TEISEL
Tecnologías para la investigación en segundas lenguas
ISSN 2696-676X
Universitat de Barcelona
Facultat de Filologia i Comunicació

# SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English

Natalia Judith LASO
Universitat de Barcelona (España)
njlaso@ub.edu
https://orcid.org/0000-0001-5553-6986

**Resumen:** This paper aims to describe the building-up of SciE-Lex (http://www.ub.edu/grelic/eng/scielex2/scielex.html), a collocational database of non-specialized terms in biomedical English, which was primarily conceived as a response to the lack of reference tools accounting for the lexicogrammatical patterning associated with non-technical terms frequently used in the health science discourse. SciE-Lex thus serves the purpose of assisting L2 English writers from the health science discourse community in their production of biomedical texts in English. This collocational database is the result of a lexicographic project carried out by the GreLiC research group at the University of Barcelona, and as such it has undergone various developmental stages since its inception. In order to evaluate its adequacy as a writing tool addressed to the Spanish biomedical community and confirm the appropriateness of the combinatorial patterning and phraseological information included in each entry, a group of language experts were asked to assess the dictionary by stressing both its weaknesses and strengths. Their feedback has stressed the suitability of SciE-Lex as a lexicographic resource and yielded significant improvement of the tool. Last but not least, SciE-Lex has also been successfully tested with targeted users in a series of "Writing for publication" workshops held at the University of Barcelona, and taught by the author, the results of which have corroborated the usefulness of this lexical database to enhance Spanish users' biomedical English published writing.

**Keywords**: databases; biomedical discourse; pattern grammar; writing resource; English for research publication purposes

*Español:*

***SciE-Lex: Creación de una base de datos de colocaciones para ayudar a la producción de textos biomédicos en inglés como L2***

***Resumen**: Este artículo tiene como objetivo describir la construcción de SciE-Lex (http://www.ub.edu/grelic/eng/scielex2/scielex.html), una base de datos combinatoria de términos no especializados en inglés biomédico, concebida principalmente para dar respuesta a la escasez de herramientas de referencia que den cuenta de los patrones léxico-gramaticales asociados a términos no especializados de uso frecuente en el discurso de las ciencias de la salud. Por lo tanto, SciE-Lex nació con el propósito de ayudar a científicos españoles, que utilizan el inglés como segunda lengua, a redactar textos biomédicos en inglés. Esta base de datos de colocaciones es el resultado de un proyecto lexicográfico llevado a cabo por el grupo de investigación GreLiC de la Universidad de Barcelona, y como tal ha experimentado diversas etapas de desarrollo. Con el fin de evaluar su idoneidad como herramienta de escritura dirigida a la comunidad biomédica española y confirmar así la pertinencia de incluir patrones combinatorios e información fraseológica en cada entrada, se le pidió a un grupo de lingüistas que evaluaran el diccionario atendiendo a las debilidades y fortalezas que identifican en la herramienta. Sus comentarios han subrayado la efectividad de SciE-Lex como recurso lexicográfico y han contribuido a mejorar la herramienta de forma significativa. Por último, SciE-Lex también ha sido utilizado en una serie de talleres de "Redacción de artículos científicos con fines de publicación" celebrados en la Universidad de Barcelona, e impartidos por la autora, cuyos resultados han corroborado la utilidad de esta base de datos léxica para mejorar la producción científica en inglés de la comunidad biomédica española.*

***Palabras clave**: bases de datos; discurso biomédico; gramática de patrones; recurso de escritura; inglés con fines de publicación*



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021 | Aceptación: 20/12/2021 | Publicación: 03/01/2021

ISSN 2696-676X


*Català:*

***SciE-Lex: Creació d'una base de dades de col·locacions per ajudar a la producció de textos biomèdics en anglès com a L2***

***Resum**: Aquest article té com a objectiu descriure la construcció de SciE-Lex (http://www.ub.edu/grelic/eng/scielex2/scielex.html), una base de dades combinatòria de termes no especialitzats en anglès biomèdic, concebuda principalment per donar resposta a l'escassedat d'eines de referència que donin compte dels patrons lèxic-gramaticals associats a termes no especialitzats d'ús freqüent en el discurs de les ciències de la salut. Per tant, SciE-Lex va néixer amb el propòsit d'ajudar científics espanyols, que utilitzen l'anglès com a segona llengua, a redactar textos biomèdics en anglès. Aquesta base de dades de col·locacions és el resultat d'un projecte lexicogràfic dut a terme pel grup de recerca GreLiC a la Universitat de Barcelona, i com a tal ha experimentat diverses etapes de desenvolupament. Per tal d'avaluar la idoneïtat de SciE-Lex com a eina d'escriptura adreçada a la comunitat biomèdica espanyola i confirmar així la pertinència d'incloure patrons combinatoris i informació fraseològica a cada entrada, se li va demanar a un grup de lingüistes que avaluessin el diccionari atenent a les debilitats i fortaleses que identificaven a l'eina. Llurs comentaris han subratllat l'efectivitat de SciE-Lex com a recurs lexicogràfic i han contribuït a millorar l'eina de manera significativa. Finalment, SciE-Lex també ha estat utilitzat en una sèrie de tallers de "Redacció d'articles científics amb fins de publicació" celebrats a la Universitat de Barcelona i impartits per l'autora, els resultats dels quals han corroborat la utilitat d'aquesta base de dades lèxica per millorar la producció científica en anglès de la comunitat biomèdica espanyola.*

***Paraules clau**: bases de dades; discurs biomèdic; gramàtica de patrons; recurs d'escriptura; anglès per a fins de publicació.*

## Introduction

This paper acknowledges the adequacy of corpus-based databases in second language teaching and learning by presenting SciE-Lex (http://www.ub.edu/grelic/eng/scielex2/scielex.html), a lexical database of the most frequently used collocations of English non-technical words in the health science discourse. This lexical resource has been developed by the Lexicology and Corpus Linguistics Research Group (GreLiC) from the University of Barcelona, specialised in corpus linguistics and lexicology. As a member of this research team and inspired by Pattern Grammar (Hunston & Francis, 2000), my research has focused on the analysis of health science discourse from a lexicogrammatical perspective for the last few years.

SciE-Lex is the result of a series of lexicogrammatical theoretical studies conducted in previous years. It also draws on a number of lexicographic studies (L'Homme, 2005 and 2008; Reimerink & Faber, 2009; Fuertes-Olivera, Niño & Sastre, 2019) which argue for the inclusion of linguistic information in terminological dictionaries. Having analysed the patterns associated with general terminology in the biomedical discourse, I came to realise that there was a shortage of reference tools offering information on the combinatorial patterns in which those terms are typically found (Laso & John, 2017).

In particular, it became noticeable that there was a need for further reference tools which provide L2 English speakers and English as an additional language (EAL) writers with the prototypical use of lexicogrammatical patterns of non-technical words, on the one hand, and with the conventionalised phraseological characteristics of their discourse community, on the other. This is key in specialised writing, where the writer has to adhere to conventional style norms and to appropriate collocations, so that the reader is not distracted by inappropriate expressions and can thus read fluently and focus on the content.

Given the apparent lack of such specialised open access lexical databases, SciE-Lex was conceived with the aim of bridging the aforementioned gap. This collocational database is therefore intended to supply not only information



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021| Aceptación: 20/12/2021| Publicación: 03/01/2021

ISSN 2696-676X


about the meanings and the grammatical and collocational patterns of general words prototypically found in biomedical journals, but also help Spanish scientists gain control of the phraseological conventions of the scientific discourse (http://www.ub.edu/grelic/eng/index.php).

As a tool that has gone through various stages of development, SciE-Lex has improved to a large extent in the last few years as it now contains lexical bundles, information about their textual distribution, and discourse function in the various sections of the biomedical research article. Additionally, following Fillmore's frame semantics' approach (1976), the GreLiC research group has been working lately on the association of the database entries with semantic frames, which is revealing remarkable differences between general and biomedical English (Verdaguer *et al.*, 2020).

## 1. HSC: Corpus Compilation and Corpus Processing

Interested as we were in the lexicogrammatical patterning of non-technical terms in biomedical English research article, and the conventionalised phraseology prototypical of such genre, the GreLiC group decided to compile a Health Science Corpus (HSC), which was funded by the Spanish Ministry of Science and Education and FEDER, as part of the project "Creation of a Database of Lexical Combinations in Scientific English" (BFF2001-2988), coordinated by Dr. Isabel Verdaguer at the University of Barcelona. SciE-Lex is based on the HSC, which currently comprises a four-million-word collection of research papers (a total of 718 published between 1998-1999) from prestige online journals that cover different disciplines such as medicine, biology, biochemistry, and biomedicine, as shown in Table 1:

**Table 1**
*Corpus size per discipline and journal*

| Discipline | Journal | Number of papers |
|---|---|---|
| **Biology** | Genes and Development | 40 articles |
| | Genetics | 54 articles |
| | Journal of Cell Biology | 26 articles |
| | The American Journal of Primatology | 20 articles |
| | Biological Control | 97 articles |
| | The Journal of Experimental Zoology | 32 articles |
| | BioEssays | 99 articles |
| | Integrative Biology | 18 articles |
| | Zoo Biology | 65 articles |
| **Biochemistry** | Biochemical Journal | 53 articles |
| | The Embo Journal | 64 articles |
| **Medicine** | Journal of Clinical Investigation | 53 articles |
| | British Medical Journal | 58 articles |
| | Journal of Bacteriology | 39 articles |
| **Total** | **718 articles** | |

Following Gries & Stefanowitsch's criteria for corpus compilation and processing (2006, p. 4), the HSC is intended to be a representative sample of the health science research article. Once the corpus was compiled, all articles were fully edited to smooth out problems concerning the typographical form of texts. Every downloaded text presented



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021| Aceptación: 20/12/2021| Publicación: 03/01/2021

ISSN 2696-676X


problems with capital letters, paragraphing, diagrams, numbers, photographs, columnar layouts, etc. Thus, they were edited manually, converted into plain text files, and eventually stored into different folders and subfolders, accounting for discipline, journal, author, and year of publication. All these steps were prior to the data processing stage.

After HSC corpus compilation and annotation, WordSmith Tools 3.0 (Scott, 1997) was used to retrieve a frequency list of lexical items in the corpus, of which I selected those with a frequency higher than five occurrences per million words. The resulting list was compared with the Academic Word List (Coxhead 2000) and the Academic Keyword List (Paquot, 2010). Lists of concordances, clusters and collocates were also obtained with WordSmith Tools 3.0, and were used to assist in the linguistic analysis of the selected lexical items. This software was particularly useful for the classification of word clusters so as to see the patterns of repeated phraseology in the concordance lines analysed and to make generalisations from the observation of repeated language events (Laso, 2009).

As already noted, one of the main aims when compiling the HSC was trying to make a representative selection of naturally occurring language in health science discourse, in order to analyse the collocational and syntagmatic structures associated with non-technical terms in biomedical discourse. Within this framework, it must be stressed that the compilation of the HSC understood as "an authoritative body of linguistic evidence which can support generalizations and against which hypotheses can be tested" (Sinclair, 1987, p. 2) has facilitated the exploration of grammatical patterning by means of corpus evidence. For many research and pedagogical purposes, the larger the corpus is, the more reliable conclusions can be drawn from the careful examination of the language shown. As pointed out by Hunston (2002), however, all observations made from a particular collection of texts "must be dealt with as deductions rather than as facts" (Hunston, 2002, p. 23), so the HSC was conceived as a sample, a cross-section of the health science discourse, and thus the information displayed in SciE-Lex is based on the evidence obtained from the in-depth analysis of the HSC data.

## 2. SciE-Lex Report

The information resulting from the analysis of the HSC corpus was codified in SciE-Lex according to the following parameters: word class; morphological variants; equivalent(s) in Spanish with clarification of the sense, if necessary; patterns of occurrence; collocates; and examples of real use.

While in its first developmental stage SciE-Lex incorporated contextual information about the use and combinatorial potential of general terms in health science discourse, in a second stage, and in line with the new corpus-based phraseological trends, SciE-Lex was supplemented with formulaic expressions, and provided explicit information about their composition, variability, discourse function and text distribution in the research article.

The inclusion of prefabricated expressions in SciE-Lex has contributed to the characterisation of the prototypical phraseology of health science discourse and to the structural organization of discourse, as shown in Figure 1.



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021| Aceptación: 20/12/2021| Publicación: 03/01/2021

ISSN 2696-676X


**Figure 1**

*Entry for "support" in SciE-Lex*

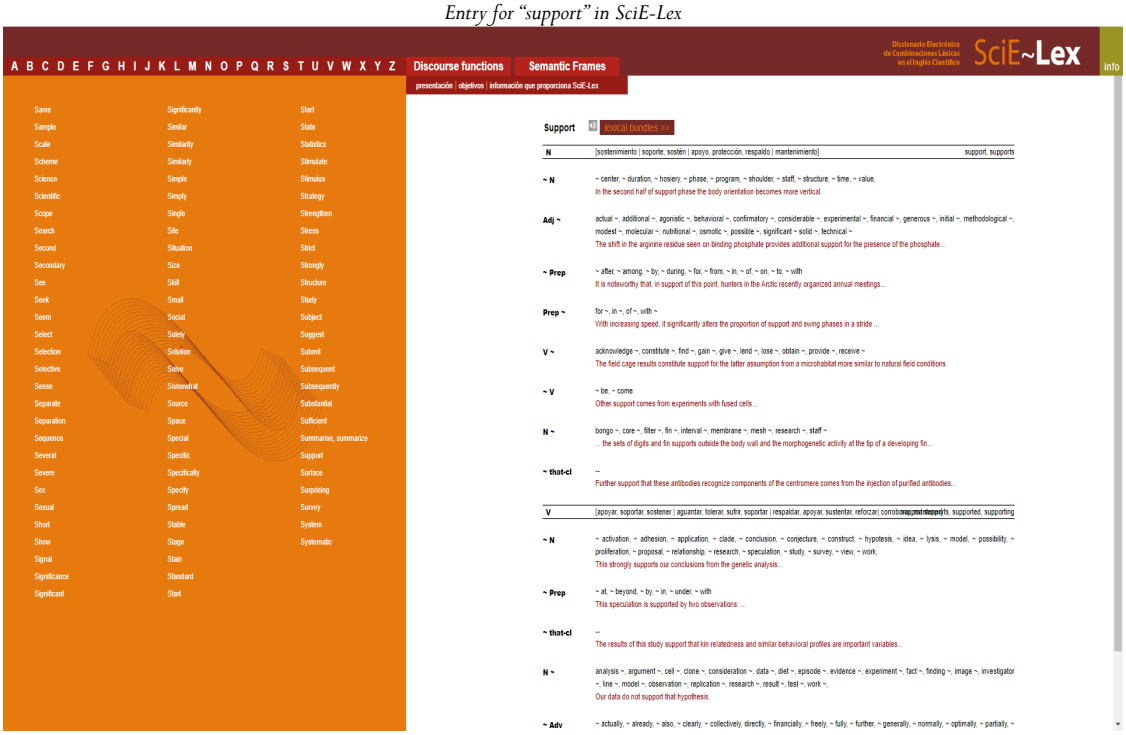



Figure 1 shows that the alphabet list can be found in the top left bar. When clicking on any letter, a drop-down menu of the entries available is displayed. If we take the example of the word entry *support*, first of all, the user finds a voice icon, which allows the user to hear the pronunciation of the search term. Below, there is information about the grammatical category of the search word. In this case it can be observed that the lemma *support* can be used as a noun and as a verb. Next to the category, there is the equivalent term in Spanish as well as morphological features, which inform about the different word forms of the lexical entry. In this case, as a noun, it can be used as both a countable and an uncountable noun (with different meanings, though), so both the base form and the inflected form for the plural are included. When used as a verb, on the other hand, we find the prototypical grammatical variants of regular verbs; that is, base form */-s* form (3rd person singular), *-ed* form (finite past, and non-finite past participle and passive participle) and finally *-ing* form (i.e., non-finite, gerund).

The equivalent terms in Spanish may need further clarification, particularly when dealing with polysemous terms, like the verb support. If a word is polysemous either in English or in Spanish, it is necessary to clarify the sense in which it is used on each occasion by means of a gloss or synonymous terms. As shown in Figure 1, the multiple senses of the verb *support* are frequently equivalent to *apoyar/sostener* in Spanish, but its various meanings (e.g., *corroborar unos resultados, reforzar conclusiones*, etc.) must be distinguished.

As discussed previously, cross-references to other entries (linked entries) can sometimes be useful; for instance, when words are morphologically and/or semantically related. In the example discussed here, it has already

ISSN 2696-676X


been noted that the search word has two uses: as a noun and as a verb, and both are morphologically and semantically related.

One of the strengths of SciE-Lex as a lexicographic tool lies in the fact that it displays information about the grammatical patterning associated with the various meanings found in its dictionary entries. It is essential to be familiar with the interaction that occurs between the meaning of a term and its valence; that is, the elements a given term subcategorises for, since in many cases the different meanings of a term are expressed through different grammatical patterns. This information is key to the correct construction of a sentence, especially when dealing with verbs.

As shown in Figure 2, for example, *support* as a verb can be followed by a noun, by a *that*-clause, by a preposition or by an adverb. Likewise, it is preceded by a noun in subject position. These grammatical constructions are illustrated by the most frequent lexical combinations and there has also been included a list of the terms that appear most recurrently in the HSC corpus in combination with the search word; for instance, *analysis/research/findings/data + support*; *generally/partially/strongly + support*. Word collocates are organized by semantic fields and, within each field, they are arranged alphabetically.

**Figure 2**

*Grammatical patterns associated with the verb "support" in SciE-Lex*

| | | |
|---|---|---|
| **V** | [apoyar, soportar, sostener \| aguantar, tolerar, sufrir, soportar \| respaldar, apoyar, sustentar, reforzar\| corroborar, mantener \| asistir \| dar fuerzas a \| acompañar] | support, supports, supporting, supported |
| **N ~** | analysis ~, argument ~, cell ~, clone ~, consideration ~, data ~, diet ~, episode ~, evidence ~, experiment ~, fact ~, finding ~, image ~, investigator ~, line ~, model ~, observation ~, replication ~, research ~, result ~, test ~, work ~,<br>Our data do not support that hypothesis. | |
| **~ N** | ~ activation, ~ adhesion, ~ application, ~ clade, ~ conclusion, ~ conjecture, ~ construct, ~ hypotesis, ~ idea, ~ lysis, ~ model, ~ possibility, ~ proliferation, ~ proposal, ~ relationship, ~ research, ~ speculation, ~ study, ~ survey, ~ view, ~ work,<br>This strongly supports our conclusions from the genetic analysis... | |
| **~ Adv** | ~ actually, ~ already, ~ also, ~ clearly, ~ collectively, directly, ~ financially, ~ freely, ~ fully, ~ further, ~ generally, ~ normally, ~ optimally, ~ partially, ~ potentially, ~ poorly, ~ rigidly, ~ still, ~ strongly, ~ thus, ~ weakly, ~ widely,<br>Our observations generally support the pioneering biochemical studies that first established... | |
| **~ Prep** | ~ at, ~ beyond, ~ by, ~ in, ~ under, ~ with<br>This speculation is supported by two observations: ... | |
| **~ that-cl** | --<br>The results of this study support that kin relatedness and similar behavioral profiles are important variables... | |



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021| Aceptación: 20/12/2021| Publicación: 03/01/2021

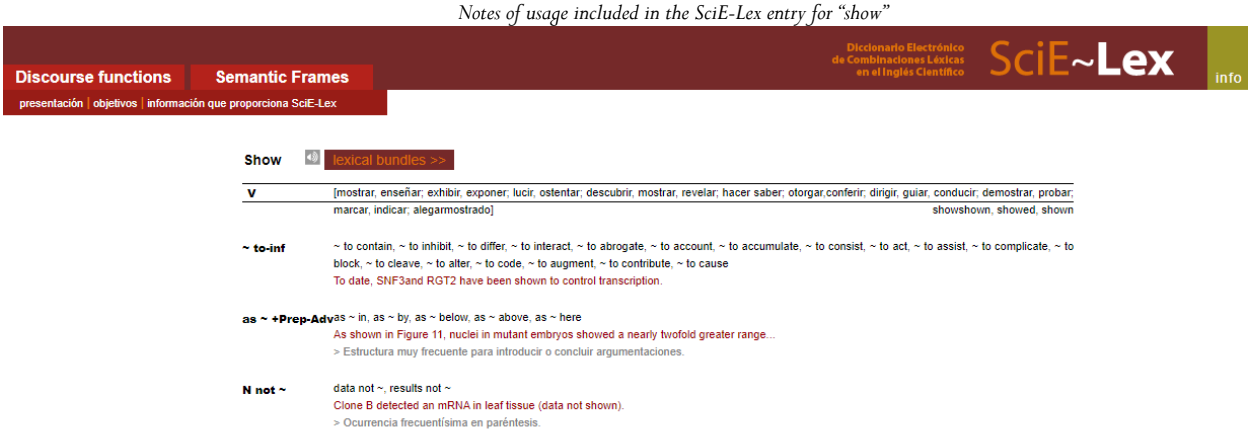
ISSN 2696-676X


Examples of actual use are particularly useful. The selected examples illustrate and complete the information provided in each entry. These examples are inspired by the HSC corpus, but they are not exact transcriptions of the phrases found, which are often of great complexity. However, the examples show authentic language in use: *it is noteworthy that*, *in support of this point*, (…) // *other support comes from experiments with fused cells*…, among others.

As shown in Figure 3, some entries also include clarifying notes to highlight special uses and / or help the dictionary user to use a term correctly. It is noted in some cases, for example, that the search term is usually used in the passive (if it is a verb) or more usually in the plural (in the case of a noun), or if it appears after a period in initial position, etc.

**Figure 3**

*Notes of usage included in the SciE-Lex entry for "show"*




In its second developmental stage, and in line with new trends in corpus-based and phraseological studies, SciE-Lex was supplemented with formulaic expressions (referred to as lexical bundles in SciE-Lex) that contribute to the characterisation of the prototypical phraseology of the health science discourse and of the structural organization of the biomedical research article. To the right of the headword, there is a hyperlink to "Lexical Bundles", which displays the different expressions in which the search term frequently appears. Some entries do not show this information because they are not associated with any specific phraseology that is noteworthy.

Figure 4 shows an example of the "Lexical Bundles" page in SciE-Lex. On the left-hand side is a list of the different prefabricated expressions in which the headword is frequently used in this domain, as well as the different variants of each expression; that is to say, those forms that are more recurrent. For example, *supports the idea that; supported by, in support of*, etc. In the central column, under the label "Discourse Function", the different discursive functions performed by each of the expressions are displayed. Finally, the right column ("Text Distribution") provides information about the distribution of these expressions in the research article; that is, if a given expression appears more frequently in the Introduction section, in the Discussion, in the Results, in the Methodology, in the Conclusions, etc. of the biomedical research article. This functionality is likely to help the user decide what type of structure is the most appropriate so as to express a certain function depending on the section of the article.



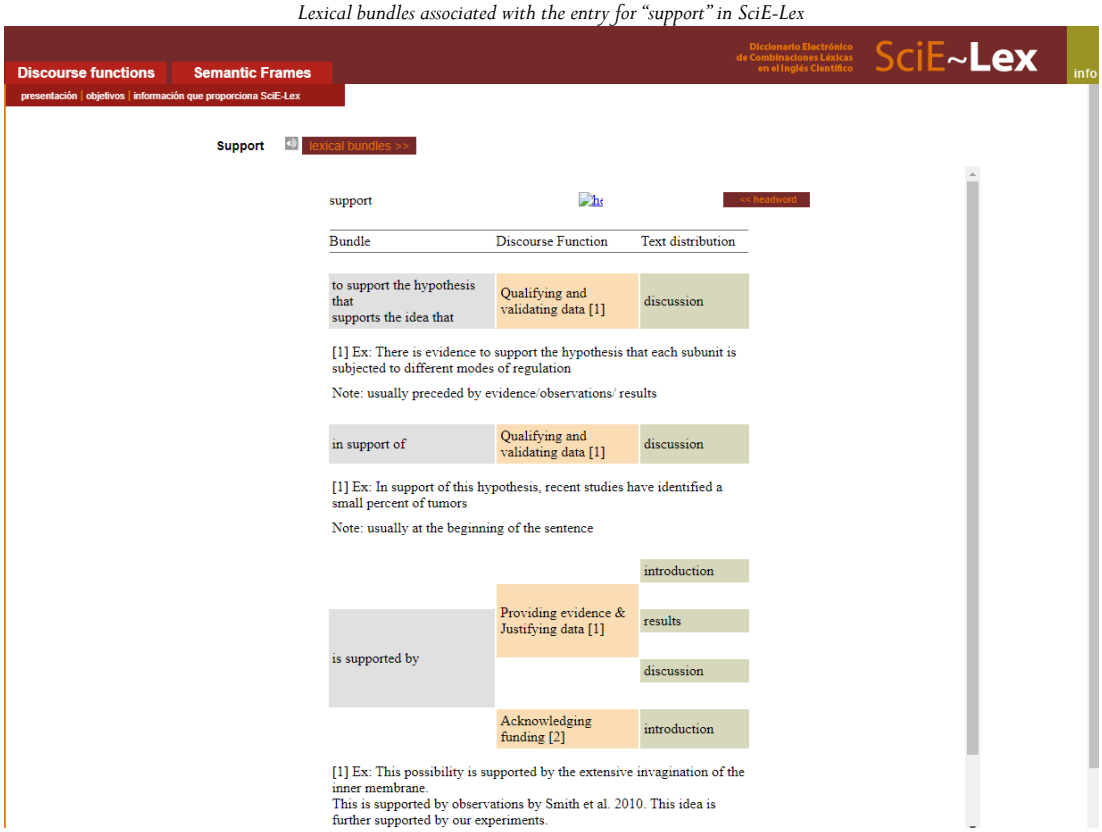
Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021| Aceptación: 20/12/2021| Publicación: 03/01/2021

ISSN 2696-676X


**Figure 4**

*Lexical bundles associated with the entry for "support" in SciE-Lex*




This section of SciE-Lex is complemented with HSC examples, which illustrate each of the selected expressions, as well as explanatory notes of usage, which provide information about the specific usage characteristics of each expression. For example, in this case it can be observed that the expression *in support of*, which is used to introduce the validation of the data that is described, frequently appears in the "Discussion" section of the biomedical research article, and usually appears in initial position introducing a sentence. Likewise, the expressions *to support the hypothesis that* and *supports the idea that* are usually preceded by nouns such as *evidence/observations/results*.

In order to proceed more systematically and efficiently, the most relevant discourse functions were included in a drop-down menu at the top of the screen. If we click on "Discourse Functions", we see that a list of functions appears on the left-hand side of the screen (see Figures 5, 6 and 7). If we click on any of them (e.g., "Bringing the reader's attention to a point"/"Describing procedures"/"Expressing lack of probability & Hedging"), on the right-hand side a list of all the expressions that are frequently used to express each one of those functions is displayed. This information can be extremely useful for SciE-Lex users since it provides them with different expressions that fulfil a similar function and that will help them avoid unnecessary repetitions of the same construction, which may have a negative impact on the quality of a text. On the contrary, the use of synonymous expressions will enhance the lexical richness of the research article that the SciE-Lex user is writing.



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021 | Aceptación: 20/12/2021 | Publicación: 03/01/2021

ISSN 2696-676X


**Figure 5**

*Lexical bundles associated with the discourse function "bringing the reader's attention to a point" in SciE-Lex*

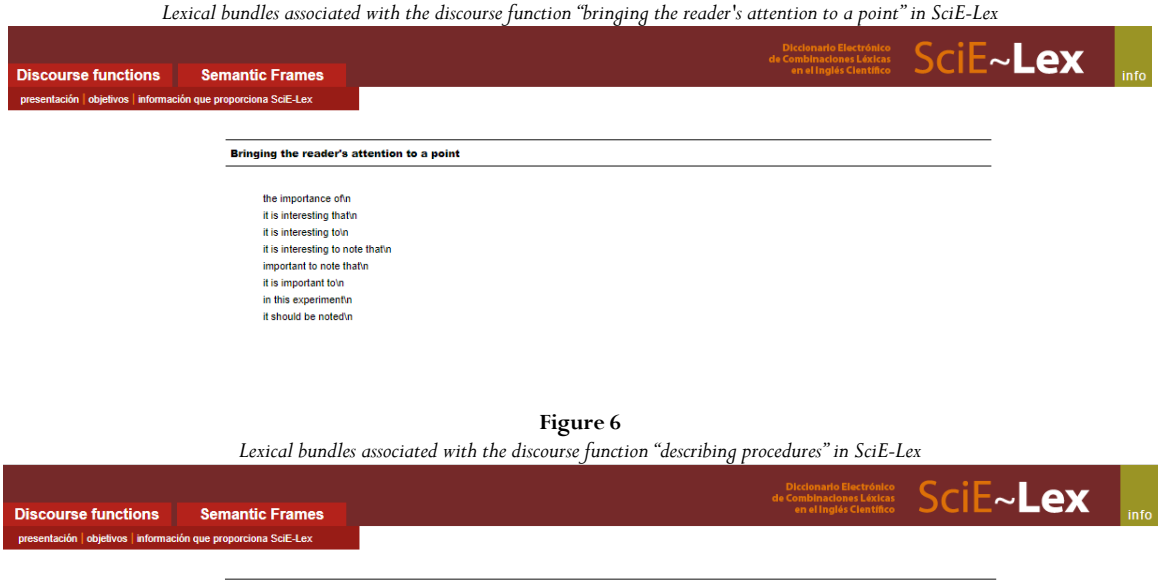



**Figure 6**

*Lexical bundles associated with the discourse function "describing procedures" in SciE-Lex*

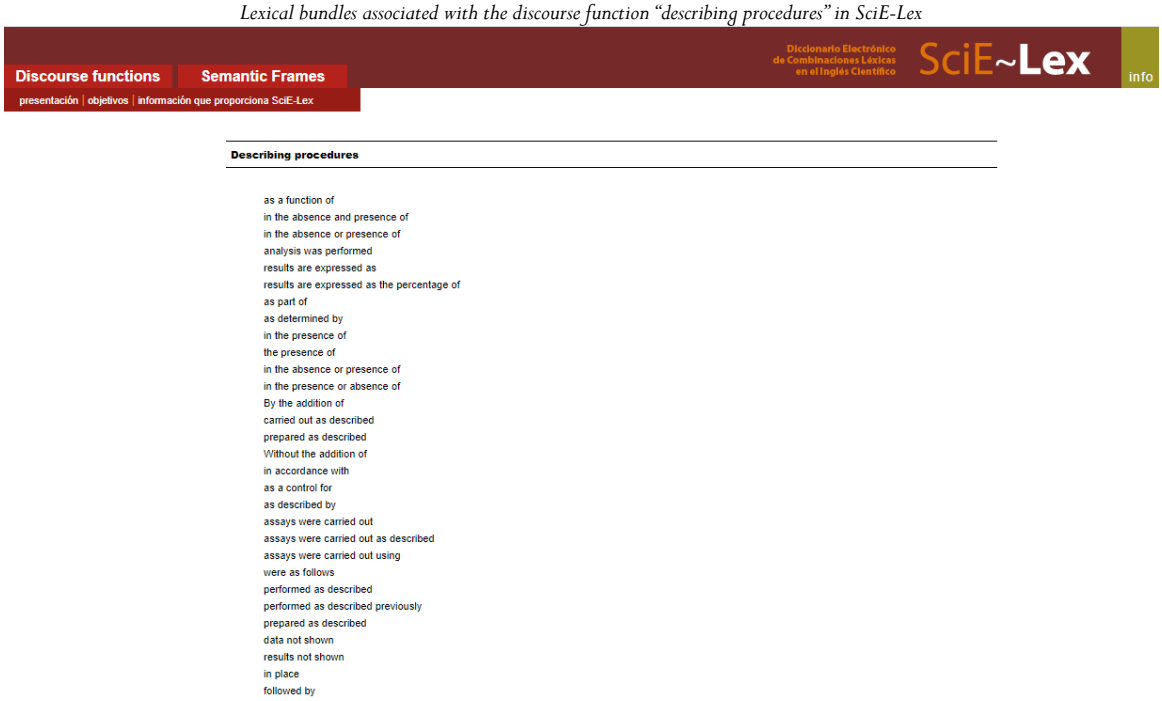
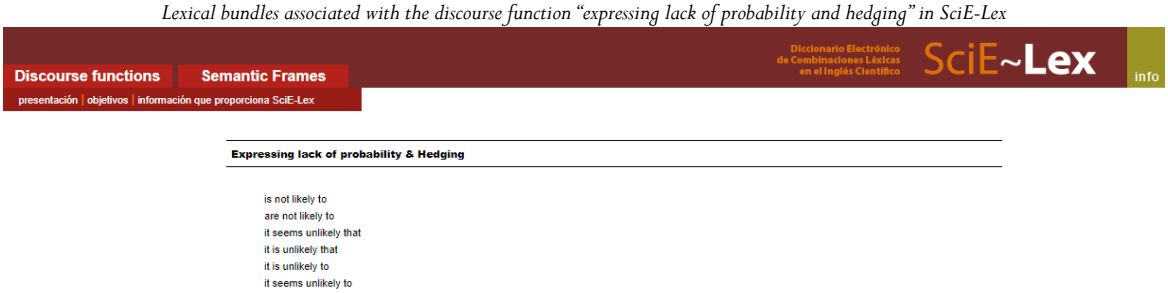



**Figure 7**

*Lexical bundles associated with the discourse function "expressing lack of probability and hedging" in SciE-Lex*






Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021 | Aceptación: 20/12/2021 | Publicación: 03/01/2021

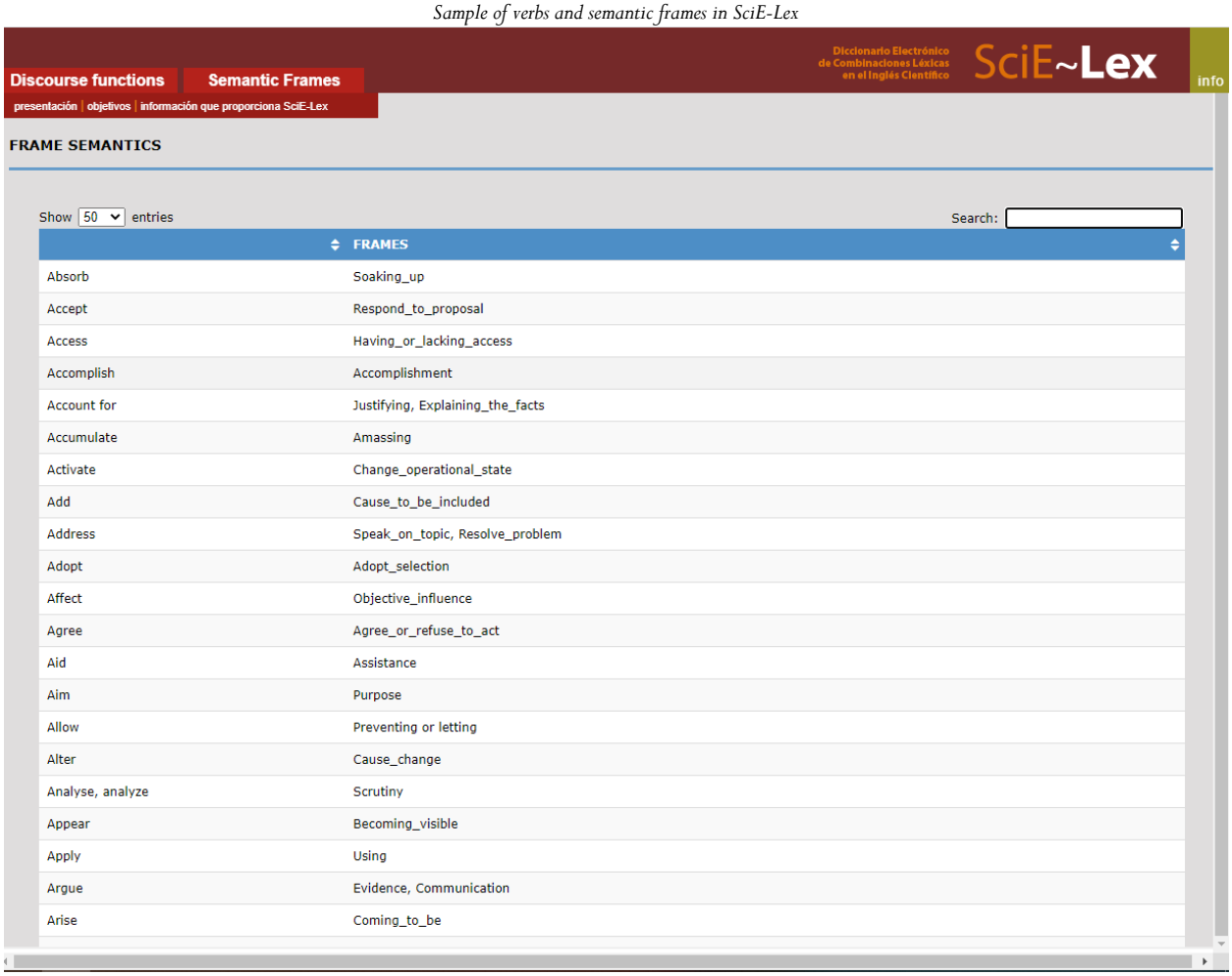
ISSN 2696-676X


Therefore, the introduction of lexical bundles and their contextual information in the SciE-Lex database has contributed to the further development of the dictionary, which has also been at a later stage supplemented with information regarding the distribution of lexical bundles across the different sections and/or moves of the academic research article as well as their discourse function.

The GreLiC group has been working lately on setting the semantic frames (FrameNet project) to which the verbs in SciE-Lex (a total number of 315 entries) belong (Verdaguer & Laso, 2020; Verdaguer *et al.*, 2020). As shown in Figures 8 and 9 below, this would allow the user to have easy access to all the information about the evoked semantic frames which FrameNet provides, and to identify the verbs which are associated with the same frame. In addition, its application reveals the semantic, syntactic, and collocational connections of the verbs that belong to the same framework and helps find similarities and differences between related words, whose command is likely to contribute to improving written production.

**Figure 8**

*Sample of verbs and semantic frames in SciE-Lex*



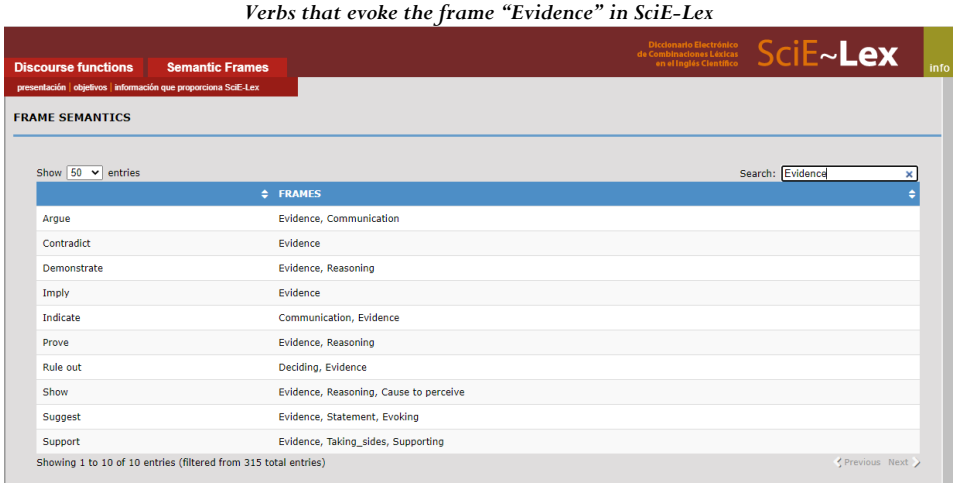



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021 | Aceptación: 20/12/2021 | Publicación: 03/01/2021

ISSN 2696-676X


**Figure 9**
*Verbs that evoke the frame "Evidence" in SciE-Lex*




## 3. Making Use of SciE-Lex to Assist the Production of Biomedical Texts

As already stated, SciE-Lex was developed with the primary aim of helping Spanish biomedical scientists to use the correct lexicogrammatical patterning of non-technical words and to conform to the conventional collocations used in the biomedical discourse. Consequently, it was of paramount importance to evaluate both the adequacy and the effectiveness of the tool to assist the production of biomedical texts in English. To this end, SciE-Lex has undergone a thorough revision process among experts in the lexicographic field. It has been tested by a range of targeted users, whose feedback has provided a first-hand account of the usefulness of this lexicographic tool.

First of all, SciE-Lex was trialled by a group of lexicographers from the University of Barcelona (4), the University of Lleida (1) and a tutor from the English for International Students Unit at the University of Birmingham, predominantly working on the area of applied linguistics. Regarding their L1, two participants in the trial were L1 English speakers, and the others were either L1 Spanish speakers (1) or L1 Catalan speakers (2).

They were all provided with some information about both the macrostructure and the microstructure of SciE-Lex, as well as the different stages involved in the dictionary-making process (see Verdaguer *et al.*, 2013). A preliminary version of the online demo of SciE-Lex was made available to all participants at http://www.ub.edu/grelic/demo/index.html and they were requested for a short online report upon consultation of the tool.

Bearing in mind that we were especially interested in testing the usefulness and the user-friendliness of the information included in SciE-Lex, raters were asked to give their opinion about the relevance of the information contained in the microstructure of the database and the manageability of the tool, on the one hand, as well as identify its strengths and any other issues in need of improvement, on the other.

As illustrated in Table 2 below, the report covered four main areas, namely: lexicogrammatical information,



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021| Aceptación: 20/12/2021| Publicación: 03/01/2021

ISSN 2696-676X


phraseological information, translation equivalents, and style refinements. Some headings were supplied to help experts focus their qualitative comments and to ensure that key areas of concern were addressed. However, experts were also encouraged to write about any further insights into SciE-Lex that could be provided.

**Table 2**
*SciE-Lex Experts' Report*

| | |
|---|---|
| **Lexicogrammatical information** | Morphological variants |
| | Patterns of occurrence |
| | List of collocates |
| | Examples of real use |
| | Usage notes |
| | Other comments |
| **Phraseological information** | Use of phraseological units (lexical bundles) |
| | Adequacy of discourse functions |
| | Examples of use |
| | Textual distribution information |
| | Usage notes |
| | Other comments |
| **Translation equivalents** | Naturalness |
| | Accuracy of equivalent terms |
| | Other comments |
| **Style refinements** | Strengths |
| | Weaknesses |
| | Suggest ways of making the dictionary more suitable for intended users |

Regarding lexicogrammatical information, experts highlighted the appropriateness of including information about the combinatorial patterning of each dictionary entry. Their comments stressed the relevance of providing phraseological combinations that are highly illustrative of the prototypical phraseologies of the health science discourse (i.e., biology, biochemistry, and medicine, mainly). This point stood out as one of the strong points of the tool since it was thought that SciE-Lex aims to cater for intended users of the dictionary. Actually, some of the raters' comments suggested that potential users who could benefit from SciE-Lex would be primarily health scientists who need to publish their papers in English, as well as translators of scientific texts. This finding was extremely reassuring as the tool was thought to be helpful for both encoding and decoding purposes.

The fact that each lemma in SciE-Lex is associated with an equivalent term in Spanish, rather than a definition, was also regarded as a strong point of the tool, because equivalent terms help users focus on a given word meaning at a time, which is particularly useful to disambiguate polysemous entries. For instance, the adjective *free* may be used with the meaning 'without cost' (Sp. *gratuito*) as in *They were allowed free access to water* or with the meaning 'not held, tied, or fixed to somewhere' as in *Free steroids could be detected*. Likewise, the translation equivalents used in the database were labelled as natural choices in Spanish.



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

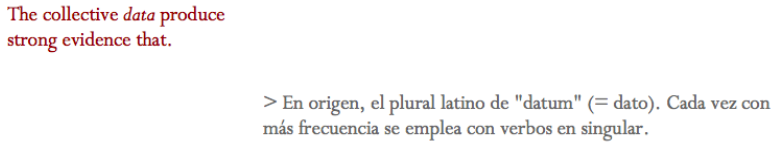
 | Recepción: 03/12/2021 | Revisión: 11/12/2021| Aceptación: 20/12/2021| Publicación: 03/01/2021

ISSN 2696-676X


The incorporation of collocational information was regarded as one of the main strengths of SciE-Lex since it covers the most prototypical contexts of occurrence of each search time, which are illustrated by corpus-based examples and thus intended users are faced with natural examples characteristic of the health science discourse, which have been extracted from the same type of journals users would be sending their manuscripts to.

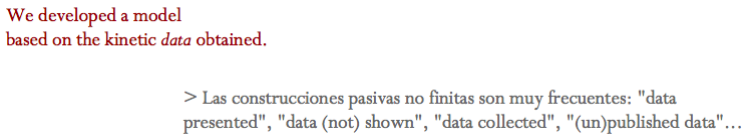
Similarly, experts also welcomed the inclusion of notes that clarify and fine-tune certain uses, since they are considered to be a great help to the users when deciding. For example, whether a given noun is more frequently used in its singular or plural form:

**Figure 10**
*Example of a note of usage in SciE-Lex (I)*




or whether a verb is more commonly found in a passive construction, to name but a few examples:

**Figure 11**
*Example of a note of usage in SciE-Lex (II)*




Supplementing dictionary entries with lexical bundles was expected to be another remarkable feature of the tool, as corroborated by experts. They valued positively the fact that the access to information about lexical bundles access was twofold; that is, they can be searched for by the link to the headword they are associated with or by discourse function. The information on lexical bundles was also considered to be very thorough as it contains the following features:

a) morphological information (i.e., inflected forms),
b) corpus-based examples that illustrate each bundle,
c) information about their most prototypical discourse functions,
d) their most frequent distribution across the health science paper (see Figure 4).

Most comments from the experts stressed the fact that the section on lexical bundles would definitely help intended users of the tool be acquainted with other possible ways of expressing their ideas, and with the section(s) of the research article where a given bundle is more likely to be found.

Concerning the weaknesses of SciE-Lex, experts, after thorough revision of the tool, highlighted a few issues that could prove challenging for users and that needed some revisiting. First, they pointed out the convenience of including more examples in each dictionary entry, since not all word combinations are illustrated by means of a corpus-



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021| Aceptación: 20/12/2021| Publicación: 03/01/2021

ISSN 2696-676X


based example. Although it is true that, unlike paper dictionaries, SciE-Lex is not constrained by space limitations, the reason behind this lexicographic decision lies in the fact that, bearing in mind that all examples come from already published articles, it was agreed that only one example would be selected to illustrate each colligation.

Some other comments referred to the fact that some meanings are missing in polysemous words, but this is because only those word senses found in the HSC (reference corpus) were included; that is, SciE-Lex entries were restricted to the prototypical word meanings that users of the tool were likely to be using in their publications.

Last but not least, some experts suggested that the inclusion of a search box in the tool would certainly provide users with searching capabilities, which is going to be adopted in for future updates of SciE-Lex.

The outcomes of the experts' reports have positively contributed not only to improving the tool, but also to validating the main aim of SciE-Lex, which was primarily "conceived as a response to the lack of reference tools that include contextual information on the lexicogrammatical patterning of non-technical terms frequently used in the scientific domain." (Laso *et al.*, 2019, p. 45).

As mentioned above, SciE-Lex has also been tested with targeted users in order to check its validity and usefulness as a tool, as well as to examine its pedagogical benefits for the biomedical community writing in English for research publication purposes (ERPP) (Cargill & Burgess, 2008; Laso *et al.*, 2019). To this end, four rounds of "Writing for Publication" workshops have been conducted by Laso & John at the University of Barcelona (during the years 2014, 2018, 2019 and 2020) and addressed to a group of biomedical researchers from leading research institutions, such as the following: *CRESA-Centre de Recerca en Salut Animal* (UAB-IRTA), a public foundation created in 1999 to conduct research on 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 of the highest international quality level which helps to improve health and quality of life and generate wealth.

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 non-native English-speaking writers of specialised discourses. Additionally, these workshops have allowed participants to familiarise themselves with SciE-Lex through a series of exercises and with formulaic language typical of the health science discourse. This experience has, consequently, served the purpose of helping participants get a better chance of having their manuscripts published in high impact international journals (Laso & John, 2017, p. 152).

Overall, "Writing for Publication" workshops have undoubtedly showed that SciE-Lex can be used to improve draft biomedical research articles from a lexicogrammatical point of view. They have also provided a strong indication of the benefits of L2 English writers being able to recognise the formulaic patterning of biomedical discourse.



Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *1*, e-a_hr03, 1-16.
https://doi.org/10.1344/teisel.v1.37444

 | Recepción: 03/12/2021 | Revisión: 11/12/2021| Aceptación: 20/12/2021| Publicación: 03/01/2021

ISSN 2696-676X


## Conclusion

This paper has provided detailed information about SciE-Lex, a resource tool aimed at assisting with the efficient production of published biomedical discourse. Spanish scientists need to master the language if they want their research to reach an international audience. This linguistic competence refers not only to lexicogrammatical issues, but also to the phraseological conventions and the typical characteristics of the language used by their discourse community.

The development of SciE-Lex has involved application of the Pattern Grammar framework (Hunston & Francis, 2000) to health science discourse. An analysis of the patterns associated with general terms in biomedical discourse (Verdaguer *et al.*, 2013) called attention to the fact that, in addition to resources that provide lexicogrammatical and discourse features of general English, further tools are required to suit the needs of specialised discourse communities (see also Hunston, 2008 and 2009). The SciE-Lex tool provides non-native English-speaking writers with the prototypical use of lexicogrammatical patterns of non-technical words, as well as the conventionalised phraseological characteristics of their discourse community. This is particularly relevant in scientific discourse, since the more acquainted health science researchers are with the written conventions of their research field, the better chances they will have to produce phraseologically competent texts and thus to get their results published in international settings.

The shortage of dictionaries and reference tools providing information about the use of non-technical terms and their preferred phraseologies in biomedical English justified thus the need for SciE-Lex. It is our hope that the dissemination of the tool coupled with the publication of this paper and further editions of "Writing for Publication" workshops will serve to inspire further investigation on useful writing resources to assist ERPP across disciplines.

## Acknowledgements

I would like to thank the GreLiC Research Group for their helpful comments during the drafting of this paper and anonymous reviewers, whose insightful comments helped me strengthen the paper considerably.

## References

Cargill, Margaret; Burgess, Sally (2008). Introduction to the Special Issue: English for research publication purposes. *Journal of English for Academic Purposes*, *7*(2), 75-76. https://doi.org/10.1016/j.jeap.2008.02.006

Coxhead, Averil (2000). A new academic Word List. *TESOL Quarterly*, *34*(2), 213-238. https://doi.org/10.2307/3587951

Fillmore, Charles J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, *280*, 20-32. https://doi-org.sire.ub.edu/10.1111/j.1749-6632.1976.tb25467.x

Fuertes-Olivera, Pedro; Niño, Marta; Sastre, Ángeles (2019). Tecnología con fines lexicográficos: su aplicación en los Diccionarios Valladolid-Uva. *Revista Internacional de Lenguas Extranjeras*, Monográfico *10*, 75-100. https://doi.org/10.17345/rile10.2556

ISSN 2696-676X


Gries, Stefan Th.; Stefanowitsch, Anatol (Eds.). (2006). *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Mouton de Gruyter. https://doi.org/10.1515/9783110197709

Hunston, Susan (2002). Pattern grammar, language teaching, and linguistic variation. Applications of a corpus-driven grammar. In Randi Reppen, Susan M. Fitzmaurice & Douglas Biber (Eds.), *Using Corpora to Explore Linguistic Variation* (pp. 167-183). John Benjamins Publishing. https://doi.org/10.1075/scl.9

Hunston, Susan (2008). Starting with the small words: Patterns, lexis, and semantic sequences. *International Journal of Corpus Linguistics*, *13*(3), 271–295. https://doi-org.sire.ub.edu/10.1075/ijcl.13.3.03hun

Hunston, Susan (2009). The usefulness of corpus-based descriptions of English for learners: the case of relative frequency. In Karin Aijmer (Ed.), *Corpora and Language Teaching* (pp. 141-156). John Benjamins Publishing. https://doi.org/10.1075/scl.33.13hun

Hunston, Susan; Francis, Gill (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins Publishing. https://doi.org/10.1075/scl.4

Laso, Natalia Judith (2009). *A corpus-based study of the phraseological behaviour of abstract nouns in medical English: A needs analysis of a Spanish medical community* [unpublished PhD dissertation, University of Barcelona]. https://dialnet.unirioja.es/servlet/tesis?codigo=92964

Laso, Natalia Judith; Comelles, Elisabet; Verdaguer, Isabel (2019). Research report on the adequacy of SciE-Lex as a lexicographic tool for the writing of biomedical papers in English. *Digital Scholarship in the Humanities*, *34*(1), 32-47. https://doi.org/10.1093/llc/fqy015

Laso, Natalia Judith; John, Suganthi (2017). The pedagogical benefits of a lexical database (*SciE-Lex*) to assist the production of publishable biomedical texts by EAL writers. *Iberica*, *33*, 147–172. http://www.aelfe.org/documents/33_06_IBERICA.pdf

L'Homme, Marie-Claude (2005). Conception d'un dictionnaire fondamental de l'informatique et de l'Internet: sélection des entrées. *Le langage et l'homme*, *40*(1), 137-154. http://olst.ling.umontreal.ca/pdf/lel-lhomme-2005.pdf

L'Homme, Marie-Claude (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire*, *217*, 78-103. https://doi-org.sire.ub.edu/10.4000/traduire.966

Paquot, Magali (2010). *Academic Vocabulary in Learner Writing*. Bloomsbury. http://doi.org/10.5040/9781474211697

Reimerink, Arianne; Faber, Pamela (2009). Ecolexicon: A frame-based knowledge base for the environment. In Jiří Hrebícek; J.H. Mírovsky; Werner Pillmann; I Holoubek; T. Bandholtz (Eds.), *European conferencee of the Czech Presidency of the Council of the EU TOWARDS eENVIRONMENT Opportunities of SEIS and SISE: Integrating Environmental Knowledge in Europe* (pp. 25-27). Masaryk University.

Scott, Mike (1997). *Wordsmith Tools version 2*. Oxford University Press.

Sinclair, John M. (Ed.). (1987). *Looking Up: An Account of the Collins COBUILD Project*. Collins ELT.

Verdaguer, Isabel; Castaño, Emilia & Laso, Natalia Judith (2020). Semantic frames in SciE-Lex. In Miguel Fuster-Márquez; Carmen Gregori-Signes; José Santaemilia Ruiz (Eds.), *Multiperspectives in analysis and corpus design* (pp. 61-72). Comares.

Verdaguer, Isabel; Laso, Natalia Judith (2020). Construcción de un diccionario combinatorio de inglés biomédico, SciE-Lex. *Revista de Lexicografía*, *XXVI*, 159-174. https://doi.org/10.17979/rlex.2020.26.0.6049

Verdaguer, Isabel; Laso, Natalia Judith; Salazar, Danica (2013). *Biomedical English: a corpus-based approach*. John Benjamins Publishing. https://doi.org/10.1075/scl.56