Exploring Terminological Collocations in Biomedical Texts

Rute Costa[1*], Margarida Ramos[1], Matilde Canelas[1], Ana Mouro [1]

[1] *Centro de Linguística da Universidade NOVA de Lisboa, Portugal*

Abstract

This paper aims to identify, describe, and analyse terminological collocations within the "HEREDITermCorpus," developed as part of the Hereditary Project. The study addresses several challenges, including the construction of the corpus, the definition and differentiation of terminological collocations from lexical collocations, and their distinction from polylexical terms. This paper provides a clearer understanding of these distinctions by focusing on the linguistic and conceptual dimensions of terminology, giving relevance to the double dimension of terminology.

Keywords

Terminological collocations, polylexical terms, biomedical corpus

1. Introduction

The goal of our paper is to identify, describe and analyse terminological collocations within the "HEREDITermCorpus" which has been built in the context of the Hereditary Project.[1] This task presents several challenges. In addition to creating the corpus and identifying terminological collocations, we must define what constitutes terminological collocations and differentiate them from lexical collocations. Another challenge involves distinguishing terminological collocations [10][4][7][3][19][20] from polylexical terms. These issues need to be addressed within both theoretical and methodological frameworks, as the distinctions we establish are essential for structuring and organising data in a terminological resource.

Terminological collocations are commonly used to refer to sets of units that frequently co-occur, either contiguously or non-contiguously, within a syntagmatic structure. In such collocations, one of the elements is a term belonging to a specific domain, used and recognised by members of a specialised community. Lexical collocation, as defined by Heid and Gouws [12], refers to "lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items in a direct syntactic relationship". A polylexical term, on the other hand, designates a concept characterised by its complex morphosyntactic structure.

Based on the aforementioned premises, Section 2 and Section 3 will outline the theoretical framework. Section 4 will focus on the constitution and treatment of the corpus, while in Section 5, we will analyse the results. The presentation will conclude with final remarks, followed by the bibliography.

4th International Conference on Multilingual digital terminology today. Design, representation formats and management systems (MDTT 2025) | 19-20 June 2025, Thessaloniki, Greece

[*] Corresponding author.

rute.costa@fcsh.unl.pt (R. Costa); mvramos@fcsh.unl.pt (M. Ramos); a55274@campus.fcsh.unl.pt (M. Canelas); a2023148474@campus.fcsh.unl.pt (A.Mouro).

0000-0002-3452-7228 (R. Costa); 0000-0001-7209-3806 (M. Ramos); 0009-0001-5010-0849 (M. Canelas); 0000-0003-0316-8987 (A. Mouro).

[1] https://hereditary-project.eu/

2. Lexical collocations: a short overview

Much has been written about lexical collocations, but comparatively less attention has been given to terminological collocations and how they differ from lexical collocations, as terminological collocations can sometimes be mistaken for polylexical terms.

The Anglo-Saxon school [8][17][1][2] and the German school [9][10] favour the term "collocation" to refer to groups of units that frequently co-occur contiguously along the syntagmatic axis. These authors stem from diverse theoretical backgrounds. Within Halliday's discourse analysis framework, particular emphasis is placed on sets of cohesive lexical units.

Sinclair [17] also addresses collocation in its double grammatical and lexical aspects, attributing an essential role to the statistical method in delimiting and more accurately describing collocation:

> *Collocation is the occurrence of two or more words within a short space of each other in the text (Sinclair, 1996, p. 170).*

Sinclair places particular emphasis on the lexical dimension of collocation from a lexicographical perspective, concentrating on its formal descriptions.

Collocation is therefore analysed using both lexical and statistical methodologies. These approaches enable the calculation and measurement of the density between the units forming the collocation, with distance understood as the length of the line segment between two points. In this context, lexical and statistical analyses serve as complementary tools. The frequency with which each unit of the collocation appears in a specific syntagmatic order serves as a key indicator for its identification, description, and classification.

Each unit within a collocation can hold varying levels of importance and value. The unit under analysis consists of a node and a collocate, with either unit capable of functioning as the node or the collocate, depending on the value attributed to it.

In 1989, Hausmann [9] defined collocation as:

> *[...] la combinaison caractéristique de deux mots dans une des structures suivantes : a) substantif + adjectif (épithète); b) substantif + verbe; c) verbe + substantif; d) verbe + adverbe; e) adjectif + adverbe; f) substantif + (prép.) + substantif (Hausmann, 1989, p. 1010).*

According to this lexicographer, a collocation is an oriented combination, meaning that the units comprising it hold different statuses. One unit serves as the core, responsible for the privileged lexical relationships it maintains with its immediate context. Consequently, it is crucial to distinguish between the base and the collocate ("Kollokator"), analogous to the notions of node and collocate. Identifying these elements is essential for linguistic description, lexicographic analysis of the collocation, and its learning and assimilation. The two components of a collocation have distinct roles, and the approach to each varies depending on whether the focus is on the base or the collocate.

A collocation is therefore composed of a base with syntactic and semantic autonomy and a collocate that adds a characteristic to the base without altering its identity.

According to the lexicologist Hoey [13], a collocation is a combination of words that frequently appear together and are primed for use in particular contexts due to repeated exposure in the language. Hoey emphasises that collocations are not purely syntactic structures but are shaped by the underlying priming effects of frequently co-occurring lexical items. These collocations reflect how language is structured and how speakers retrieve and produce language based on past experiences with similar lexical combinations.

In essence, collocations are seen as the result of cumulative exposure to patterns of word combinations, making them central to natural language use.

3. Terminological collocation vs polylexical terms[2]

Costa [6] explores the concept of terminological collocations, emphasising that they are combinations of lexical units that, although similar to terms, do not serve as designations of concepts. The author characterises terminological collocations as having a morphosyntactic structure in which one constituent is a term that influences another constituent, which may or may not be terminological. She emphasises that these collocations are non-designative, meaning they do not directly convey a concept in the same way that terms do.

The author highlights the significance of differentiating terminological collocations from polylexical terms, as they serve distinct functions within specialised discourse.

Terminological collocations are highly relevant in specialised discourse, but unlike terms, they are not defined solely by conceptual criteria. Instead, they are determined by morphological and syntactic criteria. In a terminological collocation, constituent X occurs in a privileged and frequent manner with constituent Y along a syntagmatic axis. The selection and lexical order are shared by a community of experts. The use of terminological collocations often reflects an individual's social and anthropological affiliation with a particular community.

Heid [11] partially aligns with Hausmann’s reasoning, but his perspective on collocation is distinct due to its terminological nature. He explicitly contends that collocation may refer to a term with characteristics and properties that differ from those traditionally associated with compound names. Heid [11] also discusses the polarity of collocations. A collocation consists of two lexemes, along with potential determinants, quantifiers, and prepositions — one lexeme serves as the determined element, while the other acts as the determinant. These concepts align with Sinclair’s “node” and “collocate” as well as Hausmann’s “Basis” and “Kollokator”.

Since his approach to collocation is terminological, one of the lexemes must necessarily be a term, and both lexemes may assume this status. For Heid, from a linguistic point of view, collocations are:

> *[...] a phenomenon of lexical combinatories: they involve the lexical, semantic, and syntactic properties of lexical items and their syntagmatic co-occurrence (Heid, 2001:788)*

When discussing syntactic properties, Heid draws a connection between collocation and compound words, as he believes the selection of components in compounds is influenced by the collocational perspective:

> *The choice of the components in such noun groups, like the choice of the components of the compounds, is often collocationally determined: there are clear combinatory preferences, often merely conventional, that in many cases go as far as the complete terminological “fixing” of the compounds and noun groups (Heid, 2001, p. 791).*

Heid acknowledges the challenge of distinguishing collocation from composition based solely on linguistic criteria, recognising that, from a theoretical standpoint, the boundary between them is very fine. From a terminological perspective, this distinction proves to be less practical, as in Terminology, it is the denomination[3] that underpins the identification of the linguistic reality, regardless of the label assigned to it:

> *From a terminological point of view, we may be more interested in whether the combination of term and collocate can be seen as the denomination of a new concept in its own right (Heid, 2001, p. 791).*

With respect to semantic properties, Heid draws on Mel'čuk's theory. Mel'čuk argues that speakers, in the full use of “parole”, use collocations to express generic meaning, and accordingly

[2] We prefer to use polylexical terms, whereas most authors use multiword terms or expressions, compounds, and complex terms.

[3] In ISO terminology, we use the term "designation" rather than "denomination."

describes collocations starting from the lexical functions that allow him to account for this generic character and which Heid expresses as follows:

> *In lexicography, examples of collocations are usually treated in terms of a given collocate with a given base being arbitrary phenomenon that must be memorized (Heid, 2001, p. 793).*

That phenomenon also occurs in specialised language, the difference being that the choice of collocate is usually the result of convention and not of free will.

We propose the term “terminological collocation” [4] to refer to a combination of at least two lexical units, where one must be a term, and the collective elements do not point to a single concept. This focus on conceptual indication differentiates it from the definition proposed by L'Homme [15], which distinguishes collocations from specialised lexical combinations. The former is described as “groupements de nature général”, while the latter are categorised as “groupements spécialisés”.

As shown by the definitions provided above, the distinction between polylexical terms—regardless of the terminology used—and terminological collocations is not immediately apparent.

To conclude, the differences that distinguish polylexical terms from terminological collocations are primarily conceptual. Polylexical terms, regardless their morhosyntactical structures, are designative, while terminological collocations are composed of a set of elements, where one of them exerts a morphosyntactic and/or semantic attraction over the other constituents, creating a cohesive structure within the collocation. In the case of a terminological collocation, one of its constituents is a term that in a given syntagmatic context attracts another constituent, which may be terminological or not — the whole of that morphosyntactic construction is a non-term. It is a non-term because, as a whole, it does not usually refer to a concept.

However, in situations where the relationship of dependence between the elements that constitute the construction — which began as a terminological collocation — solidifies, its status changes: the terminological collocation loses its status and becomes a terminological unit. The latter involves at least one constituent that is a term representing a concept specific to a particular domain and is used and shared by members of a domain-specific community.

Both terminological collocations and polylexical terms are regarded as specialised polylexical units. A polylexical unit “can be defined as a stable and recurrent sequence of lexemes that are perceived as an independent lexical unit by the speakers of a language” [18].

In the theoretical framework of the double dimension of terminology, terminological collocations are lexical combinations employed in domain-specific communication to convey knowledge, while polylexical terms denote concepts that belong to a particular conceptual system. Terminological collocations appear to be more variable than polylexical terms, as we will show in this paper through the corpus analysis.

Once the distinction between terminological collocations and polylexical terms has been addressed, we will focus on a specific type of terminological collocation: verbal terminological collocations. We will examine the morphosyntactic structure, where one of the constituents is a polylexical term (cf. Section 5), bridging the linguistic and conceptual analyses.

4. HEREDITermCorpus compilation[4]

For the compilation, annotation, and exploitation of corpora, we use Sketch Engine[5], a natural language processing (NLP) tool. This choice is motivated by its wide range of embedded tools, including both automatic and manual corpus annotation, as well as text-type analysis based on the statistical evaluation of metadata manually assigned to each text. The combination of annotation and

[4] https://zenodo.org/records/14628022

[5] https://www.sketchengine.eu/

advanced corpus analysis tools provides terminologists with a diverse set of approaches to working with corpora. The semi-automated terminology extraction method used in this study adopts a hybrid guiding criterion [14], integrating a combination of techniques and technologies.

The analysis of terminological data is based on a mixed approach [5], combining both conceptual and linguistic levels. This approach depends on the meaning encapsulated by relevant linguistic data (which is not necessarily statistical) or the concept-related information conveyed within a specific context. The patterns or unique features observed in texts serve as strong indicators of specialised knowledge.

Currently, the HEREDITermCorpus consists of 997 texts written in English – with a total of 5,137,643 tokens – authored by experts in the biomedical sciences. These texts, available in various editorial types (e.g., Clinical Trial, Meta-Analysis, Protocol, etc.), are catalogued with annotated metadata, specifically the “doc - Text Type” label, along other metainformation, within the NLP tool interface.

Among a set of verbs revealing a high number of occurrences in the corpus, we decided to work with two verbs in the infinitive tense, namely “increase” (Frequency=639) and “enhance” (Frequency=273).

5. Analysis of verbal terminological collocations derived from corpus

In the results from the HEREDITermCorpus, we identify verbs associated with either monolexical or polylexical terms. After reviewing the concordance, we chose to focus on verbs and applied the CQL query [lemma="XXX"&tag="VV"][tag="N."], where "XXX" represents the verb whose cotext we aim to analyse. We examined verbs such as “maintain”, “assist”, “affect”, and ultimately decided to explore the behaviour of “increase” and “enhance”. Both verbs are commonly used in general language and can be found in standard dictionaries:

1) while “increase”

used intransitively implies progressive growth in size, amount, or intensity;

used transitively it may imply simple not necessarily progressive addition, [6]

2) “enhance” is described as “to improve the quality, amount, or strength of something”.[7]

The following expressions are associated with the infinitive verb "increase":

[1.1] increase blood-brain-barrier
[1.2] increase BBB
[1.3] increase BBB permeability
[1.4] increase blood-brain-barrier permeability
[1.5] increase permeability of the blood-brain barrier
[1.6] increase gut microbiome diversity
[1.7] increase gut microbiota richness and diversity
[1.8] increase susceptibility to pathogenic infiltration

For the verb “enhance” we found the following expressions

[2.1] enhance gut barrier function
[2.2] enhance extracellular signal-regulated kinase phosphorylation
[2.3] enhance neurodegeneration

We consider these nine expressions to be terminological collocations because the verb governs either a monolexical or a polylexical term. As an example,[8] we present the analyses of [1.4]:

[[increase]v + [BBB]sigle = polylexical term]N= term + [permeability] N= term]] = terminological collocation

[6] https://www.merriam-webster.com/thesaurus/increase

[7] https://dictionary.cambridge.org/dictionary/english/enhance

[8] If the paper is accepted, we will analyse all the examples in greater depth using scientific arguments, as we do not have enough space to delve further into the analysis.

It is interesting to note the term variation present in our corpus, as illustrated below, being the example 1.1. and 1.2 and 1.3 and 1.4:

[1.1] increase blood-brain-barrier
[1.2] increase BBB
[1.3] increase BBB permeability
[1.4] increase blood-brain-barrier permeability
[1.5] increase permeability of the blood-brain barrier

There is variation in the order of the constituents, referring to the morphosyntactic structure, but the designated concept remains the same.

The verb "increase" is commonly used in scientific and medical contexts to discuss changes in quantities, effects, or conditions. "Blood-brain barrier" and "blood-brain barrier permeability" are two terms that refer to well-established, context-specific concepts in medicine.

Alongside the term "blood-brain barrier permeability," we identified other synonymous collocates such as "augment" and "elevate". "Make" is also a collocate in the collocation "make the blood-brain barrier permeable". However, this collocation falls into a different category, as "make" functions as a causative verb, whereas "increase" denotes an action that leads to a change in quantity or degree. Therefore, "increase blood-brain barrier permeability" is not a synonymous collocation of "make the blood-brain barrier permeable".

6. Final remarks

In our paper, we discuss the concept of terminological collocations, distinguishing them from both lexical collocations and polylexical terms. To clarify these two concepts, we focus on the dual dimension of terminology, defining terminological collocations by considering both their morphosyntactic structure and the designated concept that distinguishes them from lexical collocations and polylexical terms. We describe the construction of our corpus, and the methodology used to identify verbs that help recognise terminological collocations. These collocations will be analysed as semi-stable structures, and we will examine, observe, and describe their variations and synonymous expressions.

Acknowledgements

This work is supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. This work is also part of the initiatives carried out the by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UID/03213 – Linguistics Research Centre of NOVA University Lisbon (CLUNL).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 for grammar and spelling checks. The authors have subsequently reviewed and edited the content and take full responsibility for the publication's final version.

References

[1] Benson, M., Benson E., & Ilson, R. (1986) Lexicographic Description of English, John Benjamins.
[2] Benson, M., Benson E., & Ilson, R. (1997). The BBI Dictionary of English Word Combinations (pp. VII – XXXIX.) John Benjamins.
[3] Castro, M. B., Martínez, S. M., & Faber, Pamela (2014). Verb collocations and phraseology in EcoLexicon. Yearbook of Phraseology, vol. 5, no. 1, 57-94.

[4] Costa, R. (2001). Pressupostos teóricos e metodológicos para a extração automática de unidades terminológicas multilexémicas. (Doctoral dissertation). Universidade NOVA de Lisboa.

[5] Costa, R. (2006). Plurality of Theoretical Approaches to Terminology. In Modern Approaches to Terminological Theories and Applications (Coll. Linguistic Insights. Studies in Language and Communication). Peter Lang.

[6] Costa, R. (2017). Curricular unit report: Theories in Terminology Master's in Language Sciences specialisation in Terminology and Specialised Information Management. (2nd Cycle). Report submitted for admission to Provas de Agregação at Universidade NOVA de Lisboa, for the Subject Area of Linguistics - Lexicology, Lexicography, and Terminology.

[7] Costa, R. & Silva, R. (2004). The verb in the terminological collocations. Contribution to the development of a morphological analyser MorphoComp. Proceedings of Language Resources and Evaluation – LREC 2004. ELRAHalliday, M. & Hasan, R. 1976. Cohesion in English, London, Longman.

[8] Halliday, M. & Hasan, R. 1976. Cohesion in English, London, Longman.

[9] Hausmann, F. J. (1989). Le dictionnaire des collocations. In F. J. Hausmann, O. Reichmann, H. E. Wiegand, & L. Zgusta (Eds.), Wörterbücher: Ein internationales Handbuch für Lexikographie (pp. 1010–1019). Walter de Gruyter.

[10] Heid, U. (1999) Extracting terminologically relevant collocations from German technical texts. 5th International Congress on Terminology and Knowledge Engineering. (TKE '99).

[11] Heid, U. (2001). Colocations in Sublanguage Texts: extraction form Corpora, In S.E. Wright & G. Budin (Eds). Handbook of Terminology management: Volume 2: Application-Oriented terminology Management, Vol. 2, (pp 788-808). John Benjamins.

[12] Heid, U., & Gouws, R. H. (2006). A model for a multifunctional dictionary of collocations. Proceedings of the 12th EURALEX International Congress, 979–988.

[13] Hoey Michael (2005) Lexical Priming: A New Theory of Words and Language. Psychology Press. Routledge. Taylor & Francis Group.

[14] International Organization for Standardization. (2025). Management of terminology resources — Terminology extraction. (ISO/FDIS Standard No. 5078) [forthcoming]

[15] L'Homme, M.-C. & Meynard I. (1998). Le point d'accès aux combinaisons lexicales spécialisée : présentation de deux modèles Informatiques. Traduction, Terminologie, Rédaction. Association canadienne de traductologie. Volume 11 (1).

[16] Silva, R., Costa, R. & Ferreira, F. (2004). Entre langue générale et langue de spécialité : une question de collocations. Études de Linguistique Appliquée, Paris : Klincksieck, 347-35.

[17] Sinclair, J. (1996) EAGLES: Preliminary Recommendations on Corpus Typology (EAG -- TCWG --CTYP/P), Version of May (p. 25) Retrieved from em http://www.ilc.pi.cnr.it.

[18] Tasovac T., Salgado A, & Costa R. (2020). Encoding polylexical units with TEI Lex-o: A case study. Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research, 8(2), 28-57.

[19] Ramos, M., Costa, R. & Roche, C. (2019). Dealing with specialized co-text in text mining: The verbal terminological collocations. In Terminologica (2019) (pp.339-362) Grenoble : Presses Universitaires Savoie Mont Blanc.

[20] Vezzani, F. (2023) Vers une méthodologie pour l'extraction et la classification automatiques des collocations terminologiques verbales en langue médicale. Phraséologie et terminologie, (257-275) Walter de Gruyter GmbH & Co KG.