INNOVATION IN LANGUAGE LEARNING International Conference

Compiling and Exploring Health Sciences Corpora for Multilingual Language Learning

Teresa Alegre[1], Katrin Herget[2], João Paulo Silvestre[3]

University of Aveiro | Centre for Languages, Literatures and Cultures, Portugal[1]
University of Aveiro | Centre for Languages, Literatures and Cultures, Portugal[2]
University of Aveiro | Centre for Languages, Literatures and Cultures, Portugal[3]

Abstract

Effective multilingual communication is crucial in the globalized healthcare sector. This study examines the use of health sciences corpora for LSP learning and research. By implementing corpus tools such as Sketch Engine [1] and AntConc [2], Applied Linguistics students become acquainted with specialized tools. Our research implements a data-driven learning (DDL) methodology [3] for corpus building and analysis at postgraduate level, focusing on health sciences terminology, discourse, and translation. By presenting students with research queries related to medical communication, we aim to create a dynamic learning environment that allows for the exploration of language usage across various healthcare contexts and enables students to develop transversal skills and competencies. Students learn to distinguish different textual genres relevant to health sciences, comprising diverse communication levels, exploring language complexity by studying the occurrence of different linguistic patterns across languages. They gain hands-on experience in compiling and exploring corpora, including data collection, cleaning, and analysis. The approach facilitates cross-linguistic analysis, allowing students to compare and contrast medical language use in different languages. This study not only enhances students' understanding of medical language across different communication levels but also equips them with valuable skills in corpus linguistics and data analysis.

Keywords: *LSP (Languages for Specific Purposes), Corpus Linguistics, Data-driven Learning (DDL), Health Sciences*

1. Introduction

In the globalized healthcare landscape, effective communication across languages is essential for the dissemination of medical knowledge and the advancement of health practices. Medical discourse encompasses a broad range of text types, each serving different purposes and audiences. The knowledge of such discourse is not only relevant for health professionals, but also for translators, technical communicators, and applied linguists in general, who have to deal with various textual genres in different contexts.

Based on our teaching experience utilizing authentic textual material in specialized language courses, exploring corpora is indispensable for both language learning and research. Postgraduate students in Applied Linguistics can benefit significantly from engaging with specialized corpora, which allow them to interact with real-world data meaningfully. By analyzing patterns, structures, and context-specific language use, students can develop a deeper understanding of language variation and use across different fields. Furthermore, working with corpora equips students with valuable skills in data-driven analysis, fostering critical thinking. Central to this research is the Data-Driven Learning (DDL) methodology, which encourages direct engagement with authentic language data. This approach not only enhances students' understanding of medical language but also develops critical transversal skills such as analytical thinking, data interpretation, and cross-linguistic comparison.

This hands-on approach is crucial for developing a comprehensive understanding of the complexities of medical language, as it involves all stages of corpus work, from data collection, processing to detailed linguistic analysis. Through this process, students learn to distinguish between different textual genres within the health sciences and explore how these genres operate at various levels of communication, from highly specialized research articles to more accessible patient information leaflets.

In this article we focus on a classroom project proposal that can be carried out by Applied Linguistics students and adapted to different learning scenarios across various domains. The proposal is flexible, allowing for modifications in the type of data used, such as multimodal data related to internal or

external communication, different textual genres, or varying communication partners. This adaptability enables students to adapt the research to their specific interests while gaining valuable experience in analyzing diverse linguistic contexts.

2. Navigating Medical Textual Genres: The Complexity of Medical Communication

Effectively disseminating medical knowledge across diverse linguistic and cultural contexts requires an understanding of the varied nature of medical texts. As Karwacka [4] points out, in relation to medical translation, typical textual genres include "popularizations, such as textbooks for medical students, popular science books on medicine, but also research papers, conference proceedings, case studies, case histories, discharge summaries, reports and relatively simple texts for patients: information leaflets, consent forms, brochures" (p. 272). This wide range of genres highlights the inherent complexity of medical discourse and the need to address the specific characteristics of each textual genre to ensure accurate and effective communication of medical information.

Montalt and González Davies [5] identify three fundamental types of medical genres: instructional, expository, and argumentative. The authors further categorize medical genres based on their broader social functions, such as disease prevention (e.g., press releases), communicating new discoveries (e.g., newspaper articles), teaching and learning in health studies (e.g., textbooks, encyclopaedias), and promoting health products (e.g., leaflets, promotional material) (pp. 57-58). This classification reflects the heterogeneous nature of medical discourse, both in internal and external communication.

Given this complexity, it is essential to collect and compile texts from the medical field to better analyze specific patterns and structures. This can be accomplished through the DDL (Data-Driven Learning) approach, which will be discussed in the next section.

3. Data-Driven Learning (DDL) in Health Sciences

Data-Driven Learning (DDL) methodologies hold significant value in the training of post-graduate students in Applied Linguistic. Susam-Saraeva and Spišiaková [6] in The Routledge Handbook of Translation and Health, emphasize the increasing application of "technology and corpus-based approaches in supporting and improving medical translations" (p. 6).

According to Boulton and Tyne [7], DDL enhances both cognitive and metacognitive skills, increases sensitivity to authentic language use, offers an interactive approach to constructivist discovery learning, and fosters motivation through individualized learning experiences. They also highlight that DDL promotes the development of reusable and transferable skills, supports lifelong learning autonomy, and aligns with contemporary theories of second language acquisition.

Boulton and Cobb [8] further elaborate that “DDL involves learners consulting language data themselves, integrating concepts of learner autonomy, induction, exemplar-based learning, and constructivism. This approach encourages learners to discover linguistic patterns independently (with varying levels of guidance), rather than being passively taught pre-digested rules” (p. 349).

In a complementary view, Gilquin and Granger [3] broadly define DDL as "using the tools and techniques of corpus linguistics for pedagogical purposes" (p. 359). This definition underscores the value of using authentic language data to support and enrich learning processes. They emphasize that corpus tools are instrumental in advancing translation teaching and practice by facilitating access to, and analysis of, authentic language data. The integration of technology in DDL refines linguistic skills and enhances language education through data-driven insights.

4. Practical Implementation of DDL: Study Design Proposal

The practical application of DDL can be significantly enhanced by creating ad-hoc corpora in the field of Health Studies. This study aims to engage post graduate students in Applied Linguistics in the creation and analysis of such corpora. These web-based corpora will be constructed to address specific communicative contexts, enabling a detailed exploration of language use across different languages. By compiling ad-hoc corpora, students will not only acquire practical skills in corpus building but also deepen their understanding of the importance of context and audience in specialized communication.

The study will focus on both internal communication, such as medical reports and case studies, and external communication, such as patient brochures and consent forms. This will allow students to compile comparable ad-hoc corpora that reflect the diverse linguistic demands of the healthcare field.

The project will use the corpus tool Sketch Engine to compile comparable corpora in different languages.
The following sections will outline the stages of this study design, using a German-Portuguese case study as an example:

Stage 1 - Preparation: Contextualization and Pre-Corpus Building

In this case study, we focus on internal communication within the Health Sciences domain. However, in other research scenarios, external communication may also be relevant for analysis depending on the specific objectives of the study.
For the corpus-building process, websites were selected based on the following criteria: i) the websites belong to the Health Sciences domain; ii) the texts target healthcare professionals; iii) the texts are authored by health experts; iv) the texts are written in German.

a) Students discuss the advantages of different types of corpora and the process of creating ad-hoc corpora. For example, comparable corpora can provide insight into domain-specific terminology and textual conventions, while parallel corpora provide equivalents in multiple languages:

b) The importance of ad-hoc corpora for Applied Linguistics is emphasized, highlighting how specific corpora can be used to analyze lexico-grammatical patterns, language varieties, and communication levels.

Stage 2 - Project execution

Compilation of Ad-Hoc Corpora for Internal Communication

a) Selection of adequate websites;

b) Using Sketch Engine, an ad-hoc corpus is compiled from open-access papers, such as those sourced from *Deutsches Ärzteblatt* (www.aerzteblatt.de) and *Revista Portuguesa de Oncologia* (https://rponcologia.com). The WebBootCat tool in Sketch Engine facilitates this process by extracting relevant texts. Four specific keywords were used for this extraction process, both in German and Portuguese: *Medizin*/*medicina*, *Lebenswissenschaften*/*ciências da vida*, *Gesundheit*/*saúde*, and *Karzinom*/*cancro*.

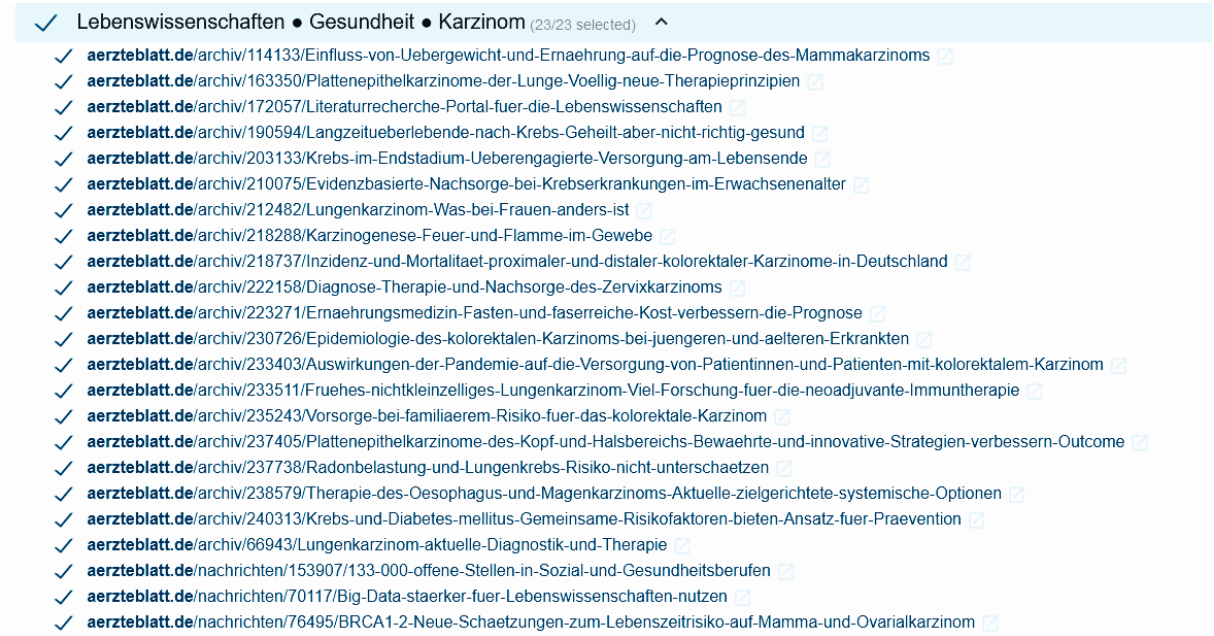


Fig. 1. Corpus compilation process and presentation of corpus statistics

c) A critical reflection on the adequacy and reliability of the selected websites is conducted.

Corpus Compilation and Concordance Analysis Using Sketch Engine

a) The corpus is built using reliable texts, following criteria such as domain, authorship, target audience, and language variety.

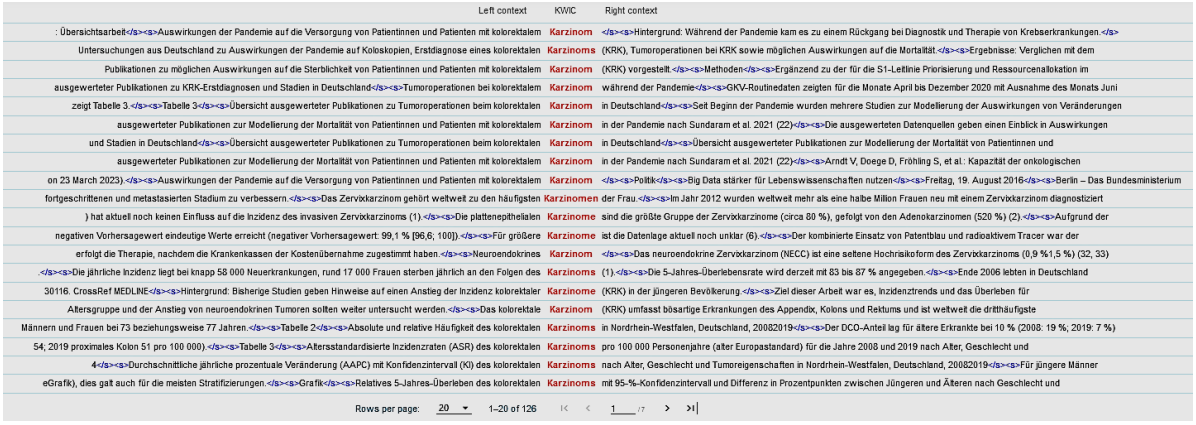
b) Keywords are extracted and a concordance analysis is performed. Table 1 presents a comparison of keyword frequency between the focus and reference corpora, with a focus on internal medical communication.

Item	Frequency (focus)	Frequency (reference)	Relative frequency (focus)	Relative frequency (reference)	Score
Lungenkarzinom	103	3111	1631,45056	0,14815	1421,82
Karzinom	126	19757	1995,755	0,94083	1028,82
Koloskopie	78	6423	1235,46741	0,30586	946,858
Inzidenz	84	19388	1330,50342	0,92326	692,317
kleinzellig	44	2493	696,93036	0,11872	623,867
Radiochemotherapie	35	1533	554,3764	0,073	517,591
Adenokarzinom	36	2524	570,21576	0,12019	509,926
KRK	38	4043	601,89441	0,19253	505,56
S3-Leitlinie	38	4710	601,89441	0,22429	492,444
AAPC	31	74	491,0191	0,00352	490,291
Zervixkarzinom	30	2292	475,17978	0,10914	429,322
Kolon	30	2968	475,17978	0,14134	417,213
Mammakarzinom	38	9979	601,89441	0,4752	408,687
Mortalität	46	21421	728,60901	1,02007	361,181
Plattenepithelkarzinom	27	4901	427,6618	0,23339	347,549
Gesamtüberleben	25	3749	395,98315	0,17853	336,847
Krebsregisterdaten	21	435	332,62585	0,02071	326,855
Chemotherapie	87	69903	1378,02136	3,32878	318,571
Nachsorge	51	39043	807,8056	1,85923	282,876

Table 1. Analysis of keyword frequency and significance within the corpus

The table indicates that the focus dataset emphasizes specialized oncology-related terms, particularly those related to lung cancer (*Lungenkarzinom*), cancer diagnostics (*Koloskopie*), and specific cancer types.

Figure 2 represents the contextual usage of the keyword *Karzinom* within the focus corpus. A concordance analysis displays how a keyword appears in various contexts, which helps to understand its semantic and syntactic roles in medical discourse.



Left context	KWIC	Right context
: Übersichtsarbeit</s><s>Auswirkungen der Pandemie auf die Versorgung von Patientinnen und Patienten mit kolorektalem	Karzinom	</s><s>Hintergrund: Während der Pandemie kam es zu einem Rückgang bei Diagnostik und Therapie von Krebserkrankungen.</s>
Untersuchungen aus Deutschland zu Auswirkungen der Pandemie auf Koloskopien, Erstdiagnose eines kolorektalen	Karzinoms	(KRK), Tumoroperationen bei KRK sowie möglichen Auswirkungen auf die Mortalität.</s><s>Ergebnisse: Verglichen mit dem
Publikationen zu möglichen Auswirkungen auf die Sterblichkeit von Patientinnen und Patienten mit kolorektalem	Karzinom	(KRK) vorgestellt.</s><s>Methoden</s><s>Ergänzend zu der für die S1-Leitlinie Priorisierung und Ressourcenallokation im
ausgewerteter Publikationen zu KRK-Erstdiagnosen und Stadien in Deutschland</s><s>Tumoroperationen bei kolorektalem	Karzinom	während der Pandemie</s><s>GKV-Routinedaten zeigten für die Monate April bis Dezember 2020 mit Ausnahme des Monats Juni
zeigt Tabelle 3.</s><s>Tabelle 3</s><s>Übersicht ausgewerteter Publikationen zu Tumoroperationen beim kolorektalen	Karzinom	in Deutschland</s><s>Seit Beginn der Pandemie wurden mehrere Studien zur Modellierung der Auswirkungen von Veränderungen
ausgewerteter Publikationen zur Modellierung der Mortalität von Patientinnen und Patienten mit kolorektalem	Karzinom	in der Pandemie nach Sundaram et al. 2021 (22)</s><s>Die ausgewerteten Datenquellen geben einen Einblick in Auswirkungen
und Stadien in Deutschland</s><s>Übersicht ausgewerteter Publikationen zu Tumoroperationen beim kolorektalen	Karzinom	in Deutschland</s><s>Übersicht ausgewerteter Publikationen zur Modellierung der Mortalität von Patientinnen und
ausgewerteter Publikationen zur Modellierung der Mortalität von Patientinnen und Patienten mit kolorektalem	Karzinom	in der Pandemie nach Sundaram et al. 2021 (22)</s><s>Arndt V, Doege D, Fröhling S, et al.: Kapazität der onkologischen
on 23 March 2023).</s><s>Auswirkungen der Pandemie auf die Versorgung von Patientinnen und Patienten mit kolorektalem	Karzinom	</s><s>Politik</s><s>Big Data stärker für Lebenswissenschaften nutzen</s><s>Freitag, 19. August 2016</s><s>Berlin – Das Bundesministerium
fortgeschrittenen und metastasierten Stadium zu verbessern.</s><s>Das Zervixkarzinom gehört weltweit zu den häufigsten	Karzinomen	der Frau.</s><s>Im Jahr 2012 wurden weltweit mehr als eine halbe Million Frauen neu mit einem Zervixkarzinom diagnostiziert
) hat aktuell noch keinen Einfluss auf die Inzidenz des invasiven Zervixkarzinoms (1).</s><s>Die plattenepithelialen	Karzinome	sind die größte Gruppe der Zervixkarzinome (circa 80 %), gefolgt von den Adenokarzinomen (520 %) (2).</s><s>Aufgrund der
negativen Vorhersagewert eindeutige Werte erreicht (negativer Vorhersagewert: 99,1 % [96,6; 100]).</s><s>Für größere	Karzinome	ist die Datenlage aktuell noch unklar (6).</s><s>Der kombinierte Einsatz von Patentblau und radioaktivem Tracer war der
erfolgt die Therapie, nachdem die Krankenkassen der Kostenübernahme zugestimmt haben.</s><s>Neuroendokrines	Karzinom	</s><s>Das neuroendokrine Zervixkarzinom (NECC) ist eine seltene Hochrisikoform des Zervixkarzinoms (0,9 %1,5 %) (32, 33)
.</s><s>Die jährliche Inzidenz liegt bei knapp 58 000 Neuerkrankungen, rund 17 000 Frauen sterben jährlich an den Folgen des	Karzinoms	(1).</s><s>Die 5-Jahres-Überlebensrate wird derzeit mit 83 bis 87 % angegeben.</s><s>Ende 2006 lebten in Deutschland
30116. CrossRef MEDLINE</s><s>Hintergrund: Bisherige Studien geben Hinweise auf einen Anstieg der Inzidenz kolorektaler	Karzinome	(KRK) in der jüngeren Bevölkerung.</s><s>Ziel dieser Arbeit war es, Inzidenztrends und das Überleben für
Altersgruppe und der Anstieg von neuroendokrinen Tumoren sollten weiter untersucht werden.</s><s>Das kolorektale	Karzinom	(KRK) umfasst bösartige Erkrankungen des Appendix, Kolons und Rektums und ist weltweit die dritthäufigste
Männern und Frauen bei 73 beziehungsweise 77 Jahren.</s><s>Tabelle 2</s><s>Absolute und relative Häufigkeit des kolorektalen	Karzinoms	in Nordrhein-Westfalen, Deutschland, 20082019</s><s>Der DCO-Anteil lag für ältere Erkrankte bei 10 % (2008: 19 %; 2019: 7 %)
54; 2019 proximales Kolon 51 pro 100 000).</s><s>Tabelle 3</s><s>Altersstandardisierte Inzidenzraten (ASR) des kolorektalen	Karzinoms	pro 100 000 Personenjahre (alter Europastandard) für die Jahre 2008 und 2019 nach Alter, Geschlecht und
4</s><s>Durchschnittliche jährliche prozentuale Veränderung (AAPC) mit Konfidenzintervall (KI) des kolorektalen	Karzinoms	nach Alter, Geschlecht und Tumoreigenschaften in Nordrhein-Westfalen, Deutschland, 20082019</s><s>Für jüngere Männer
eGrafik), dies gilt auch für die meisten Stratifizierungen.</s><s>Grafik</s><s>Relatives 5-Jahres-Überleben des kolorektalen	Karzinoms	mit 95-%-Konfidenzintervall und Differenz in Prozentpunkten zwischen Jüngeren und Älteren nach Geschlecht und

Rows per page: 20 1–20 of 126 1 /?

Fig. 2. Results of concordance search for the keyword *Karzinom*

Similarly, Table 2 presents the most frequent keywords in the Portuguese corpus.

Item	Frequency	Relative frequency
doente	604	5432,92496
tratamento	489	4398,51044
cancro	438	3939,77009
oncologia	249	2239,73231
risco	242	2176,76795
estudo	215	1933,90541
doença	209	1879,93596
trastuzumab	203	1825,9665
tumor	164	1475,16506
grupo	136	1223,30761
sobrevivente	134	1205,31779
quimioterapia	121	1088,38397
terapêutica	116	1043,40943
idade	98	881,50107
cirurgia	98	881,50107
diagnóstico	97	872,50616
revisão	97	872,50616
mulher	95	854,51634

Table 2. Frequent keywords in the Portuguese ad-hoc corpus

A key observation from the table is that the most frequent words, such as *doente* (*patient*), *tratamento* (*treatment*), and *cancro* (*cancer*), though important in health sciences, are not exclusive to medical terminology. Words like *risco* (*risk*), *estudo* (*study*), and *grupo* (*group*) also exhibit a wide semantic range. These terms are frequently used in various contexts, not limited to health or medical sciences. For instance, *risco* can be associated with many fields, such as finance or environmental studies.
On the other hand, terms such as *trastuzumab* and *quimioterapia* (*chemotherapy*) are more closely tied to the medical domain. These terms are integral to oncological treatment and specialized medical discourse, representing concepts and drugs specific to the field of oncology.

c) The Word Sketch feature in Sketch Engine is used. Students analyze the most frequent noun modifiers associated with *Karzinom*, such as descriptors that specify the type, location, or cellular characteristics of the carcinoma.

Fig. 3. Frequent modifiers of *Karzinom* in the German Medizin corpus

d) Creating a Wordcloud: Using the distributional thesaurus in Sketch Engine, students can identify words that frequently co-occur with *Karzinom* / *carcinoma*, analyzing collocate frequency, thematic associations, contextual usage, and other relevant aspects.

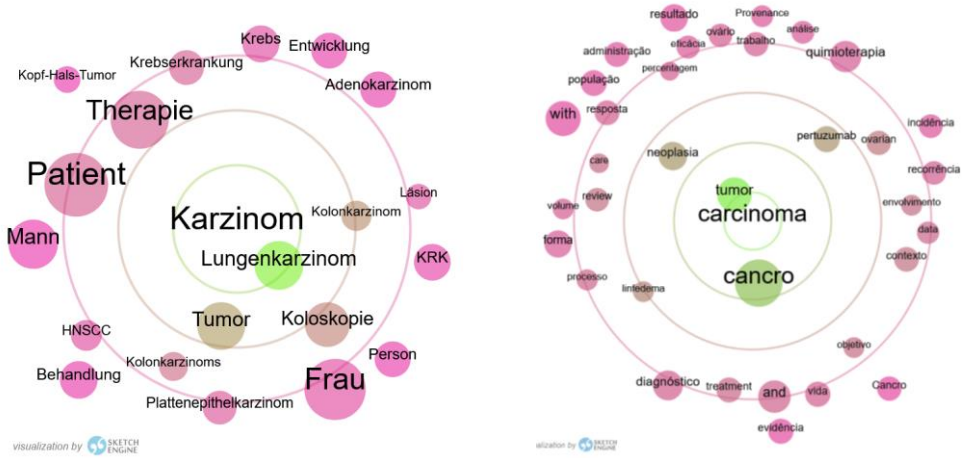


Fig. 4. Wordclouds with frequent co-occurrences of *Karzinom* and *carcinoma*

Stage 3 - Evaluation

a) The project is evaluated through a portfolio that includes the results of the concordance search, keyword extraction, and analysis, along with reflections on the value of corpus-based research in Applied Linguistics.

b) Students complete an online questionnaire to provide feedback on their experience with data-driven project work, offering insights into their perceptions of the process.

5. Conclusion

This study proposal outlines a framework for implementing DDL methodologies within Applied Linguistics, with a specific emphasis on Health Studies.
The primary objective is to actively engage post-graduate students in the creation and analysis of ad-hoc corpora using the corpus tool Sketch Engine. By focusing on health-related texts, such as those from the field of internal medical communication, this proposal demonstrates how DDL can support the development of students' critical cognitive and metacognitive skills, enhance their sensitivity to authentic language use, and provide an interactive and constructivist learning environment. The frequent occurrence of both general terms and specialized terminology in both corpora indicates that students must become skilled in distinguishing between every day and specialized language. Those

with backgrounds in health and medical sciences, in particular, play a key role in interpreting the nuances of these terms, as they are better equipped to recognize when a word functions as part of specialized lexicon within specific contexts. This analysis allows students, particularly in applied linguistics or medical studies, to examine how common and specialized terms interact within different communicative settings, thereby enhancing their understanding of medical terminology. Moreover, this aligns with the broader goal of the study: to develop transversal skills in language analysis, data interpretation, and corpus-based learning.

The proposed study design sets the stage for future research and practical application of DDL methodologies in the health studies domain. By integrating corpus-based approaches into the study of medical texts, this framework offers a model for enhancing both linguistic analysis and pedagogical practices within specialized fields.

REFERENCES

[1] Kilgarriff, A., Baisa, V., Bušta, J. et al. (2014). The Sketch Engine: ten years on. *Lexicography ASIALEX* 1, 7–36. https://doi.org/10.1007/s40607-014-0009-9.

[2] Anthony, L. (2022). *AntConc* (Version 4.2.0) [Computer Software]. Waseda University.

[3] Gilquin, G., & Granger, S. (2010). How can DDL be used in language teaching? In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 359-370). Routledge.

[4] Karwacka, W. (2015). Medical translation. In Ł. Bogucki, S. Goźdź-Roszkowski, P. Stalmaszczyk (Eds.) *Ways to Translation*. (pp: 271-298). Wydawnictwo Uniwersytetu Łódzkiego.

[5] Montalt, V. & González Davies, M. (2014). *Medical Translation Step by Step. Learning by Drafting*. Routledge

[6] Susam-Saraeva, S., & Spišiaková, E. (2021) (Eds.). *The Routledge Handbook of Translation and Health*. Routledge.

[7] Boulton, A., & Tyne, H. (2013). Corpus linguistics and data-driven learning: a critical overview. *Bulletin suisse de Linguistique appliquée* (Neuchâtel: Institut de Linguistique de l'Université) 97. 97–118. https://hal.archives-ouvertes.fr/hal-01208263.

[8] Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning* 67(2). 348–393. https://doi.org/10.1111/lang.12224.