10th International Conference on Higher Education Advances (HEAd'24)
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/HEAd24.2024.17159

Exploring the collocation database CCDB in the LSP classroom

Katrin Herget
Department of Languages and Cultures/CLLC, University of Aveiro, Portugal.

How to cite: Herget, K. 2024. Exploring the collocation database CCDB in the LSP classroom. In: 10th International Conference on Higher Education Advances (HEAd'24). Valencia, 18-21 June 2024. https://doi.org/10.4995/HEAd24.2024.17159

Abstract

This paper presents a proposal for the integration of the collocation database CCDB in German as Language for Specific Purposes (LSP) classes at the Master's level. The primary focus of this study is to outline a conceptual framework for its implementation in the classroom setting. The proposed approach emphasizes the importance of incorporating data-driven learning (DDL) methodologies to equip students with the necessary skills to extract insights from large language datasets. The use of corpora in language teaching and learning has been extensively researched and debated over the past few decades. The increasing integration of Information and Communication Technology (ICT) in LSP classes has significantly contributed to advocating for the adoption of corpora in language instruction. The German database CCDB (https://corpora.ids-mannheim.de/ccdb/), based on a 2.2 billion-word subset of the German Reference Corpus, is an invaluable resource for exploring a large-scale corpus through data-driven methods. It comprises rich collocation profiles that can be utilized in LSP classes to enhance students' linguistic competences. This study presents a teaching proposal integrated into the 'Languages and Business Relations' Master's course at the University of Aveiro.

Keywords: *Languages for Specific Purposes; corpus linguistics; data-driven learning; collocation database CCDB; German.*

1. Introduction

The demand for expertise in multiple languages across various fields has increased significantly. This has led to the need for language specialists with specialized knowledge in their respective domains (Gollin-Kies, Hall, Moore, 2016, p. 35). Multilingual communication has assumed a crucial role in industries that require language skills for specific purposes, facilitating information exchange in diverse linguistic environments. As the professional environment becomes increasingly complex and career paths continue to diversify, individuals are finding themselves in greater demand to adapt to changing requirements and specialize accordingly. Considering the diverse landscape of ICT today, attention will be directed toward the

exploration of online corpora in Language for Specific Purposes (LSP) classes, due to their growing relevance in Natural Language Processing. By adopting a data-driven learning (DDL) approach, even novice students in corpus work may gain useful insights and become familiar with practices for exploring large amounts of language data. In this paper, we present a proposal for the integration of online corpora, specifically the collocation database (CCDB), into LSP classes. While the practical implementation of this proposal is yet to be conducted, we aim to outline a conceptual framework for its potential application in the classroom setting. Drawing on previous studies (Herget, 2018, 2020) that have explored project-based learning scenarios involving approaches to implementing the Translation Management System Phrase (form. Memsource) in LSP classes or conducting a localization project with a group of Master students in "Languages and Business Relations" at the University of Aveiro, our aim is to elucidate how the CCDB could be effectively utilized in the LSP classroom. Particularly in the context of LSP instruction, there exists a pressing need for continuously updated methods and strategies for the organization and management of knowledge due to the evolving complexity of professional profiles.

2. Languages for Specific Purposes – working definition and current challenges

LSP teaching focuses on developing language skills for specific professional contexts, such as business, law, medicine, and so on. In an era of immense technological potentialities for the educational field, LSP teaching offers personalized learning experiences that adapt to individual needs and accelerate language acquisition like never before. Firstly, we should clarify what LSP stands for and then discuss some major trends impacting the study of LSP. According to Basturkmen & Elder (2004, p. 672) "LSP is generally used to refer to the teaching and research of language in relation to the communicative needs of speakers of a second language in facing a particular workplace, academic, or professional context". In terms of a working definition, in the following, we perceive LSP from a broad perspective, as outlined by Koskela & Isohella, (2018, p. 101) rather than focusing on a specific language (e.g. GSP). According to the authors, a broad approach "can be applied to education offered on any language or to multilingual education". Following Brandt (2006, p. 14), *specialness*, or the adjective *special* or *specific* refers to different levels of the communicative process, involving:

- context (domain, subject-matter, setting)
- discourse partners (expertise and status)
- message (specific functions)
- channel and medium (discourse path and text type)
- code (syntactical, morphological and lexical features

Gollin-Kies et al. (2015) summarize key trends in learning, teaching and researching LSP, which include inter & cross-cultural communication, computer-based language research and

independent learning, among others. Each of these categories intersects with one another, reflecting the confluential dynamics within the domain of LSP.

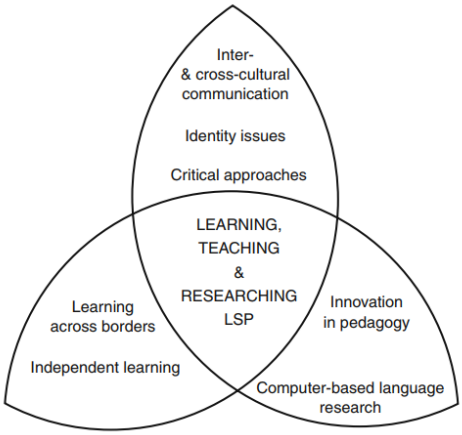

Figure 1: Key topics in LSP research, learning and teaching (Gollin-Kies et al., 2015, p. 51)

3. Self-directed Language Learning – a DDL approach

The integration of ICT in LSP classes has transformed language education, enhancing interactive learning and providing students with valuable digital skills for a specific domain. According to Arnó-Macià (2015, p. 5), "Developments in IT have influenced LSP, not only in facilitating access to specialized discourse and communication, but also as a result of the evolution of technology as a language learning tool (in turn influenced by evolving educational paradigms)". ICT broadly encompasses technologies used for the acquisition, processing, storage, and dissemination of information, as well as communication and collaboration through various electronic means. The integration of corpus analysis into LSP classes can be considered "potentially highly motivating as they allow exploration of individual questions and are thus learner-centered, fostering autonomy with potential for life-long learning" (Boulton & Tyne, 2014, p. 30X). It is a major condition to familiarize learners with strategies to discover grammatical patterns, word meanings or other aspects of language. Within the context of corpus application in LSP classes, data-driven learning (DDL) can be considered a major paradigm, being an "essentially constructivist, inductive approach" (Boulton, 2017, p. 182) by which learners explore large text collections according to their individual needs. According to DDL, teachers function as mediators supporting students' cognitive understanding and processing of L2 learning, which, consequently, "lead[s] to longer retention than simply 'being taught'" (Boulton, 2017, p. 182). Corino & Onesti (2019, p. 2) put it the following: "The DDL approach

in teaching vocabulary and grammar leads […] to a relevant consciousness-raising of the learners, drawing the student's attention to the formal properties of the target language." In our study, we propose the DDL approach to explore the collocation database CCDB in the LSP classroom, providing Master's students of German as an LSP with insights into a search word's co-occurrence profile. Hence, the DDL approach implies "the hands-on use of authentic corpus data (concordances) by advanced, sophisticated foreign or second language learners in higher education for inductive, self-directed language learning of advanced usage" (Boulton, 2011, p. 572).

4. The Case of CCDB: a brief contextualisation

The collocation database CCDB was created at the Institute for the German Language (IDS) in Mannheim in 2001, allowing for a variety of collocation analyses, based on empirical data that were established from a subset of the Mannheim German Reference Corpus (DeReKo) (Keibel & Belica, 2007). According to the Keibel & Belica (2009, p. 54), DeReKo "is to serve as an empirical basis for the scientific study of contemporary written German", comprising fictional, scientific and newspaper texts, among others. Co-occurrence analysis enables the identification of significant patterns in the usage of word combinations across large corpora. This is achieved through statistical analysis and clustering techniques, which assess the immediate context of a chosen search word within a particular corpus.

5. Exploring CCDB for classroom use

The CCDB can be explored in LSP classes in multiple way, presenting a promising tool for language learning and exploration. As a proposal, this study suggests leveraging the collocational database to provide direct access to DeRoKoVecs (Fankhauser & Kupietz 2017, 2019; Kupietz et al. 2018), a platform that offers insights into paradigmatic and syntagmatic relations between words based on large-scale corpora.

Syntagmatic Info

#	w'	max(a)	⟨a⟩	Σa/Σw'	⊥(a/c)	Σa/Σw	Collocate (W2V)
1		0.996	0.100	2.411e-4	2.411e-4	5.946e-5	stillgelegter
2		0.985	0.098	2.384e-4	2.384e-4	4.909e-5	bestehender
3		0.982	0.098	2.377e-4	2.377e-4	3.783e-5	kerntechnischer
4		0.986	0.359	2.367e-4	5.829e-4	6.916e-5	oberirdischer
5		0.976	0.098	2.362e-4	2.362e-4	4.879e-5	vorhandener
6		0.971	0.097	2.350e-4	2.350e-4	3.533e-5	innerörtlicher
7		0.968	0.097	2.342e-4	2.342e-4	2.750e-5	gemeindeeigener
8		0.965	0.097	2.336e-4	2.336e-4	3.907e-5	veralteter
9		0.963	0.096	2.331e-4	2.331e-4	3.866e-5	ungenutzter
10		0.955	0.095	2.311e-4	2.311e-4	3.625e-5	maroder
11		0.950	0.095	2.300e-4	2.300e-4	2.659e-5	leerstehender
12		0.949	0.095	2.298e-4	2.298e-4	4.688e-5	denkmalgeschützter

Figure 2: Co-occurrences of Rückbau (DeReKoVecs)

Figures 2 and 3 represent the syntagmatic relations of the node word *Rückbau* and its collocates, illustrating the contextual associations and usage patterns surrounding the term. Analyzing these figures provides language learners with valuable insights into how *Rückbau* is commonly employed in various contexts, offering information on specific language patterns and proper usage. Exploring co-occurrences enables students to proficiently use vocabulary in specific authentic contexts.

WPD17/C93/67147	lagen sowie der **Rückbau stillgelegter** Industrieanlagen kommen als neue Kompetenzen hinzu. Seit 2010 entwickelt CSD den Bereich Energie. 2015 verstär
WPD17/D60/85976	tete das auf den **Rückbau stillgelegter** Kernkraftwerke spezialisierte Unternehmen Energiewerke Nord (EWN) bis Ende 2010. Werdegang Rittscher ist gelernt
WPD17/E10/64028	nehmen für den **Rückbau stillgelegter** Kernkraftwerke Einwahlrufnummer, eine Rufnummer zur Anwahl eines Modems über das Telefonnetz, siehe Rufnumm
WPD17/E02/42973	, ist ein auf den **Rückbau stillgelegter** Kernkraftwerke (KKW) spezialisiertes Unternehmen. Es ist auch als ein bundeseigenes Eisenbahninfrastrukturunterne
WPD17/K04/96909	e Kosten für den **Rückbau stillgelegter** Kernkraftwerke bei Insolvenz der Betreiber trägt, ist offen. Im Entwurf der Arbeitsgruppe Umwelt für den Koalitionsve
WPD11/L05/02057	bln-Nippes Gbf (**stillgelegter Rangierbahnhof, z. Zt. Rückbau** von stillgelegten Gleisanlagen außer der in eine Abstellanlage für ca. 20 S-Bahnfahrzeuge u
WPD11/D60/85976	itet das auf den **Rückbau stillgelegter** Kernkraftwerke spezialisierte Unternehmen Energiewerke Nord (EWN). Werdegang Rittscher ist gelernter Elektriker u
WPD11/E10/64028	nehmen für den **Rückbau stillgelegter** Kernkraftwerke {{Begriffsklärung}} {{DEFAULTSORT:Ewn}} EWN
WPD11/E02/42973	sind ein auf den **Rückbau stillgelegter** Kernkraftwerke (KKW) spezialisiertes Unternehmen. Sie sind auch als ein bundeseigenes Eisenbahninfrastrukturunter

Figure 3: Concordance search of node word Rückbau + 1R collocate stillgelegter (DeReKoVecs)

A very interesting feature is the possibility of creating self-organizing maps (SOM) that allow students to interpret co-occurring profiles of a specific search word. Figure 4 shows a two-dimensional self-organizing map (SOM) that "renders in a simplified way the complex multidimensional similarity relations within a set of collocation profiles" (Vachková & Belica, 2009). By compiling and examining a SOM, students get insights into the immediate lexical environment of a node word and its co-occurring items. The interpretation of the lexical feature map reveals that the lexical unit *Rückbau* is predominantly used in the context of physical deconstruction (*Abriss*, *Stilllegung*), redevelopment (*Neubau*, *Umbau*, *Renovierung*), environmental remediation (*Renaturierung*, *Rekultivierung*), infrastructure (*Gleisanlage*, *Straße*), and nuclear facilities (*Kernkraftwerk*, *Reaktor*). The lexeme *Demontage* is used both in

the context of physical disassembly and broader conceptual transformation. Lexical units such as *demontieren*, *Entsorgung*, *Zerlegung*, *Rückbau*, *Abriss*, and *Stilllegung* directly relate to physical disassembly of machinery or structures. In a figurative sense, lexemes like *Kollaps*, *Entmündigung*, *Verfall*, *Destabilisierung* represent the dismantling of systems, institutions, or social constructs.

Another feature for exploration in the classroom involves near-synonyms (Figure 5). The CCDB allows the modeling of semantic proximity by contrasting two near-synonyms, which can be visualized through a two-dimensional color-marked grid. According to Keibel & Belica (2007, p. 4), "[t]he distribution of colors generally provides a reliable idea of how similar the two near-synonyms really are with respect to their usage properties; moreover, it points to the particular usage aspects that the two words do not have in common." In the following figure, the yellow

Rückbau *Demontage*

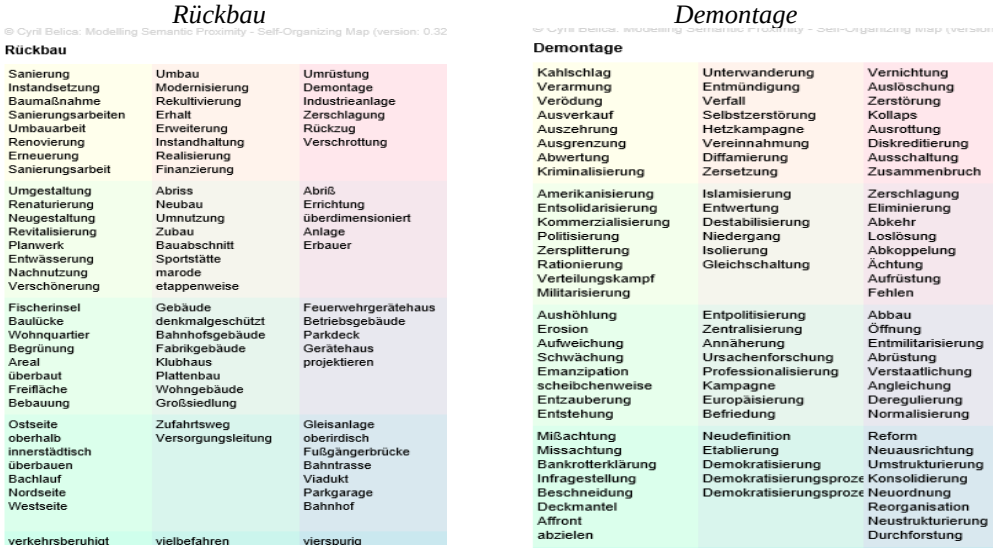

Figure 4: Section of topographic profile of Rückbau *and* Demontage *according to CCDB*

color refers to typical co-occurrences in combination with the lexical unit *Rückbau*, while the red color indicates typical collocates of *Demontage*. The mixed colors indicate co-occurrences where both lexical units are present. The grid shows a gradual transition from one color tone to another, revealing the existing continuum between both lexical units. The different shades are particularly important for the lexical-semantic understanding that students are expected to develop in LSP classes. Pure colors, such as pure yellow and pure red, indicate the most probable usage of both lexical units. For example, the lexical unit *Rückbau* is more commonly used in the context of traffic and infrastructure (e.g. *Verkehrsberuhigung*, *Durchgangsverkehr*,

Umgehungsstraße), as well as within the context of structures or facilities (*Wohnquartier*, *Kernkraftwerk*). The topographic profile of the lexeme *Demontage* reveals different usages, comprising the decay of social structures (*Niedergang*, *Verfall*), as well as the disregard of values (*Missachtung*, *Infragestellung*). The thematic field where the quasi-synonyms *Rückbau* and *Demontage* overlap relates to the context of transformation (*Umstrukturierung*, *Reorganisation*, *Umstellung*), discontinuation (*Schließung*, *Ablösung*, *Auflösung*), as well as political and economic changes (*Deregulierung*, *Abkopplung*, *Aufweichung*).

Figure 5: Topographic profile of near-synonyms Rückbau *and* Demontage

6. Conclusions

In conclusion, the proposal for integrating the collocational database CCDB into LSP classes presents numerous potential applications for linguistic exploration and learning. The examples discussed in this paper aim to broaden the lexical-semantic knowledge of students studying German as LSP, providing valuable insights into the utilization of self-organizing maps to

comprehend complex lexical relationships and usage patterns. By contrasting two near-synonyms, the proposal suggests a pathway for students to develop a deeper understanding of how words may co-occur in natural language contexts, allowing them to identify specific usage aspects of a lexical unit, as well as to enhance their linguistic proficiency in understanding subtle differences in meaning. With the help of the lexemes *Rückbau and Demontage*, various methods for potentially integrating the CCDB in LSP classes have been explored, employing a DDL approach that encourages learners' awareness of corpus usage but also equips them with valuable skills for independent language exploration and analysis.

References

Arnó-Macià, E. (2014). Information technology and languages for specific purposes in the EHEA: options and challenges for the knowledge society. In E. Bárcena, T. Read, & J. Arús (eds.). *Languages for specific purposes in the digital era*. (pp. 3-25). Springer.

Basturkmen, H., & Elder, C. (2004).The practice of LSP. In A. Davies A., & C. Elder (eds.) *The Handbook of Applied Linguistics* (pp. 672–694). Blackwell.

Boulton, A. (2017). Data-driven learning and language pedagogy. In S. Thorne & S. May (eds.), *Language, Education and Technology: Encyclopedia of Language and Education.* (pp. 181-192). New York: Springer. DOI 10.1007/978-3319-02328-1_15-1

Boulton, A., & Tyne, H. (2014). Corpus-based study of language and teacher education. In M. Bigelow & J. Ennser-Kananen (eds.), *The Routledge Handbook of Educational Linguistics*. New York: Routledge, 301-312.

Boulton, A. (2011). Data-driven learning: The perpetual enigma. In S. Goźdź-Roszkowski (Ed.), *Explorations across languages and corpora* (pp. 563-580). Frankfurt: Peter Lang.

Brand, C. (2008). Lexical Processes in Scientific Discourse Popularization. Frankfurt: Peter Lang.

Corino, E., & Onesti C. (2019) Data-Driven Learning: A Scaffolding Methodology for CLIL and LSP Teaching and Learning. *Front. Educ.* 4:7. doi: 10.3389/feduc.2019.00007

Derekovecs database: https://korap.ids-mannheim.de/gerrit/plugins/gitiles/ids-kl/derekovecs

Fankhauser, P., & Kupietz, M. (2019). Analyzing domain specific word embeddings for a large corpus of contemporary German. *International Corpus Linguistics Conference*, Cardiff, Wales, UK, July 22-26.

Gollin-Kies, S., Hall, D. R., & Moore, S. H. (2016). *Language for Specific Purposes. Research and Practice in Applied Linguistics.* Palgrave Macmillan.

Herget, K. (2018). Lokalisierung von Firmen-Websites im Fach 'Angewandtes ProjektDeutsch'. In M. Ellison, M. Pazos Anido, P. Nicólas Martínez & S. Valente Rodrigues (eds.), *As línguas estrangeiras no ensino superior: Propostas didácticas e casos em estudo* (pp. 111-124). Porto: APROLÍNGUAS, FLUP e-DITA.

Herget, K. (2020). Project-based learning: A practical approach to implementing Memsource in the classroom. In *6th International Conference on Higher Education Advances (HEAd'20)* (pp. 717-724). https://doi.org/10.4995/HEAd20.2020.11133

Keibel, H. & Belica, C. (2007). CCDB: A Corpus-Linguistic Research and Development Workbench. *Proceedings of the 4th Corpus Linguistics conference*.

Koskala, M., & Isohella, S. (2018). *Teaching LSP to Technical Communicators.* In J. Humbley, G. Budin, & C. Laurén (Eds.). Languages for Special Purposes: An International Handbook (96-110). Boston, Berlin: DeGruyter.

Kupietz, M., Lüngen, H., Kamocki, P., Witt, A. (2018): German Reference Corpus DeReKo: New Developments – New Opportunities. In: Calzolari, N. et al (eds), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: ELRA, 4353-4360

Vachková, M. & Belica, C. (2009). Self-Organizing Lexical Feature Maps Semiotic Interpretation and Possible Application in Lexicography. In I. Rauch, & R. Seymour (eds.). *IJGLSA* 13, 2 *[Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis*, - Berkeley: IJGLSA/University of California Press], pp. 223-260.