Corpus-based selection of frequent noun + noun sequences for an English for Specific Purposes course for chemistry students

Carolina Beatriz Panza
Universidad Nacional de Río Cuarto (UNRC), Argentina
cbpanza@hum.unrc.edu.ar

(First received: September 1, 2024; accepted: October 24, 2024)

Abstract

This classroom account describes the use of a small corpus of laboratory safety manuals as a source of relevant linguistic realisations for materials design in a course of English for undergraduate students of chemistry. The corpus was used to extract frequent noun + noun sequences to design pedagogical applications which can help students understand the logical relations between the two nouns, with the ultimate aim of enhancing students' reading comprehension of disciplinary texts in English.
Key words: ESP, chemistry, noun sequences, laboratory safety manuals, corpus design

Resumen

Este relato áulico describe el uso de un pequeño corpus de manuales de seguridad de laboratorio como fuente de exponentes lingüísticos relevantes para el diseño de materiales en un curso de inglés para estudiantes universitarios de Química. El corpus fue utilizado para extraer secuencias frecuentes de sustantivo + sustantivo para diseñar aplicaciones pedagógicas que puedan ayudar a los estudiantes a comprender las relaciones lógicas entre los dos sustantivos, con el objetivo final de mejorar la comprensión lectora de textos disciplinarios en inglés.
Palabras claves: inglés para propósitos específicos, química, secuencias de sustantivos, manuales de seguridad de laboratorio, diseño de corpus

Context

English has become the language for communication among scholars working in different fields and disciplines (Hyland, 2006; Swales, 1990, 2004; Wood, 2001), to the extent that in some disciplines over 90% of the most prestigious journals are published in this language. Reading fluently in English is considered, then, a crucial skill for undergraduate students to advance in their studies, keep updated with the latest findings in their research areas, and eventually participate in their disciplinary discourse communities. To address this need, in Argentina, most undergraduate programs prepare students for reading disciplinary texts in English. In this line, at the National University of Río Cuarto, students of chemistry have to take two compulsory courses of English in the first year of their undergraduate program: English 1 and English 2.

The classroom account presented here describes how a genre-based corpus specifically designed to address the objectives of the course English 1 was used to select noun + noun sequences (hereafter, NN sequences) relevant to students' needs. This linguistic resource is particularly frequent in science-related texts and its use and evolution in this register has been reported in detail by Biber and Gray (2011).

The rationale underlying syllabus organization and materials design for English 1 draws on principles of genre analysis from the English for specific purposes school (Bhatia, 1993, 2004, 2012; Dudley-Evans, 1994; Swales, 1990, 2004). The present account describes the use of a genre-based corpus as a suitable source of linguistic examples for pedagogical purposes. Frequent NN sequences were identified in a corpus of laboratory safety manuals in order to illustrate a range of logical relations that can be expressed by this pattern. The ultimate purpose of this selection was to explicitly teach students different meaning relations that can be expressed by these sequences, particularly focusing on how to express these meanings in Spanish using suitable prepositions or paraphrasing.

NN sequences

Academic and scientific texts are usually lexically dense, concentrating information on complex noun groups (Halliday, 1993, 1998, 2004). This usually poses a challenge to students of English as a foreign language, since meanings which are usually expressed by one or more clauses can be packed into a single complex noun group (Halliday & Matthiessen, 2014). One particularly productive resource in English for creating complex noun groups is the use of a noun as a premodifier of another noun. When comparing this pattern in English and Spanish—the native language of students in the course English 1—, it is important to understand that there are differences in "the amount of structure (syntactic or phonological) that the modifier position allows in these languages" (Marqueta-Gracia, 2017, p. 15). In Spanish, NN sequences mainly occur in compounds and can be written as a single word (e.g., *telaraña*), with a hyphen (*lavadora-secadora*) or as two separate words (e.g., *experiencia piloto*, *fecha limite, copia pirata*) (Real Academia Española, 2009). This last option is considered a syntagmatic compound, as its components have certain prosodic and morphological independence, i.e., words keep their accent as single words and the words are written separately (Real Academia Española, 2009). These syntagmatic compounds in Spanish are usually left-headed, e.g., ***palabra*** *clave*, ***copia*** *pirata*, whereas in English they are typically right-headed, e.g., *lab* ***coat,*** *eye* ***protection***. However, this is not the most striking difference between NN sequences in both languages. The main difference lies in the possible syntactic configurations within the noun group, i.e., in English, nouns (one or more) can function as premodifiers in the noun group. Thus, in English, the NN pattern is particularly productive, even allowing for recursivity with sequences of several nouns, e.g., *materials safety data sheet.* The borderline between noun compounds and noun modifier + noun head in English is in fact a cline (Biber et al., 1999). Biber et al., (1999) used the criterion of orthographic separation of words to identify NN sequences and found that 30% of all premodifiers in academic prose were nouns (p. 589). Their criterion of orthographic separation between the two nouns has been employed here to identify relevant NN sequences for materials development for English 1.

These NN sequences, which only contain content words, require readers to infer the logical relation between the premodifying noun and the head noun (Biber, et al.,1999). Frequent logical relations between the two nouns identified by Biber et al. (1999, pp. 590–591) are C*omposition, Purpose, Identity, Content, Source, Objective Type 1, Objective Type 2, Subjective Type 1, Subjective Type 2, Time, Location Type 1, Location Type 2, Institution, Partitive, and Specialization*. The description of each logical relation is presented below, where N1 refers to the premodifying noun and N2 refers to the head noun.

Composition: N2 is made from N1; N2 consists of N1. Example: *tomato sauce.*

Purpose: N2 is for the purpose of N1; N2 is used for N1. Example: *war fund.*
Identity: N2 has the same referent as N1 but classifies it in terms of different attributes. Example: *men workers.*
Content: N2 is about N1; N2 deals with N1. Example: *algebra test.*
Source: N2 is from N1. Example: *plant residues.*
Objective Type 1: N1 is the object of the process described in N2 or of the action performed by the agent described in N2. Example: *egg production.*
Objective Type 2: N2 is the object of the process described in N1. Example: *discharge water.*
Subjective Type 1: N1 is the subject of the process described in N2; N2 is nominalized from an intransitive verb. Example: *child development.*
Subjective Type 2: N2 is the subject of the process described in N1. Example: *labour force.*
Time: N2 is found at the time given by N1. Example: *Summer conditions.*
Location Type 1 N2 is found or takes place at the location given by N1. Example: *Paris conference.*
Location Type 2 N1 is found or takes place at the location given by N2. Example: *notice board.*
Institution: N2 identifies an institution for N1. Example: *insurance companies.*
Partitive: N2 identifies parts of N1. Example: *cat legs.*
Specialization: N1 identifies an area of specialization for the person or occupation given in N2; N2 is animate. Example: *Education Secretary.*

Students whose first language is Spanish often find it difficult to understand the meanings expressed by NN sequences, as pointed out by Carrió Pastor (2008), who suggested that one possible cause may be the fact that many complex noun groups are "juxtaposed nouns without any preposition that would identify their semantic connection" (p. 27). This supports the idea that explicit teaching of NN sequences to undergraduate students is necessary. In a previous study in a similar context to that of the present account—a course of English for specific purposes for undergraduate students of Chemical, Natural, and Exact Sciences in a university in Argentina—Benassi et al. (2011) found that explicit theoretical instruction on the complex noun group enhanced students' comprehension of these groups.

Corpus design

A corpus is "a collection of naturally occurring texts used for linguistic study" (Hyland, 2006, p. 58). Specialized genre-based corpora, usually comprising between 20,000 and a million words, are considered "a body of relevant and reliable evidence" (Sinclair, 2001, p. xi) which can be used to answer specific research questions and to inform pedagogy in English for specific purposes (ESP) applications (Egbert et al. 2022; Flowerdew, 2002; Hunston, 2002; Sinclair, 1991, 2001, 2004). For the present classroom account, a corpus of ten laboratory safety manuals was compiled in order to be able to provide chemistry students with real examples of NN sequences which they will frequently encounter in the reading material of English 1. Laboratory safety manuals present relevant information for all personnel carrying out activities at university laboratories, from directors and supervisors to students and cleaning personnel. This genre was selected as reading material in English 1 as it is particularly suitable to be used with first year students who still do not have enough disciplinary knowledge to understand abstract ideas presented in scientific texts such as

research articles but can understand language referring to concrete practices in their disciplinary fields. Students can make sense of the context of production and use of this particular genre as they become lab users as soon as they start their careers. Therefore, undergraduate students of chemistry are a natural audience for these manuals.

To compile the corpus, ten laboratory safety manuals were selected from the online websites of universities around the world. Manuals were downloaded when the complete text was available in PDF version. A particular effort was made to select manuals from different countries, even if English was not the official language of the country (4 from USA, 2 from Canada, 1 from Spain, 1 from India, 1 from Turkey, and 1 from Australia). (For more details, see Corpus references). The PDF versions were saved as txt files so that they could be processed using the free linguistic software *AntConc 3.5.8* (Anthony, 2019). A list of all the words appearing in the corpus was created using the Wordlist Tool. The total number of types (i.e., distinct words) was 23,325 and the total number of words (tokens) was 339,580.

Identification of NN sequences

After creating a comprehensive list of all the words in the corpus, the most frequent nouns were analysed using the Concordance tool. For each analysed noun (NODE), a sorting was made, first to the left in order to see what frequent combinations were possible when the node was the head noun (see Figure 1) and then to the right to see what NN combinations were possible when the node word functioned as premodifier (See Figure 2).

Figure 1. Example of sorting to the left of the noun *equipment*

Figure 2. Example of sorting to the right of the noun *safety*

Examples of NN sequences were extracted when they appeared with a minimum frequency of five times in at least three different manuals. The examples were categorised using Biber et al.'s (1999) descriptions of logical relations. This procedure served to provide relevant examples to illustrate a range of logical relations expressed by these sequences in laboratory safety manuals. The logical relations more frequently expressed by NN sequences are presented below, illustrated with examples from the corpus.

Composition: N2 is made from N1; N2 consists of N1. Examples: *glass pipettes, metal powder, cardboard box, plastic bags, plastic containers.*
Purpose: N2 is for the purpose of N1; N2 is used for N1. Examples: *safety device, safety equipment, safety footwear, safety glasses, safety gloves, safety goggles, safety shield, evacuation plan.*
Content: N2 is about N1; N2 deals with N1. Examples: *safety chapter, safety considerations, safety data, safety guidelines, safety handbook, safety instructions, safety manual, safety section, safety regulations.*
Objective Type 1: N1 is the object of the process described in N2 or of the action performed by the agent described in N2. Examples: *tissue damage, eye protection, equipment decontamination, waste collection, waste disposal, waste management, waste minimization, laboratory use, lab inspection, radiation protection.*
Location Type 1 N2 is found or takes place at the location given by N1. Examples: *door signs, laboratory work, laboratory practices, laboratory refrigerators.*
Institution: N2 identifies an institution for N1. Examples: *safety department, safety council, safety commission.*
Partitive: N2 identifies parts of N1. Examples: *lab shelves, lab doors, lab sink.*
Specialization: N1 identifies an area of specialization for the person or occupation given in N2; N2 is animate. Examples: *lab technician, lab manager, lab director, lab supervisor, radiation committee, safety officer.*

Pedagogical applications

The list of frequent NN sequences identified in the corpus of laboratory safety manuals was used to design activities to raise students' awareness of the logical relationships holding between the nouns in the sequence. By the time this teaching point was introduced to students, they had already worked with laboratory safety manuals to understand the purpose of the genre, its typical rhetorical organization, the possible topics dealt within the genre, and the writers and intended readers of these manuals. Students have also worked with the basic structure of the noun group in English, and its pre- and post-modification.

The activities presented here aimed to expand students' knowledge of possible modifiers within the noun group, focusing on the use of nouns as premodifiers in the frequent NN sequences identified for laboratory safety manuals. All the activities were carried out in Spanish, using the input text in English. First of all, two frequent NN sequences (*safety manuals* and *laboratory practices*) were presented to students in order to illustrate two strategies that we can use to express these sequences in Spanish. One was the insertion of a preposition between the second and the first noun. Students were presented with a set of prepositions in Spanish that could be used to explicitly signal the meaning relationship between the two nouns (*a, ante, bajo, con, contra, de, desde, durante, en, entre, hacia, hasta, mediante, para, por, según, sin, sobre*). Students had to select the prepositions that could most accurately express the relationship between the two words. Students were shown that *safety manual* could be interpreted as "manual **sobre** seguridad" and *laboratory practices* as "prácticas **en** el laboratorio". The second option to help students uncover the meaning relationship between the two nouns was to teach them to paraphrase the NN sequence as suggested by Nakov (2008). In this case, students were shown that the NN sequence *safety manual* could be interpreted as "manual que trata acerca de la seguridad" and *laboratory practices* as "prácticas que se realizan en el laboratorio".

Students were explained that different types of meaning relationships were possible between the two nouns and that the most frequent ones in NN sequences found in laboratory safety manuals were *composition, purpose, content, objective, location, institution, partitive and specialization*. The following material was used in class to explain and illustrate different types of meaning relationships in NN sequences.

Secuencias de dos sustantivos: Relaciones de significados frecuentes

1- **Composición:** El sustantivo de la derecha está **hecho o consiste** del sustantivo de la izquierda. Ejemplo: *glass pippettes* (pipetas de vidrio/pipetas hechas de vidrio)
2- **Propósito:** El sustantivo de la derecha **se utiliza para lograr el propósito** indicado por el sustantivo de la izquierda. Ejemplo: *safety glasses* (anteojos de/para seguridad)
3- **Contenido:** El sustantivo de la derecha **trata sobre/es acerca** del sustantivo de la izquierda. Ejemplo: *safety manual* (manual que trata acerca de la seguridad)
4- **Objeto:** El sustantivo de la izquierda es el **objeto del proceso o acción implícita** en el sustantivo de la derecha. Ejemplo: *tissue damage* (daño a los tejidos/daño causado a los tejidos)
5- **Localización:** El sustantivo de la derecha **se encuentra o tiene lugar** en la localización indicada por el sustantivo de la izquierda. Ejemplo: *laboratory work* (trabajo en el laboratorio/trabajo que se realiza en el laboratorio)

6- **Institución:** El sustantivo de la derecha identifica una institución a cargo del sustantivo de la izquierda. Ejemplo: *safety department* (departamento de/para seguridad/departamento que está a cargo de la seguridad)
7- **Partitivo:** El sustantivo de la derecha **identifica partes** del sustantivo de la izquierda. Ejemplo: *lab doors* (puertas del laboratorio)
8- **Especialización:** El sustantivo de la izquierda **identifica un área de especialización de la persona u ocupación** señalada por el sustantivo de la derecha. Ejemplo: *lab technician* (Técnico de laboratorio/técnico que está especializado en el trabajo de laboratorio)

After this, students were presented with several excerpts from laboratory safety manuals selected because they presented NN sequences identified as frequent in the corpus. They had to read the excerpts and infer which section of the manual these may have been extracted from. This was done to make sure students understood the context of production of each segment. Then, students were guided to pay attention to the underlined NN sequences to try to discover the type of meaning relationship between the two nouns and to provide a version of the NN sequence in Spanish, either by paraphrasing or by inserting a suitable preposition. Some of the excerpts used are shown below. Because of space constraints not all the excerpts were included here. The source of each excerpt is included in parentheses.

Excerpts

1- A Principal Investigator (PI), Laboratory Supervisor, or their delegate may want to reassign an existing and unassigned lab coat to a new lab member rather than purchasing a new lab coat. (University of California Riverside)
2- Effective training is a critical component to facilitating a safe environment and for the prevention of laboratory accidents. All employees must be trained in general safe work practices and be given specific instructions on hazards unique to their job. (University of California Riverside)
3- Tissue damage begins immediately when a corrosive chemical comes in contact with the eyes or skin. (…) Lab coworkers are encouraged to guide the victim of a chemical splash to the appropriate emergency shower. Multiple copies of the relevant chemical material safety data sheet (MSDS) should be printed out and presented to emergency responders. (Harvard Department of Chemistry and Chemical Biology)
4- Identify the location of emergency eyewashes and safety showers, fire extinguishers, and other safety equipment (spill kit, etc.) before bringing hazardous materials to the new lab. (City University of New York)

The activity was done in groups and, in case students needed more examples of the same NN sequence, the corpus was used to extract several examples of the same sequence in different manuals. The activity served to provide students with relevant examples of NN sequences that they will encounter in their reading material, raise awareness of the complexities of this pattern, and help them develop strategies to understand the meaning expressed by these NN sequences. Many students commented that they were not aware of the relationships between nouns in the NN sequence before this class and that both using prepositions and paraphrasing helped them understand these sequences better. It is expected

that students will eventually use this knowledge to enhance their reading ability in English of other discipline-specific genres.

Conclusions

The present classroom account described one of the many uses that a specialized genre-based corpus can have for ESP teaching purposes. As described by Granath (2009),

> Corpora are invaluable for teachers, in that they can employ them in a number of ways, such as, for example, to create exercises, demonstrate variation in grammar, show how syntactic structures are used to signal differences in meaning and level of style, discuss near synonyms and collocations, and last (but not least) to give informed answers to students' questions. (p. 49)

For ESP courses, genre-based corpora are, then, extremely helpful resources that teachers can use to design materials which are relevant to address their students' specific needs.

References

Anthony, L. (2019). *AntConc Version 3.5.8* [Computer Software]. Waseda University. https://www.laurenceanthony.net/software

Benassi, C., Flores, S, Sobrero, M., Stefañuk, M, Benassi, M., & May, C. (2011). The impact of the complex noun phrase in reading comprehension of English scientific texts. *Revista de Ciencia y Tecnología*, *16*, 13–20.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Longman.

Bhatia, V. K. (2004). *Worlds of Written Discourse: A Genre-Based View*. Continuum International.

Bhatia, V. K. (2012). Critical reflections on genre analysis. *Ibérica*, *24*, 17–28.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.

Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: the influence of written language. *English Language and Linguistics*, *15*(2), 223–250.

Carrió Pastor, M. (2008). English complex noun phrase interpretation by Spanish readers. *Revista Española de Lingüistica Aplicada*, 27–44.

Dudley-Evans, T. (1994). Genre Analysis: An approach to text analysis for ESP. In M. Coulthard (Ed.), *Advances in Written Text Analysis* (pp. 219-228). Routledge.

Egbert, J, Biber, D., & Gray, B. (2022). *Designing and evaluating Language Corpora. A Practical Framework for Corpus Representativeness*. Cambridge University Press.

Flowerdew, L. (2002). Corpus-based analysis in EAP. In J. Flowerdew (Ed.). *Academic Discourse* (pp. 95–114). Longman.

Granath, S. (2009). Who benefits from learning how to use corpora? In K. Aijmer (Ed.) *Corpora and Language Teaching*. John Benjamins Publishing Co.

Halliday, M. A. K. (1993). Some grammatical problems in scientific English. In M. A. K. Halliday & J. R. Martin (Eds.), *Writing Science: Literacy and Discursive Power* (pp. 76–94). Routledge.

Halliday, M. A. K. (1998). Things and Relations: Regrammaticizing Experience as Technical Knowledge. In J. Martin & R. Veel (Eds.), *Reading Science: Critical and Functional Perspectives on Discourses of Science* (pp. 185–235). Routledge.

Halliday, M. A. K. (2004). *The Language of Science*. Continuum.

Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's Introduction to Functional Grammar* (Fourth Edition). Routledge.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.
Hyland, K. (2006). *English for academic purposes. An advanced resource book.* Routledge.
Marqueta-Gracia, B. (2017). Restrictions in the semantic interpretation of English and Spanish compounds. *Iberia: An International Journal of Theoretical Linguistics*, *9*, 1–35.
Nakov, P. (2008). Noun Compound Interpretation Using Paraphrasing Verbs: Feasibility Study. In: Dochev, D., Pistore, M., Traverso, P. (eds*) Artificial Intelligence: Methodology, Systems, and Applications. AIMSA 2008. Lecture Notes in Computer* Science, vol 5253. Springer. https://doi.org/10.1007/978-3-540-85776-1_10
Real Academia Española (2009). *Nueva Gramática de la Lengua Española*. Espasa Calpe. Asociación de Academias de la Lengua Española.
Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford University Press.
Sinclair, J. (2001). Preface. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small Corpus Studies and ELT* (pp. vi–xv). John Benjamins Publishing Co.
Sinclair, J. (2004). *How to use Corpora in Language Teaching*. John Benjamins Publishing Co.
Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.
Swales, J. (2004). *Research genres. Explorations and applications.* Cambridge University Press.
Wood, A. (2001). International scientific English: The language of research scientists around the world. In J. Flowerdew and M. Peacock (Eds.). *Research Perspectives on English for Academic Purposes* (pp. 71–83). Cambridge University Press.

Corpus references

Harvard Department of Chemistry and Chemical Biology (2012). Laboratory Safety Manual. https://www.chemistry.harvard.edu/files/chemistry/files/2012_1_9_safetymanual.pdf
Laboratory Safety Committee (LASAC). CSIR-Central Electrochemical Research Institute. (2022). Lab Safety Manual. https://cecri.res.in/portals/0/news_files/CECRILabSafetyManual.pdf
Macquarie University (2014). General Laboratory Safety Guidelines. https://bio.mq.edu.au/wp-content/uploads/2014/10/GENERAL-LABORATORY-SAFETY-GUIDELINES-V2-2014.pdf
McGill University (2021). Laboratory Safety Manual. https://www.mcgill.ca/ehs/files/ehs/laboratory_safety_manual_v_2.0.pdf
Sabancı University. Faculty of Engineering and Natural Sciences. (2023). Laboratory Safety Handbook. https://fens.sabanciuniv.edu/sites/fens.sabanciuniv.edu/files/2023-11/su-fens-laboratory-safety-handbook.pdf
Texas Tech University (2017). Laboratory Safety Manual. https://www.depts.ttu.edu/che/includes/LabSafetyManual.pdf
The City University of New York (2018). Laboratory Safety Manual. https://www.cuny.edu/wp-content/uploads/sites/4/media-assets/LAB-MANUAL-final-draft-for-committee.pdf

Toronto Metropolitan University. Department of Chemical Engineering. (2022). Laboratory Safety Manual. https://www.torontomu.ca/content/dam/chemical/forms-resources/general/Laboratory-Safety-Manual.pdf

Universidad de León. (2021). Laboratory safety and best practice handbook. https://www.unileon.es/files/2022-07/Manual_de_seguridad_GLP.pdf

University of California Riverside (2020). Laboratory Safety Manual. https://ehs.ucr.edu/document/laboratory-safety-manual