acknowledge all the participants in the workshops who willingly submitted first and second drafts of their work for this research study. We would also like to thank the reviewers for the helpful comments and suggestions on this paper.

Article history:
Received 21 March 2016
Received in revised form 11 August 2016
Accepted13 August 2016

References

Anthony, L. (2014). *AntConc* (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. URL: http://www.laurenceanthony.net/ [21/03/2016].

Bennet, G. (2010). *Using Corpora in the Language Learning Classroom*. Ann Arbor: Michigan University Press.

Bhatia V.K., N. Langton & J. Lung (2005). "Legal discourse: Opportunities and threats for corpus linguistics" in U. Connor & T. Upton (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*, 203-231. Amsterdam: John Benjamins.

Bianchi, F. & R. Pazzaglia (2007). "Student writing of research articles in a foreign language: Metacognition and corpora" in R. Facchinetti (ed.), *Corpus Linguistics 25 Years On*, 259-287. Amsterdam: Rodopi.

Biber D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Bloch, J. (2008). *Technologies in the Second Language Composition Class*. Ann Arbor: University of Michigan Press.

Bloch, J. (2009). "The design of an online concordancing program for teaching about reporting verbs". *Language Learning and Technology* 13,1: 59-78.

Bloch, J. (2010). "A concordance-based study of the use of reporting verbs as rhetorical devices in academic papers". *Journal of Writing Research* 2,2: 219-244.

Boas, M. C. (2005). "Semantic frames as interlingual representation for multilingual databases". *International Journal of Lexicography* 18,4: 445-478.

Cargill, M & S. Burgess (2008). "Introduction to the Special Issue: English for research publication purposes". *Journal of English for Academic Purposes* 7,2: 75-76.

Cargill, M. & P. O'Connor (2013). *Writing Scientific Research Articles*. Oxford: Wiley-Blackwell.

Charles, M. (2007). "Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions". *Journal of English for Academic Purposes* 6,4: 289-302.

Charles, M. (2011). "Using hand-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*, 26-43. London: Continuum.

Cohen, A.D., H. Glasman, P. R. Rosenbaum-Cohen, J. Ferrera & J. Fine (1988). "Reading English for specialised purposes: Discourse analysis and the use of student informants" in P.

Carrell, J. Devne & D.E. Eskey (eds.), *Interactive Approaches to Second Language Reading*, 152-167. Cambridge: Cambridge University Press.

Etherington, S. (2008). "Academic writing and the disciplines" in P. Friedrich (ed.), *Teaching Academic Writing*, 26-58. London: Continuum.

Ferguson, G., C. Pérez-Llantada & R. Plo (2011). "English as an international language of scientific publication: A study of attitudes". *World Englishes* 30: 41-59.

Flowerdew, J. (2003). "Signalling nouns in discourse". *English for Specific Purposes* 22: 329-346.

Flowerdew, J. (2012). *Corpora and Language Education*. London: Palgrave Macmillan.

Flowerdew, L. (2015). "Corpus-based research and pedagogy in EAP: From lexis to genre". *Language Teaching* 48: 99-116.

Friginal, E. (2013). "Developing research report writing skills using corpora". *English for Specific Purposes* 32: 208-220.

Gledhill, C. (2000). *Collocations in Science Writing*. Tübingen: Gunter Narr.

Granger, S. & F. Meunier (eds.) (2008). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins.

Granger, S. & M. Paquot. (2015). "Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid". *Lexicographica: international annual for lexicography* 31,1: 118-141.

Howarth, P.A. (1996). *Phraseology in English Academic Writing: Some Implications for Language Learning and Dictionary Making*. Tübingen: Max Niemeyer Verlag.

Howarth, P.A. (1998). "Phraseology and second language proficiency". *Applied Linguistics* 19,1: 24-44.

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

Hyland, K. (2016). "Academic publishing and the myth of linguistic injustice". *Journal of Second Language Writing* 31: 58-69.

Janssen, M. (2004). "Multilingual databases, lexical gaps, and SIMuLLDA". *International Journal of Lexicography* 17,2: 137-154.

Johns, T. (1990). "From printout to handout: Grammar and vocabulary teaching in the context of data driven learning". *CALL Austria* 10: 14-34.

Kaszubski, P. (2000). *Selected Aspects of Lexicon, Phraseology and Style in the Writing of Polish Advanced Learners of English: A Contrastive, Corpus-Based Approach*. URL: http://www.staff.amu.edu.pl/~przemka/rsearch.html#PhD [19/03/2016].

Kennedy, G. (2014). *An introduction to corpus linguistics*. London: Routledge.

Laso, N.J. (2009). *A Corpus-Based Study of the Phraseological Behaviour of Abstract Nouns in Medical English: A Needs Analysis of a Spanish Medical Community. PhD dissertation,* University of Barcelona. URL: http://www.tesisenred.net/bitstream/handle/10803/1671/NJLM_THESIS.pdf?sequence=1

Laso, N.J. & S. John (2013). "A corpus-based analysis of the collocational patterning of adjectives with abstract nouns in medical English" in I. Verdaguer, N.J. Laso & D. Salazar. (eds.), *Biomedical English: A Corpus-based Approach*, 55-71. Amsterdam/Philadelphia: John Benjamins Publishing.

Laso, N.J. & S. John (2013). "An exploratory study of NNS medical writers' awareness of the collocational patterning of abstract nouns in medical discourse". *Revista Española de Lingüística Aplicada (RESLA)* 26: 307-331.

Meunier, F. (2012). "Formulaic Language and Language Teaching". *Annual Review of Applied Linguistics* 32: 111-129.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross & K.J. Miller (1990). "Introductions to WordNet: An on-line database". *International Journal of Lexicography* 3,4: 235-244.

Oakey, D. (2002). "Lexical phrases for teaching academic writing in English: Corpus evidence" in S. Nuccorini (ed.), *Phrases and Phraseology – Data and Descriptions*, 85-105. Bern: Peter Lang.

Paquot, M. (2012). "The LEAD dictionary-cum-writing aid: An integrated dictionary and corpus tool" in S. Granger & M. Paquot (eds.), *Electronic Lexicography*, 163-185. Oxford University Press: Oxford.

Pellicer-Sánchez, A. (2015). "Learning L2 collocations incidentally from reading". *Language Teaching Research*, 1-22.

Pérez-Llantada, C. (2014). "Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage". *Journal of English for Academic Purposes* 14: 84-94.

Reppen, R. (2011). "Using corpora for pedagogy: Does it make a difference?" Plenary paper presented at the *American Association for corpus*

linguistics conference 2011, Atlanta, GA.

Simpson, R.C. (2004). "Stylistic features of academic speech: The role of formulaic expressions" in U. Connor & T.A. Upton (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*, 37-64. Amsterdam: John Benjamins.

Sinclair, J.M. (ed.) (1987). *Looking Up: An Account of the Collins COBUILD Project*. London: Collins ELT.

Sinclair, J.M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Tribble, C. (2013). "Introduction. Corpora in the language-teaching classroom" in C.A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*.

Vossen, P. (2004). "EuroWordNet: A multilingual database of autonomous and language specific Wordnets connected via an inter-lingual index". *International Journal of Lexicography* 17,2: 161-173.

Williams, G. (2005). "Challenging the native-speaker norm: A corpus-driven analysis of scientific usage" in G. Barnbrook, P. Danielsson & M. Mahlberg (eds.), *Meaningful Texts. The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, 115-127. London: Continuum.

Wray, A. (1999). "Formulaic language in learners and native speakers". *Language Teaching* 32: 213-31.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Natalia J. Laso is a lecturer at the University of Barcelona (UB). She holds a PhD in English Philology from the UB and is also a member of the GRELIC-Lexicology and Corpus Linguistics Research Group. Her research is focused on science writing for research publication purposes as well as on the use of corpora in the linguistics classroom.

Suganthi John is a Senior Lecturer in English Language in the Department of English Language and Applied Linguistics at the University of Birmingham. Her research focuses on self-representation and identity in academic texts. She is also interested in writing development across boundaries (undergraduate to postgraduate; postgraduate to workplace) and in writing for research publication purposes.

NOTES

[1] Grup de Recerca en Lexicologia i Lingüística de Corpus (Lexicology and Corpus Linguistics Research group); (2014SGR1374).

[2] Recognising that there are a number of different definitions for multi-word units, in this paper we use "lexical bundles" as our preferred term because this is the terminology used in *SciE-Lex*, the pedagogical resource discussed in this paper.

Appendix 1

Writing for Publication Workshop 1

Worksheet 1

Source: adapted from "Academic writing and the disciplines" in Friedrich, P. (ed.) Teaching Academic Writing. London: Continuum.

1. Read through the HSC samples below and answer the following questions:

a) Do the titles contain a common pattern? How long is a typical title? Does it include any punctuation, such as a colon, semi-colon, or dash?

1. *Interactions between the Escherichia coli cAMP receptor protein and the Cterminal domain of the a subunit of RNA polymerase at Class I promoters*
2. *A New Concept in Artificial Diets for Chrysoperla rufilabris: The Efficacy of Solid Diets*
3. *Comparative genomics: the key to understanding the Human Genome Project*
4. *Development and Behavior of Spodoptera exigua (Lepidoptera: Noctuidae) Larvae in Choice Tests with Food Substrates Containing Toxins of Bacillus thuringiensis*
5. *Effects of Temperature on Eggs, Fecundity, and Adult Longevity of Hylobius transversovittatusGoeze (Coleoptera: Curculionidae), a Biological Control Agent of Purple Loosestrife*
6. *Hybrid Zones and the Genetic Architecture of a Barrier to Gene Flow Between Two Sunflower Species*
7. *Fecundity and Longevity of Green Vegetable Bug, Nezara viridula, Following Parasitism by Trichopoda giacomellii*
8. *Genetic Identification of Three ABC Transporters as Essential Elements for Nitrate Respiration in Haloferax volcanii*
9. *Habitat Preferences of Three Congeneric Braconid Parasitoids: Implications for Host-Range Testing in Biological Control*
10. *Developmentally programmed assembly of higher order telomerase complexes with distinct biochemical and structural properties*

b) Do the enclosed articles use any headings or sub-headings? If so, are they general (e.g., Intro, Method, Conclusion) or text-specific?

c) Do(es) the author(s) use first person pronouns (*I/my/me* or *we/our/us*) at all? If so, when, how often and why?

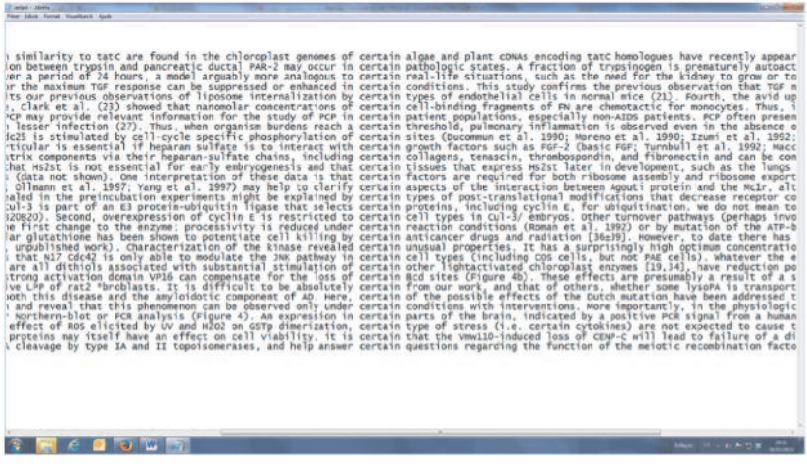
d) When referring to other work, do(es) the author(s) use any evaluative language, such as adjectives (e.g., *useful*, *successful*, *positive/negative*, *harmful*), adverbs (e.g., *effectively*, *satisfactorily*, *inadequately*, (*un*)*successfully*) or verbs with evaluative connotations (e.g., *succeed*, *fail*, *prove*). List these and indicate whether they are positive or negative.

Appendix 2

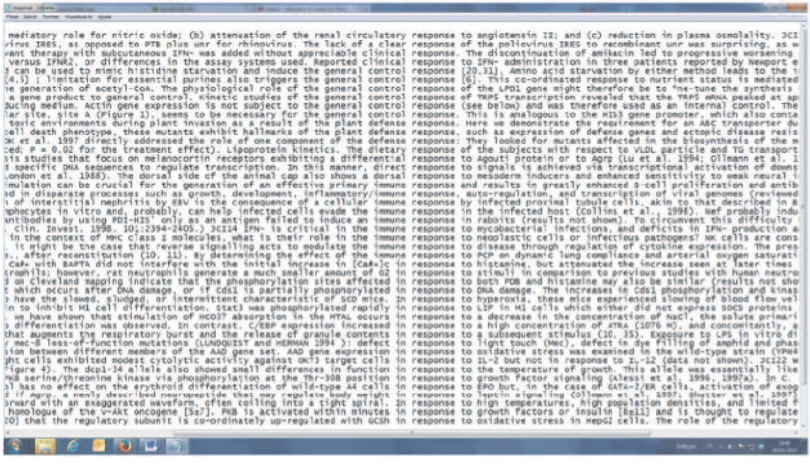
Writing for Publication Workshop 2

Worksheet 2

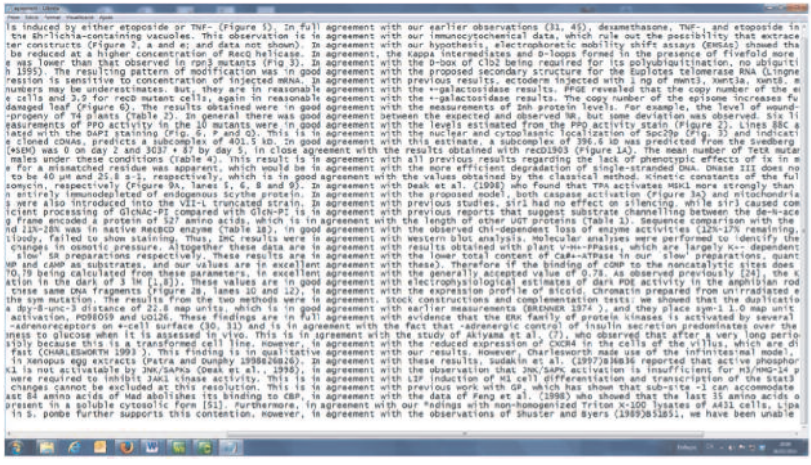
1. Study the concordance data below and find the instances of the word certain used (1) for referring to someone or something without being specific about exactly what or who they are and (2) with the sense "definitely true":



2. From the concordance data below, supply the adjectives and the prepositions that collocate with the word *response*.



3. From the concordance data, identify (1) the verbs that collocate with the word *agreement*, (2) the adjectives that precede it, and (3) the prepositions that follow it.



Appendix 3

*SciE-Lex*_Satisfaction Questionnaire

We would be extremely grateful if you could spend just a few minutes of your time completing this short questionnaire about your use of *SciE-Lex*.

Please tell us something about yourself:

I am a postgraduate student (PhD)
I am a researcher
I am a university lecturer
Other

Please select the best description of your field of study/work:

Engineering
Life Sciences
Medicine and Nursing
Physical Sciences
Psychological Sciences
Social Sciences

What is your first language?

Please indicate how useful you feel *SciE-Lex* is for scientific writing

	1	2	3	4	5	6	
Not useful at all	○	○	○	○	○	○	Extremely useful

If you have found *SciE-Lex* to be useful, please indicate how you feel it has helped you:

	Disagree	Not sure	Agree	Strongly agree
***SciE-Lex* gives me the language I need for my scientific writing**	○	○	○	○
***SciE-Lex* gives me a wider choice of language for my writing**	○	○	○	○
***SciE-Lex* helps me to see which phrases are generic and can be re-used**	○	○	○	○
***SciE-Lex* helps me to organise my writing**	○	○	○	○
***SciE-Lex* gives me ideas for my writing**	○	○	○	○
***SciE-Lex* has helped/helps me to feel more confident about my writing**	○	○	○	○
***SciE-Lex* has helped/helps me to improve my writing style**	○	○	○	○

Can you think of any other ways that *SciE-Lex* could help you?

Could you suggest a few ways in which you think *SciE-Lex* could be improved?

Many thanks for completing this questionnaire