DOI: https://doi.org/10.17234/SRAZ.70.6

UDC: 81'276
UDC: 81'42
Original scientific paper
Received on 17 June 2025
Accepted for publication on 10 November 2025

Compiling a Corpus for Term Extraction Aimed at the Socioterminological Systematization of Mechatronics Terminology

Ivana Jurković
Faculty of Mechanical Engineering and Naval Architecture
University of Zagreb
ivana.jurkovic@fsb.unizg.hr

Dalibor Vrgoč
Institute for the Croatian Language
dvrgoc@ihjj.hr

This paper explores the methodological framework for compiling a specialized corpus intended for term extraction, with the objective of supporting the socioterminological systematization of mechatronics terminology. The primary aim of the study is to develop empirically grounded strategies for corpus compilation that enhance the reliability and efficiency of term extraction. It was hypothesized that a corpus-driven approach, when applied exclusively to a didactic subcorpus (mechatronics textbooks), allows for a more precise extraction of term candidates than the approach involving an academic subcorpus (scientific papers in the field of mechatronics). To test this hypothesis, two subcorpora of mechatronics texts written in English were compiled (didactic and academic). Terms were extracted from both the didactic and the academic subcorpus of mechatronics texts and compared to the items extracted by other extraction functions available in Sketch Engine. The comparative statistical analysis has demonstrated that the didactic subcorpus allows for the extraction of more term candidates, while the academic subcorpus produces more noise. Thus, it is argued that a smaller, balanced didactic corpus is more suitable for term extraction aimed at the socioterminological systematization of mechatronics terminology than a larger corpus involving the academic subcorpus. The contribution of the outlined methodology is reflected in greater efficiency in socioterminological systematization of mechatronics terminology. Thus, it is proposed that it may be effectively extended to other interdisciplinary domains.

Key words: corpus-driven approach, mechatronics, socioterminology, term extraction, terminology systematization

1. Introduction

The history of terminology research, systematization and standardization is extensive and has spanned several centuries. A significant contribution to scientifically grounded terminology studies was made in the 20th century by Eugen Wüster, an engineer who presented a systematic methodology for terminological work (Wüster, 1974: 63; [3]1991: 8-61), which continues to be regarded as a foundational framework for contemporary terminological research. Wüster's contribution is reflected in the General Theory of Terminology (GTT) (Wüster, 1974: 63), whose onomasiological approach underlies the majority of international standards and manuals for terminological and terminographic work (cf. Felber, 1984: 3; Hudeček/Mihaljević, 2012: 29). GTT is rooted in structuralist principles, and in response to it, alternative approaches to terminology studies have since emerged. These newer frameworks address terminology from sociolinguistic, cognitive, sociocognitive and cultural perspectives (cf. Guespin, 1995: 209; Cabré, 1999: 10; Temmerman, 2000: 38; Faber, 2009: 113-120; Diki-Kidiri, 2022: 5). Post-Wüsterian reactions to the General Theory of Terminology have been twofold in nature; both supportive and critical (Humbley, 2022: 15). However, a careful analysis of Wüster's works suggests that GTT was more progressive than it is often assumed to be, a view also supported by Trojar (2017: 81) and Vrgoč (2021: 36). Moreover, Wüster demonstrated a keen awareness of the importance of developing computational tools and their potential for storing and processing information (Wüster 1974: 98) as a prerequisite for contemporary linguistic research grounded in corpus linguistics methodologies. Nevertheless, more recent theoretical approaches such as socioterminology have proven to be valuable enhancements to Wüster's original framework.

In contrast to more recent approaches, the early development of terminology was marked by efforts toward standardization, primarily that of technical terminology. The father of terminology, Eugen Wüster, based his terminological work on the study and standardization of technical terms[1]. Due to the rapid and continuous development of new technologies, technical terminology still involves numerous research and practical challenges.

The systematization, maintenance and care of contemporary technical terminology require continuous engagement of both field experts and linguists due to the rapid advancements of new technologies. Precisely because of technological progress and the need to apply an interdisciplinary approach in the technical field, mechatronics emerged; a young discipline within the field of technical sciences characterized by a high degree of interdisciplinarity. Although numerous definitions of mechatronics may be found in literature (cf. Auslander, 1996: 5; Tomizuka, 2000: 1; Cintra Faria & Barbalho, 2023: 1), most of them identify it as an interdisciplinary or synergistic integration of various technical disciplines, such as mechanical engineering, electrical engineering and computer engineering. The term *mechatronics* was introduced in 1969 by the Japanese company Yaskawa

[1] See Wüster's dictionary *The Machine Tool*, 1968.

Electric Corporation[2] (Bishop, [2]2008: 1-1; Purković/Salopek, 2015: 9). According to Bishop ([2]2008: 1-1), who cites the original patent filed by Yaskawa, the term *mechatronics* in Japanese was formed by combining the first part of the Japanese word for *mechanism* and the second part of the Japanese word for *electronics*. As Kurosawa (1983: 1475) notes, Yaskawa Electric Corporation registered the term as a protected trademark in 1972. Bearing in mind that it was not before 1982 that unrestricted use of the term *mechatronics* was permitted (Purković/Salopek, 2015: 9) and that the European universities started offering study programs in mechatronics in the 1990s (Grimheden/Hanson, 2005: 187), it becomes clear how young mechatronics in fact is, especially if compared with traditional technical disciplines such as mechanical engineering. Thus, mechatronics is a very interesting field from the perspective of terminological research because its terminology is still underexplored both in English and in other languages. Furthermore, because of its interdisciplinary nature, mechatronics terminology offers potential for conducting variation studies.

The main aim of this paper is to present empirically grounded insights into the methodology of compiling a corpus for term extraction in the scope of the systematization of mechatronics terminology that may be applied to terminology systematization of other emerging and interdisciplinary fields. For this purpose, we have collected a corpus of mechatronics texts written in the English language. We hypothesize that a corpus-driven approach to term extraction aimed at terminology systematization yields more effective results when the corpus is limited to the didactic subcorpus of mechatronics texts (such as handbooks and university textbooks) compared to when it is broadened to include scientific papers belonging to the academic subcorpus. Thus, the following research questions were asked:

1. Is there a significant difference between term lists extracted from the didactic and the academic subcorpora of mechatronics texts?

2. Which of the two subcorpora (didactic or academic) allows for a more precise extraction of mechatronics term candidates?

2. Theoretical and Methodological Considerations

It is generally considered that socioterminology emerged in the 1990s. However, sociolinguists have identified the potential of approaching terminological research from the sociolinguistic perspective at least a decade earlier (cf. Bugarski, 1986a: 22). Furthermore, it was around the same period that scholars started to recognize the value of approaching terminological research from the standpoint of applied linguistics (cf. Bugarski, 1986b: 101-

[2] Different names of the Yaskawa Electric Corporation may be found in literature. In this paper we refer to the official name that may be found on the company's website: https://www.yaskawa-global.com/ (accessed on: 16 June 2025).

108). According to the socioterminological theory, a term acquires the status of a term only if it is embedded within discourse, whereby the circumstances of its emergence, dissemination, usage, and interpretation are the factors that grant terminological status to a given lexical or multi-word unit (Delavigne/ Gaudin, 2022: 191). We consider the socioterminological approach to be an appropriate choice for the systematization of mechatronics terminology, as mechatronics is an interdisciplinary field in which reterminologization seems to be particularly frequent. In other words, mechatronics has adopted a number of terms from traditional technical disciplines. At the same time, the development of mechatronics has given rise to original mechatronic terms that have started circulating in other technical disciplines.

The development of socioterminology and the increasing use of computational tools have significantly influenced the evolution of contemporary methods for terminology analysis (Grčić Simeunović/Frleta, 2012: 231). In describing the contributions of the Canadian school of terminology, L'Homme (2006: 60-61) identifies two modern terminological approaches grounded in corpus analysis: the creation of terminological databases, which formed the basis for Meyer's (2001: 289-299) method of extracting relevant conceptual information from context, and the lexical semantic approach, which adopts a semasiological perspective and does not consider standardization as an element of terminological work (L'Homme, 2022: 241). Canadian terminologists have also played a key role in developing a systematic approach to concept analysis, encompassing four categories of terminological activities: corpus selection, term creation, preparation of terminological records, and quality control (Meyer, 2022: 114). Bearing in mind that socioterminological analysis depends on adequate corpus selection, it is essential that the corpus is selected and compiled in a well-considered manner. Furthermore, it should be clarified which approach is more suitable for the systematization of mechatronics terminology: a corpus-based or a corpus-driven approach.

2.1. The Corpus Approach

A corpus may broadly be defined as a large, naturally-occurring collection of texts, i.e. words in context or complete sentences (Borucinsky, 2023: 22). Sinclair defines a corpus as "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair, 1991: 171). In other words, a corpus should be understood as a sample of language. Similar to practices in the natural and social sciences, sample selection depends on the specific aims of the research.

Anthony (2020: 563-565) identifies three defining characteristics of a corpus:

- The language variety the corpus aims to model (e.g. a corpus may model general language or the language of a specific genre or discipline);

- Corpus design (depending on their purpose, corpora can be "large" or "small"; for the purpose of comparison, corpora may be compiled as comparable corpora, and as parallel corpora[3]);

- The way that a corpus is stored and accessed (corpora may be public or private, stored on a local computer or a network server).

While discussing the function of the corpus, Tognini-Bonelli (2001: 10) addresses the three main issues in corpus compilation, which may be considered from both the theoretical and the practical perspective:

- Authenticity;

- Representativeness;

- Sampling criteria.

Authenticity reflects the fact that if corpora are to be used as a source of evidence, it must be presumed that a corpus comprises genuine, naturally occurring communication. If this is not the case, it should be clearly stated (e.g. if a corpus is compiled for specific purposes, such as specialized corpora for terminological research). Representativeness of the language included in the corpus refers to whether "the findings based on its contents can be generalized to a larger hypothetical corpus" (Leech, 1991: 27). In recent publications one may find the terms *balance* and *diversity* to be used instead of the term *representativeness* in this context (Corpas & Seghiri, 2009: 80; Stefanowitsch, 2020: 28; Borucinsky, 2023: 23). As related to sampling criteria, if a corpus is to serve as the basis for generalizations about language, a central concern in corpus design is the precise definition of the target population the corpus is intended to reflect (Biber, 1994: 378). One of the most debated issues in the context of sampling pertains to determining the appropriate sample size (Tognini-Bonelli, 2001: 62). Thus, in the initial stages of the research it is of paramount importance to take all sampling criteria into account and compare them with previous research for benchmarking purposes.

Most corpus linguists agree that corpora as linguistic samples may be observed by taking two major analytical approaches: a corpus-based and a corpus-driven approach (Biber, 2015: 196). Essentially, a corpus-based approach uses a corpus to test, exemplify, or refine pre-existing theories or hypotheses, while a corpus-driven approach aims to let the data itself guide the theoretical insights (Tognini-Bonelli, 2001: 84). In other words, a corpus-driven approach is data-led and inductive[4], which allows for minimizing researcher bias by not imposing external theoretical assumptions. Furthermore, there is another approach that should be mentioned in this context: a corpus-illustrated approach

[3] For the distinction between comparable and parallel corpora see also Teubert, 1996: 245-247.

[4] Further details on the methodology of using corpus data in linguistics are presented in Gries, ²2017.

(Mihaljević, 2021: 17), in which the researcher starts from their own intuition and then seeks confirmation in the corpus (Tummers et al., 2005: 227). Examples of the implementation of the corpus-illustrated approach may be found in areas such as lexicography, but according to Borucinsky (2023: 21) it does not belong to corpus linguistics because data are not collected systematically. Thus, in this paper we shall consider only the two major corpus approaches, the corpus-based and the corpus-driven approaches[5], respectively.

2.2. Selection of a Suitable Methodological Approach for the Socioterminological Systematization

As already pointed out, socioterminology is interested in terms as elements of discourse. This means that instead of observing terms in isolation, they are analyzed in context, and for this purpose specialized corpora are selected and compiled. The question remains which methodological approach is the most suitable for socioterminological systematization purposes. Both corpus-based and corpus-driven approach add a particular value to terminological research by offering different insights and perspectives. On the other hand, both have their limitations. The primary advantages of the corpus-based approach are its reliability and validity (Biber, 2015: 197). However, its main limitation lies in its focus on the specific linguistic phenomenon under investigation, which may result in the researcher overlooking language features beyond the immediate scope of the study (Borucinsky, 2023: 21). In contrast, the corpus-driven approach formulates hypotheses based on the findings that emerge from corpus analysis. Scholars continue to debate whether a research approach can be fully driven by corpus data and whether texts can be examined without any subjective bias or preconceived theoretical influence. Corpus-driven studies often integrate aspects of corpus-based research, which is why Biber (2009: 302) supports the implementation of a hybrid methodological framework. Borucinsky (2023: 22) argues that it is also not possible to assert that one of the two approaches is inherently superior, as the choice between them depends on the specific focus and objectives of the research.

Building on the methodologies proposed by Biber (2012: 19) and Borucinsky/Pritchard (2022: 8-10), we argue that a corpus-driven approach is more appropriate for term extraction aimed at the socioterminological systematization, as it enables the data to reveal patterns independently, thereby reducing potential bias resulting from the researcher's subjective intuition. This is especially useful in cases when the terminologist is not an expert in the field whose terminology is being systemized.

Furthermore, contrary to some views related to corpus representativeness and sampling[6], we hypothesize that a small, carefully compiled didactic corpus

[5] For a more detailed clarification of the differences between the corpus-based and the corpus-driven approach see also Gries, 2010: 328-330.

[6] The frequently quoted general rule of corpus sizing is "the bigger, the better" (see also Łukasik, 2014: 78).

is more effective in term extraction for terminology systematization purposes than an academic corpus or a combination of both. In the following sections it is shown how this hypothesis was tested on the example of mechatronics texts written in the English language.

3. Materials and Methods

The corpus for this study was compiled following the theoretical framework presented in Tognini-Bonelli (2001: 54-62). The corpus, named MECH, contains authentic texts in the domain of mechatronics. As the main purpose of compiling the corpus is aimed at socioterminological systematization of mechatronics terminology, we included mechatronics textbooks and scientific articles, i.e. we did not include texts that belong to the category of popular science, as they are more suitable for variation studies (cf. Grčić Simeunović, 2021: 98-101). Table 1 shows the description of the corpus.

Table 1. *Corpus compiled for the purposes of this study.*

No.	Domain	Genre	Language	Corpus name	Number of tokens	Number of unique words	Number of documents
1.	Mechatronics	Textbooks Scientific articles	English	MECH	1,551,298	72,831	55

Even though there is no clear consensus among scholars as to the recommended size of a specialized corpus, most studies quote LSP corpora consisting of 500,000 to 1,000,000 tokens to be sufficient (cf. Bergenholtz and Tarp, 1995: 95; Pearson, 1998: 56-57; Grčić Simeunović et al., 2020: 625; Cigan, 2023: 84). In the sampling phase we followed Bowker and Pearson's (2002: 47-48) argument that a small, balanced, and carefully selected LSP corpus can yield more valuable insights than a larger, non-specialized corpus that is not specifically adapted to the research objectives. Regarding corpus size, we adopted Biber's (2006: 252) position that a corpus should be sufficiently large to capture the linguistic phenomenon under investigation, while also aligning with Anthony's (2020: 563) perspective that the true value of a corpus is not related to its size, but to the quality and relevance of the information it provides[7].

To achieve balance from the aspect of genre, we compiled two subcorpora: the didactic subcorpus consisting of textbooks (MECHD) and the academic subcorpus consisting of scientific articles (MECHA). While compiling the corpus, particular attention was paid to achieving a balanced representation of the two subcorpora. Following Zanettin (1998: 4-5) and Bowker and Pearson's (2002: 71),

[7] See also Borucinsky/Pritchard, 2022: 8.

we took the publication date criterion into account, so we selected recent texts published between 2005 and 2025.

To assure that the two subcorpora may be used as comparable corpora, we aimed at achieving both size-related and reliability-related balance. As mechatronics is a highly interdisciplinary field, one of the challenges was connected to assuring that the content of the selected texts reflects the purposes of this study. In other words, the aim was to select texts strategically so as to reduce, to the greatest extent possible, the extraction of terms outside the mechatronics domain. Thus, only textbooks and scientific articles that include the words *mechatronics* and *mechatronic* in their titles were included in the corpus, excluding the ones that were strictly connected to the education in mechatronics, as their content does not belong to the area of technical, but rather to social sciences. Regarding the choice of texts, Bowker (2003: 161) highlights the importance of quality and appropriateness of texts to be included in the corpus. Following Łukasik's (2014: 6) argument that "a specialised corpus must represent the style and value of mainstream academic writing of a given field", we selected mechatronics texts based on the corpus compilation criteria according to Pearson's corpus design methodology (1998: 58-62). In other words, all texts included in our corpus are written, published by high rank publishing houses and journals, produced by acknowledged individuals and factual from the standpoint of mechatronics. The didactic subcorpus includes two acknowledged mechatronics textbooks (Bishop, 2008; Jouaneh, 2013), while the academic subcorpus includes a total of 53 articles published by MDPI open access journals (MDPI, 2025) in the last five years that include the lexemes *mechatronics* and *mechatronic* in their titles, excluding articles dealing with education in mechatronics. The search was conducted by entering the aforementioned keywords into the search tool on the MDPI platform. Most of the selected articles were published in the MDPI journals *Applied Sciences* (33.96%) and *Machines* (18.87%).

Furthermore, to ensure that the corpus contains term-rich texts (cf. Pearson on factuality and technicality, 1998: 61), we consulted a field expert during the corpus compilation phase. The final outcome of the described sampling process is a balanced corpus of mechatronics texts involving two subcorpora that may be used as comparable corpora. Descriptions of the two subcorpora are shown in Table 2.

Table 2. *Subcorpora compiled for the purposes of this study.*

No.	Domain	Genre	Language	Corpus name	Number of tokens	Number of unique words	Number of documents
1.	Mechatronics	Textbooks	English	MECHD	850.318	31.114	2
2.	Mechatronics	Scientific articles	English	MECHA	700.980	41.717	53
		Total			1.551.298	72.831	55

As to the disparity between the number of documents included in the two subcorpora, MECHD is made up of a significantly smaller number of texts than MECHA because the selected textbooks are very long. To mitigate the effects of the aforementioned imbalance and to reduce the influence of individual authors' idiosyncrasies, one of the textbooks included in the MECHD subcorpus was chosen due to its composition as a compilation of chapters on mechatronics topics written by multiple authors.

Both subcorpora were manually collected and automatically processed (i.e. lemmatized, tokenized, POS tagged) in *Sketch Engine* (*SkE*) (Kilgarriff et al., 2004: 24). The two subcorpora were compared based on the results they yield using three extraction functions available in *SkE*: Wordlist (i.e. frequency list extraction), N-grams (i.e. extraction of multi-word expressions), and Keywords (i.e. terminology extraction). Bearing in mind that single-word terms are mostly nouns (cf. Nahod, 2020: 187), the extracted single-word terms using the Keywords function are expected to mostly belong to the category of nouns. Thus, to ensure that the results of the Wordlist and the Keywords functions are as comparable as possible, the settings in the Wordlist function were adjusted to extract nouns only. For benchmarking purposes, as both the Wordlist and the N-grams functions operate with simple frequencies in the focus corpus, the N-grams function was used to extract 4-grams that belong to the category of lexical bundles. As opposed to the Wordlist and the N-grams functions, the Keywords function operates using a combination of normalized frequencies and two corpora (focus and reference corpora). We used English Web 2021 (enTenTen21) (Jakubíček et al., 2013: 3) as the reference corpus, as it is an all-purpose English corpus covering the largest possible variety of genres, topics, text types and web sources.

As the Wordlist and the N-grams functions operate using simple frequencies only, while the Keywords function operates with normalized frequencies and compares the focus corpus with the reference corpus, we hypothesized that using the Wordlist and N-grams functions would result in a significantly greater overlap in extracted items between MECHD and MECHA than the Keywords function would. In other words, our starting point was that both subcorpora would yield similar extraction results when comparing the most frequent words and lexical bundles in mechatronics texts, but that the extracted terms using the Keywords functions (which is of relevance for socioterminological systematization purposes) would differ significantly depending on whether they are extracted from MECHD or MECHA.

Thus, we proceeded with the extraction of the fifty most frequent nouns using the Wordlist function, the fifty most frequent lexical bundles using the N-grams function and fifty single-word terms with the highest keyness score using the Keywords function, from both MECHD and MECHA. Data were processed using *SkE* and downloaded in *MS Office Excel*. To reduce noise as much as possible, all six tables were manually checked, and noise was eliminated following the exclusion methodology presented in Borucinsky and Pritchard (2022: 11). From the wordlists and term lists we excluded proper names (e.g. *Laplace*), initialisms (e.g. *IEEE*), symbols (e.g. θ) and miscellaneous (e.g. *xx*), while from the 4-gram

list we excluded 4-grams that include initialisms (e.g. "*PM brush DC motor*") and symbols ("*where T is the*"), topic specific bundles (e.g. "*architecture of the system*"), lexical bundles beginning with *and* (e.g. "*and the Internet of*"), lexical bundles that include the combination of more than two lexical words (e.g. "*electronic devices and digital*") and miscellaneous (e.g. "*a b a b*").

Having obtained the final lists, we compared the results using descriptive statistics and the chi-square test with Yates's correction for continuity.

4. Results

To answer our research questions, we extracted the fifty most frequent words, the fifty most frequent lexical bundles and the top fifty single-word terms with the highest keyness scores from MECHD and MECHA.

Lexical bundles are pragmatic multi-word lexical units that tend to co-occur with high frequency within a given register (Pritchard, 2015: 6). They represent formulaic sequences that belong to the category of lexical items, but essentially function as grammatical units (e.g. "*is determined by the*") (Borucinsky & Pritchard, 2023: 7). As they are semantically transparent, lexical bundles provide "the building blocks of coherent discourse" (Hyland, 2008: 6) and are, therefore, important in the LSP context. As they are formulaic in nature and both subcorpora belong to the same domain, we expected that the greatest overlap between MECHD and MECHA would be found in the results extracted using the N-grams function, followed by results extracted using the Wordlist function. Finally, we expected to find the greatest differences between MECHD and MECHA in the results extracted using the Keywords function. The reason why we expected the Wordlist function to yield results with a greater overlap between MECHD and MECHA than the ones extracted by the Keywords function is because the Wordlist results are only based on frequency. In other words, they are expected to contain both domain-specific and cross-disciplinary terms, as well as general language words.

Table 3 shows the fifty most frequent lexical bundles extracted from MECHD and MECHA. Lexical bundles found in both subcorpora are printed in bold. As may be seen from Table 3, of the fifty lexical bundles in MECHD, 21 are also present in MECHA (42%).

Table 3. *Fifty most frequent four-word lexical bundles found in the two subcorpora.*

MECHD			MECHA		
Item	f	$f_{rel.}$	**Item**	f	$f_{rel.}$
can be used to	151	177.58062	**is shown in figure**	81	115.55251
is shown in figure	125	147.00383	in the field of	45	64.19584
a function of the	71	83.49817	**as well as the**	45	64.19584
the output of the	69	81.14611	**as shown in figure**	44	62.76927
is proportional to the	64	75.26596	**on the other hand**	44	62.76927
as well as the	58	68.20978	the position of the	39	55.63639
in the form of	56	65.85771	**can be used to**	38	54.20982
is a function of	56	65.85771	in the case of	37	52.78325
on the other hand	55	64.68168	**a wide range of**	37	52.78325
is defined as the	51	59.97756	**it is possible to**	37	52.78325
as a function of	48	56.44947	**is based on the**	35	49.9301
with respect to the	45	52.92138	it is necessary to	32	45.65038
the value of the	44	51.74535	**with respect to the**	32	45.65038
as shown in figure	43	50.56932	**at the same time**	30	42.79723
can be used for	42	49.39329	**in the form of**	29	41.37065
is illustrated in figure	38	44.68916	with the use of	25	35.66436
in terms of the	38	44.68916	**are shown in figure**	25	35.66436
can be found in	36	42.3371	a crucial role in	25	35.66436
the use of a	36	42.3371	**the value of the**	24	34.23778
the magnitude of the	35	41.16107	**can be found in**	24	34.23778
the use of the	35	41.16107	**in terms of the**	24	34.23778
is equal to the	35	41.16107	in the context of	22	31.38463
are shown in figure	34	39.98504	**can be used for**	22	31.38463
the ratio of the	34	39.98504	the design of the	22	31.38463
can be used in	34	39.98504	of the system is	22	31.38463
is based on the	33	38.80901	with the help of	22	31.38463
is connected to the	32	37.63298	as illustrated in figure	21	29.95806
a wide range of	32	37.63298	the movement of the	20	28.53148
it is possible to	32	37.63298	it is important to	19	27.10491
at the same time	31	36.45695	the performance of the	19	27.10491
is the number of	31	36.45695	the output of the	19	27.10491
is said to be	31	36.45695	the dynamics of the	19	27.10491

one of the most	30	35.28092	**in the case of**	19	27.10491
can be used as	30	35.28092	can be expressed as	18	25.67834
is applied to the	29	34.10489	**is illustrated in figure**	18	25.67834
that can be used	29	34.10489	the motion of the	17	24.25176
the sum of the	29	34.10489	**is the number of**	17	24.25176
is determined by the	28	32.92886	the accuracy of the	17	24.25176
the operation of the	28	32.92886	**the use of the**	17	24.25176
it should be noted	28	32.92886	the case of the	16	22.82519
an example of a	28	32.92886	at the end of	16	22.82519
is given by the	28	32.92886	of the system and	16	22.82519
the difference between the	28	32.92886	is applied to the	15	21.39861
in the case of	27	31.75283	as a result of	15	21.39861
can also be used	27	31.75283	it is crucial to	15	21.39861
is related to the	26	30.5768	the state of the	15	21.39861
if and only if	26	30.5768	**the use of a**	15	21.39861
in the presence of	26	30.5768	the distance between the	14	19.97204
is one of the	26	30.5768	in the design of	14	19.97204
may be used to	25	29.40077	the implementation of the	14	19.97204

Table 4 shows the fifty most frequent nouns extracted from MECHD and MECHA using the Wordlist function. Nouns found in both subcorpora are printed in bold. As may be seen from Table 4, of the fifty nouns in MECHD, more than a half are also present in MECHA (54%).

Table 4. *Fifty most frequent nouns found in the two subcorpora.*

MECHD			MECHA		
Item	f	$f_{rel.}$	**Item**	f	$f_{rel.}$
system	4,378	5,148.66203	**system**	4,758	6,787.64016
control	2,195	2,581.3872	**figure**	2,470	3,523.63833
output	2,121	2,494.36093	**control**	2,354	3,358.15572
signal	2,042	2,401.45451	**model**	2,262	3,226.9109
figure	1,989	2,339.12489	**design**	1,421	2,027.16197
sensor	1,735	2,040.41312	robot	1,304	1,860.25279
function	1,700	1,999.25204	**sensor**	1,198	1,709.03592
input	1,695	1,993.37189	**process**	1,096	1,563.52535
time	1,643	1,932.2183	**method**	1,043	1,487.91692

voltage	1,606	1,888.70517	parameter	1,024	1,460.81201
state	1,540	1,811.08715	**time**	997	1,422.2945
circuit	1,528	1,796.97478	**controller**	975	1,390.90987
frequency	1,355	1,593.52148	**measurement**	923	1,316.72801
example	1,334	1,568.82484	**value**	843	1,202.60207
motor	1,220	1,434.75735	**datum**	833	1,188.33633
value	1,208	1,420.64498	**application**	833	1,188.33633
design	1,118	1,314.80223	machine	783	1,117.00762
device	1,108	1,303.04192	approach	779	1,111.30132
model	1,070	1,258.35276	simulation	757	1,079.91669
equation	1,044	1,227.77596	result	739	1,054.23835
application	946	1,112.52496	technology	698	995.74881
force	909	1,069.01183	**force**	669	954.37816
temperature	895	1,052.5474	**motor**	662	944.39214
power	894	1,051.37137	algorithm	653	931.55297
measurement	887	1,043.13915	level	648	924.4201
datum	838	985.51365	**analysis**	599	854.51796
type	834	980.80953	review	595	848.81166
number	827	972.57732	**function**	569	811.72073
element	782	919.65594	**case**	550	784.61582
process	760	893.78327	development	546	778.90953
code	760	893.78327	**component**	542	773.20323
controller	755	887.90311	research	532	758.93749
method	751	883.19899	motion	531	757.51091
speed	747	878.49487	**element**	529	754.65776
case	730	858.50235	**software**	523	746.09832
variable	712	837.3338	**energy**	518	738.96545
flow	702	825.57349	part	518	738.96545
torque	698	820.86937	position	515	734.68573
table	679	798.52479	engineering	507	723.27313
response	673	791.4686	study	499	711.86054
energy	671	789.11654	structure	491	700.44794
range	640	752.65959	industry	485	691.8885
form	627	737.3712	**signal**	483	689.03535
component	619	727.96295	performance	474	676.19618
operation	593	697.38615	**information**	464	661.93044
software	591	695.03409	solution	452	644.81155
actuator	586	689.15394	section	446	636.2521

Table 5 shows the top fifty single-word terms with the highest keyness scores extracted from MECHD and MECHA using the Keywords function. Terms found in both subcorpora are printed in bold. As may be seen from Table 5, of the fifty single-word terms found in MECHD, only nine may be found in MECHA (18%).

Table 5. *Top fifty single-word terms ordered by keyness score found in the two subcorpora.*

MECHD				MECHA			
Item	$f_{rel.\ (focus)}$	$f_{rel.\ (ref.)}$	**Keyness**	**Item**	$f_{rel.\ (focus)}$	$f_{rel.\ (ref.)}$	**Keyness**
mechatronics	291.6556	0.22739	238.437	**mechatronics**	513.56671	0.22739	419.237
flowmeter	250.4945	0.23145	204.227	coiler	166.90918	0.02582	163.683
actuator	689.1539	2.55444	194.166	cobot	176.8952	0.16713	152.421
accelerometer	341.0489	1.0487	166.959	**actuator**	476.4758	2.55444	134.332
transducer	430.4272	1.81478	153.272	pendulum	355.21698	1.96103	120.302
stator	272.8391	0.86351	146.948	repeatability	178.32178	0.5667	114.458
op-amp	170.5244	0.26029	136.1	manipulator	228.25188	1.02931	112.97
microcontroller	381.0339	1.82634	135.169	gait	318.12604	1.89621	110.187
thermocouple	171.7005	0.57003	109.998	EtherCAT	122.68539	0.15901	106.716
actuation	151.708	0.58158	96.554	machine	275.32883	2.1146	88.72
micromotor	94.08245	0.02522	92.744	decoupling	131.24483	0.62815	81.224
microprocessor	245.7904	1.73195	90.335	robot	1,860.25281	24.98433	71.63
voltage	1888.705	21.13365	85.377	morphism	94.1539	0.33305	71.381
micromachining	89.37833	0.06329	84.998	harvester	158.34973	1.3222	68.62
stepper	164.6443	0.95429	84.759	subsystem	231.10503	2.6053	64.379
thyristor	94.08245	0.13532	83.749	biomimetics	61.34269	0.02835	60.624
transistor	455.1238	4.5617	82.012	**torque**	584.89545	9.09101	58.061
torque	820.8694	9.09101	81.446	oscillogram	52.78325	0.00745	53.385
gage	226.9739	2.00115	75.962	robotics	166.90918	2.20971	52.313
potentiometer	112.8989	0.64999	69.03	controller	1,390.90991	25.97793	51.594
capacitance	154.06	1.34571	66.104	**inertia**	158.34973	2.10736	51.281
inductance	110.5469	0.71086	65.199	maintainability	62.76926	0.29781	49.136
resistor	398.6744	5.20247	64.438	servomotor	51.35667	0.08081	48.442
displacement	488.0527	6.59364	64.403	covariance	81.31473	0.72559	47.702
microtransducer	62.32962	0.00065	63.289	simulation	1,079.91675	22.3487	46.295
encoder	218.7417	2.59114	61.19	**sensor**	1,709.03589	36.99648	45.005
nonlinearity	82.32214	0.36653	60.974	locomotion	77.03501	0.75209	44.538
impedance	230.502	2.83524	60.362	figure	82.7413	0.89493	44.192
flip-flop	123.4832	1.06544	60.27	obsolescence	71.32871	0.63802	44.156
armature	102.3147	0.71971	60.077	conceptualization	74.18186	0.7988	41.796

winding	149.3559	1.50648	59.987	spindle	179.74835	3.47089	40.428
inertia	184.6368	2.10736	59.741	**armature**	67.04899	0.71971	39.57
rotor	346.929	4.84758	59.5	reflectivity	38.51751	0.00153	39.457
electroplate	77.61802	0.32692	59.248	cubature	38.51751	0.007	39.243
H-bridge	63.50565	0.0925	59.044	femur	68.47556	0.77561	39.128
causality	145.8278	1.49055	58.954	kinematics	59.91612	0.59577	38.174
photodiode	75.26596	0.29997	58.667	friction	282.4617	6.48017	37.895
microdevice	57.6255	0.02048	57.449	waypoint	85.59445	1.29122	37.794
servovalve	56.44947	0.00309	57.273	anisotropy	58.48954	0.58317	37.576
oscillator	214.0376	2.86367	55.656	**microcontroller**	99.8602	1.82634	35.686
cantilever	102.3147	0.86589	55.37	echo-doppler	34.23778	0.00179	35.175
hysteresis	76.44199	0.39928	55.344	gripper	52.78325	0.53116	35.126
diode	272.8391	4.00727	54.688	**excitation**	85.59445	1.52726	34.264
equation	1,227.776	21.51499	54.576	axis	473.62265	13.1279	33.595
dynamometer	62.32962	0.16674	54.279	measurement	1,316.72803	38.51643	33.346
interfacing	104.6667	0.961	53.884	cascade	246.79735	6.52044	32.95
sensor	2,040.413	36.99648	53.726	estimation	255.35678	6.80234	32.856
eigenvalue	95.25848	0.80474	53.336	parameter	1,460.81201	43.54779	32.814
thermistor	64.68169	0.34735	48.749	**actuation**	49.9301	0.58158	32.202
excitation	119.9551	1.52726	47.86	stiffness	114.12594	2.61451	31.851

As may be seen from the results presented in Tables 1, 2 and 3, our hypotheses were mostly correct. Both the number of lexical bundles extracted by the N-grams function and the number of nouns extracted by the Wordlist function that are found in both subcorpora are higher than the number of single-word terms that were found in both subcorpora. As opposed to our expectations, there is a greater overlap between MECHD and MECHA as related to nouns extracted by the Wordlist function than the lexical bundles. However, by testing the statistical significance of this difference using the chi-square test with the significance level amounting to .05 and the Yates's correction (cf. Table 6), we conclude that the difference between the two functions (Wordlist and N-grams, respectively) as related to the item overlap in both subcorpora is not statistically significant at $p < .05$.

Table 6. *χ2 test results: lexical bundles (N-grams) vs nouns (Wordlist) found in both subcorpora*

χ2 with Yates correction	**Significance level**	***p*-value**
1.0016	.05	.316923

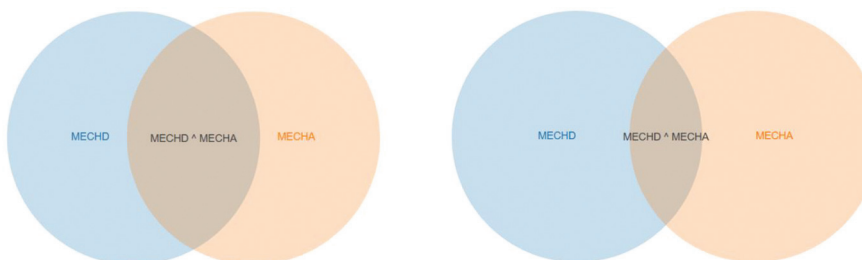
Further steps in the analysis involved subjecting the Keywords function to the comparison with the N-grams and the Wordlist functions using the same

statistical test. The results are shown in Table 7 (N-grams vs Keywords) and Table 8 (Wordlist vs Keywords). For illustration purposes, a visualization of the overlaps of lexical bundles and terms in MECHD and MECHA is shown in Figure 1 below Table 7.

The results indicate that in both cases there is a statistically significant difference between the results rendered by the functions being compared. In other words, the overlap of terms found in both MECHD and MECHA is significantly smaller than the overlap of lexical bundles and nouns extracted by the Wordlist function in both subcorpora.

Table 7. *χ2 test results: lexical bundles (N-grams) vs terms (Keywords) found in both subcorpora*

χ2 with Yates correction	**Significance level**	***p*-value**
5.7619	.05	.016377



(a) (b)

Figure 1. *Proportionally adjusted visualization of the overlap differences between MECHD and MECHA: (**a**) lexical bundles (N-grams); (**b**) single-word terms (Keywords).*

Table 8. *χ2 test results: nouns (Wordlist) vs terms (Keywords) found in both subcorpora*

χ2 with Yates correction	**Significance level**	***p*-value**
12.5434	.05	.000398

To test our hypothesis that the overlap of extracted items in both subcorpora and the type of *SkE* extraction function being used are statistically dependent variables, we ran the χ2 test for the 2 x 3 contingency table involving all three functions (N-grams, Wordlist and Keywords). For this purpose we formulated the following null hypothesis: The overlap of extracted items found in both subcorpora and the *SkE* extraction function being used are not statistically correlated. The results of the χ2 test are shown in Table 9. As the results indicate, the aforementioned correlation is statistically significant, which means that the null hypothesis was rejected.

Table 9. *χ2 test results: the correlation between the overlap of extracted items in MECHD and MECHA and the SkE extraction function being used*

χ2	**Significance level**	***p*-value**
14.2615	.05	.0008

Having confirmed that the overlap of extracted terms in both subcorpora is minimal as compared to the overlap of items obtained by other extraction functions in *SkE*, we carried out a qualitative check of the differences between the term lists obtained from MECHD and MECHA. A closer inspection of the extracted terms has shown that the differences between the subcorpora are not due to terminological variation, but rather that MECHA contains more noise than MECHD. While the MECHD term list contains only four lexemes that are not purely technical terms (*displacement*, *nonlinearity*, *causality*, *equation*), the MECHA term list contains nineteen of them (e.g. *repeatability*, *gait*, *harvester*, *covariance*), which amounts to 38% of the top fifty terms. This may be attributed to the fact that scientific papers in the area of mechatronics frequently present results of applied and interdisciplinary research, such as the application of mechatronics in medicine. This finding is further supported by the fact that the majority of the texts included in MECHA are published in the journal *Applied Sciences*, which also explains why medical terms such as *femur* are found among the top fifty terms extracted from MECHA.

5. Discussion

Based on the presented findings, it may be concluded that the inclusion of the academic subcorpus of mechatronics texts into the corpus compiled for term extraction aimed at the corpus-driven systematization of mechatronics terminology would result in introducing too much noise. Thus, we argue that the MECHD subcorpus alone is a better option for this purpose. This conclusion is in line with the corpus design methodology implemented in Cigan's study (2023: 84), which involved a didactic corpus to explore collocations in mechanical engineering texts. On the other hand, Grčić Simeunović (2021: 100) compiled a corpus consisting of academic, didactic and popular science texts to make conclusions about the classificatory role of adjectives in the domain of karstology. However, it must be underlined that this was a study on terminological variation, which has different objectives as compared to terminology systematization studies.

If our corpus were to be used for ESP or EAP teaching purposes (cf. Borucinsky/Kegalj, 2023: 39-40), and where N-grams (cf. Borucinsky/Pritchard, 2022: 7-8;) and Wordlist (cf. Rinder on "frequently encountered lexis", 2017) would be the preferred extraction functions, both corpora could be used independently or combined. Similarly, after extracting the terms for mechatronics terminology systematization purposes, the didactic and the academic subcorpora should be

combined for corpus-based research purposes to allow for the identification of genre-based variation (cf. Halliday, 1992: 84; Webster, 2005: 40; Grčić Simeunović, 2021: 79-87) or syntactic variation (cf. Ibekwe-Sanjuan, 1998: 564).

Following the premise that technical knowledge (Špiranec, 2012: 132) belongs to the overall encyclopedic knowledge (cf. Žic Fuchs, 1991: 6), a mechatronics corpus may be perceived as a knowledge-rich context (Marshman, 2022: 291) that may serve the purpose of not only extracting and analyzing mechatronics terms, but also preparing terminological definitions. In light of this, we argue that the most suitable corpus design methodology involves the inclusion of full textbooks, as concept definitions may be found in all sections of a mechatronics textbook (cf. Pearson, 1998: 1; Łukasik 2014: 77).

One of the limitations of this study is its exclusive focus on single-word terms. Thus, future studies could take multi-word terms into account, which would be useful not only from the terminological standpoint, but also in the context of LSP teaching.

Another limitation of this study is reflected in the fact that it involves only English corpora. There are, however, several reasons for this. As a global *lingua franca,* being perceived as a language of prestige (Crystal, 2003: 126; Bogunović, 2023: 251), English now permeates all principal spheres of human activity (Borucinsky/Bogunović, 2022: 436). A pronounced influx of English elements can be observed across European languages, primarily motivated by a speaker's aim to project topic awareness and credibility (Petrović, 2024: 196). Finally, English has become the language of academia (Nikolić-Hoyt, 2005: 180), which means that it influences other languages from the aspect of their terminological development (Kereković, 2021: 5; Vrgoč, 2023: 98). Thus, we argue that the most suitable approach to the socioterminological systematization of mechatronics terminology in other languages is the contrastive socioterminological approach with English being the starting language.

6. Conclusion

Based on the results of our study, we conclude that there is a significant difference between the term lists extracted from the didactic and the academic subcorpora of mechatronics texts. Furthermore, we conclude that the didactic subcorpus yields more precise results regarding the extraction of mechatronics term candidates. Thus, we advocate that the selection of a smaller, balanced didactic corpus made up of mechatronics textbooks written by acknowledged authors reflects the most suitable corpus design methodology for term extraction aimed at the socioterminological systematization of mechatronics terminology. Bearing in mind that interdisciplinarity is a common trait of most current research, we believe that the findings of this study may be of use for terminologists working on terminology systematization of other emerging and interdisciplinary fields.

7. Acknowledgments

The authors wish to thank Zoran Vrhovski, PhD, Assistant Professor for providing field expert consultation in the corpus selection phase.

8. Bibliography

Anthony, Laurence (2020). Resources for researching vocabulary, in: *The Routledge Handbook of Vocabulary Studies* [ed. Stuart Webb], London / New York: Taylor and Francis, pp. 561–590.

Auslander, David M. (1996). What is mechatronics, in: *IEEE-ASME Transactions on Mechatronics*, 1(1), pp. 5–9.

Bergenholtz Henning / Tarp Sven (1995). *Manual of Specialised Lexicography*. Amsterdam: John Benjamins.

Biber, Douglas (1994). Representativeness in Corpus Design, in: *Current Issues in Computational Linguistics: in Honour of Don Walker* [ed. Antonio Zampolli / Nicoletta Calzolari / Martha Palmer], Pisa: Giardini Editori e Stampatori in Pisa, pp. 377–407.

Biber, Douglas (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers* (*Studies in Corpus Linguistics Issue 23*). Amsterdam / Philadelphia: John Benjamins.

Biber, Douglas (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing, in: *International Journal of Corpus Linguistics*, 14(3), pp. 275–311.

Biber, Douglas (2012). Register as a predictor of linguistic variation, in: *Corpus Linguistics and Linguistic Theory*, 8, pp. 9–37.

Biber, Douglas (^{2}2015). Corpus-Based and Corpus-driven Analyses of Language Variation and Use, in: *The Oxford Handbook of Linguistic Analysis* [ed. Bernd Heine / Heiko Narrog], Oxford: Oxford University Press, pp. 159–192. [first edition 2012].

Bishop, Robert H. (Ed.). (^{2}2008). *The Mechatronics Handbook*. Boca Raton / London / New York / Washington: CRC Press [first edition 2002].

Bogunović, Irena (2023). Engleske riječi u hrvatskome: Jezično posuđivanje i dvojezična leksička obrada, in: *Suvremena lingvistika,* 49 (96), pp. 251-280.

Borucinsky, Mirjana (2023). *Primjena metoda korpusne lingvistike u jezikoslovnim istraživanjima*. Rijeka: Pomorski fakultet Sveučilišta u Rijeci.

Borucinsky, Mirjana / Bogunović, Irena (2022). Crpljenje engleskih riječi iz korpusa hrvatskoga jezika, in: *FLUMINENSIA,* 34 (2), pp. 435–461.

Borucinsky, Mirjana / Kegalj, Jana (2023). Sastavljanje korpusa za poučavanje jezika struke, in: *Proceedings of the 6th International Conference Contemporary Challenges in LSP Teaching* [ed. Brankica Bošnjak Terzić / Snježana Kereković / Mirna Varga], Zagreb: The Association of LSP Teachers at Higher Education Institutions / Faculty of Humanities and Social Sciences, University of Zagreb, pp. 34–46.

Borucinsky, Mirjana / Pritchard, Boris (2022). Lexical bundles in maritime texts, in: *ICAME Journal*, 46(1), pp. 5-17.

Bowker, Lynne (2003). Specialised lexicography and Specialised Dictionaries, in: *A Practical Guide to Lexicography* [ed. Piet van Starkenburg], Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 154–164.

Bowker, Lynne / Pearson, Jennifer (2002). *Working with Specialized Language*. London / New York: Routledge.

Bugarski, Ranko (1986a). *Jezik u društvu*. Beograd: Prosveta.

Bugarski, Ranko (1986b). *Lingvistika u primeni*. Beograd: Tumačenje književnosti.

Cabré, M. Teresa (1999). *Terminology: theory, methods, and applications*. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Cigan, Vesna (2023). *Kolokacijski odnosi u engleskome, njemačkome i hrvatskome strojarskom strukovnom jeziku*. Zagreb: Faculty of Humanities and Social Sciences, University of Zagreb. [Doctoral Thesis].

Cintra Faria, Ana Carolina / Barbalho, Sanderson César Macêdo (2023). Mechatronics: A Study on Its Scientific Constitution and Association with Innovative Products, in: *Applied System Innovation, 6(4)*:72, pp. 1-42.

Corpas, Gloria, & Seghiri, Miriam (2009). Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish), in: *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate* [ed. Allison Beeby / Patricia Rodríguez-Inés / Pilar Sánchez-Gijón], Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 75–107.

Crystal, D. ([2]2003). *English as a global language*, Cambridge / New York / Melbourne / Madrid / Cape Town / Singapore / São Paulo: Cambridge University Press [first edition 1997].

Delavigne, Valérie & Gaudin, François (2022). Founding principles of Socioterminology, in: *Theoretical Perspectives on Terminology. Explaining terms, concepts and specialized knowledge* [ed. Pamela Faber / Marie-Claude L'Homme], Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 177–195.

Diki-Kidiri, Marcel (2020). La terminologie culturelle, une branche majeure de la linguistique de développement, in: *Linguistique du développement: Prolégomènes à un champ disciplinaire émergent. Revue Mosaïques, Hors-série numéro 6: Mélanges en hommage au Professeur Henry Tourneux* [ed. Mahamat Adam], Paris: Editions des archives contemporaines, pp. 53-62.

Faber Benítez, Pamela (2009). The Cognitive Shift in Terminology and Specialized Translation, in: *MonTI,* 1, pp. 107–134.

Felber, Helmut (1984). *Terminology Manual*. Paris: UNESCO – INFOTERM.

Grčić Simeunović, Larisa / Frleta, Tomislav (2012). Kriteriji za prevođenje složenih naziva i kolokacija na hrvatski jezik (pravna terminologija EU-a), in: *Aktualna istraživanja u primijenjenoj lingvistici* [ed. Leonard Pon / Vladimir Karabalić / Sanja Cimer], Osijek: Hrvatsko društvo za primijenjenu lingvistiku, pp. 231–244.

Grčić Simeunović, Larisa / Stepišnik, Uroš / Vintar, Špela (2020). Klasifikacijska uloga pridjeva u području geomorfologije krša, in: *Rasprave Instituta za hrvatski jezik*, 46(2), pp. 619-633.

Gries, Stefan. (2010). Corpus linguistics and theoretical linguistics: A love–hate relationship? Not necessarily, in: *International Journal of Corpus Linguistics*, 15(3), pp. 327-343.

Gries, Stefan Th. ([2]2017). *Quantitative corpus linguistics with R: A practical introduction*. New York / London: Routledge Taylor & Francis Group [first edition 2009].

Grimheden, Martin / Hanson, Mats (2005). Mechatronics: the Evolution of an Academic Discipline in Engineering Education, in: *Mechatronics,* 15(2), pp. 179–192.

Guespin, Louis (1995). La circulation terminologique et les rapports entre science, technique et production, in: *Meta: Translators' Journal,* 40, pp. 206–215.

Halliday, Michael Alexander Kirkwood (1992). Language as system and language as instance: the corpus as a theoretical construct, in: *Computational and Quantitative Studies. Volume 6 in the Collected Works of M.A.K. Halliday* (2005) [ed. Jonathan J. Webster], London / New York: Continuum, pp. 76-92.

Hudeček, Lana / Mihaljević, Milica (2012). *Hrvatski terminološki priručnik*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Humbley, John (2022). The reception of Wüster's General Theory of Terminology, in: *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge* [ed. Pamela Faber / Marie-Claude L'Homme], Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 15-36.

Hyland, Ken (2008). As can be seen: Lexical bundles and disciplinary variation, in: *English for Specific Purposes*, 27(1), pp. 4–21.

Ibekwe-SanJuan, Fidelia (1998). Terminological Variation, a Means of Identifying Research Topics from Texts, in: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, Montreal: Association for Computational Linguistics, pp. 564–570.

Jakubíček, Miloš / Kilgarriff, Adam / Kovář, Vojtěch / Rychlý, Pavel / Suchomel, Vít (2013). The TenTen Corpus Family, in: *Proceedings of the 7th International Conference on Corpus Linguistics CL 2013* [ed. Andrew Hardie / Robbie Love], Lancaster: UCREL, pp. 125-127.

Jouaneh, Musa (2013). *Fundamentals of Mechatronics*. Stamford: Cengage Learning.

Kereković, Snježana (2021, October 28). Može li hrvatsko tehničko nazivlje preživjeti? [Oral presentation with published abstract], in: *3. terminološki okrugli stol: Hrvatska terminologija u europskome kontekstu: stanje i perspektive*, Zagreb, Croatia.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2004). The Sketch Engine: ten years on, in: *Lexicography*, 1, pp. 7–36.

Kurosawa, Toyoki (1983). Development of Mechatronics Technology and Future Needs, in: *Journal of the Japan Society of Precision Engineering,* 49(11), pp. 1475–1480.

Leech, Geoffrey (1991). The State of the Art in Corpus Linguistics, in: *English Corpus Linguistics: Studies in Honor of Jan Svartvik* [ed. Karin Aijmer / Bengt Altenberg], New York / London: Routledge Taylor & Francis Group, pp. 8-29.

L'Homme, Marie-Claude (2006). A Look at some Canadian Contributions to Terminology, in: *Modern approaches to Terminological Theories and Applications*

[ed. Heribert Picht], Bern / Berlin / Bruxelles / Frankfurt am Main / New York / Oxford / Wien: Peter Lang, pp. 55–75.

L'Homme, Marie-Claude (2022). Terminology and Lexical Semantics, in: *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge* [ed. Pamela Faber / Marie-Claude L'Homme], Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 237–259.

Łukasik, Marek (2014). Compiling a Corpus for Terminographic Purposes, in: *Komunikacja Specjalistyczna*, 7, pp. 71-83.

Marshman, Elizabeth (2022). Knowledge patterns in corpora, in: *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge* [ed. Pamela Faber / Marie-Claude L'Homme], Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 291-310.

MDPI (2025). Available online: https://www.mdpi.com/ (accessed on: 25 May 2025).

Meyer, Ingrid (2001). Extracting knowledge-rich contexts for terminography. A conceptual and methodological framework, in: *Recent Advances in Computational Terminology* [ed. Didier Bourigault / Christian Jacquemin / Marie-Claude L'Homme], Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 279–302.

Meyer, Ingrid (2022). Concept management for Terminology: A Knowledge Engineering approach, in: *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge* [ed. Pamela Faber / Marie-Claude L'Homme], Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 111-126.

Mihaljević, Josip (2021). *Konceptualni okvir igrifikacije hrvatskoga mrežnoga rječnika*. Zagreb: Faculty of Humanities and Social Sciences, University of Zagreb. [Doctoral Thesis].

Nahod, Bruno (2020). O nazivnosti pridjeva: komparativna analiza pridjeva u Struni i rječnicima općega jezika, in: *Svijet od riječi – Terminološki i leksikografski ogledi* [ed. Ana Ostroški Anić / Ivana Brač], Zagreb: Institut za hrvatski jezik i jezikoslovlje, pp. 185-197.

Nikolić-Hoyt, Anja (2005). Hrvatski u dodiru s engleskim jezikom, in: *Hrvatski jezik u dodiru s europskim jezicima: Prilagodba posuđenica* [ed. Lelija Sočanac / Orsolya Žagar-Szentesi / Dragica Dragičević / Ljuba Dabo-Denegri / Antica Menac / Anja Nikolić-Hoyt], Zagreb: Nakladni zavod Globus, pp. 179-205.

Pearson, Jennifer (1998). *Terms in Context*. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Petrović, Bernardina (2024). *Leksička i kontekstna sinonimija u hrvatskome jeziku*. Zagreb: Hrvatska sveučilišna naklada.

Pritchard, Boris (2015). On multiword lexical units and their role in maritime dictionaries, in: *Iranian Journal of English for Academic Purposes*, 1(4), pp. 40–64.

Purković, Damir / Salopek, Goran (2015). *Osnove mehatronike: Za početno učenje i buduće nastavnike*. Rijeka: Sveučilište u Rijeci.

Rinder, Jamie (2017, June 28-30). Vocabulary and LSP for Global Engineers [Oral presentation with published abstract], in: *21st Conference on Language for Specific Purposes*, Bergen, Norway.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford / New York / Toronto: Oxford University Press.

Stefanowitsch, Anatol (2020). *Corpus linguistics: A Guide to the methodology*. Berlin: Language Science Press.

Špiranec, Ivana (2013). Međudjelovanje jezične i izvanjezične informacije u jeziku struke, in: *Jezik kao informacija - Zbornik radova s međunarodnog skupa Hrvatskog društva za primijenjenu lingvistiku* [ed. Anita Peti-Stantić / Mateusz-Milan Stanojević], Zagreb: Srednja Europa / Croatian Applied Linguistics Society, pp. 131–142.

Temmerman, Rita (2000). *Towards New Ways of Terminology Description. The Sociocognitive Approach.* Amsterdam / Philadelphia: John Benjamins Publishing Company.

Teubert, Wolfgang (1996). Comparable or Parallel Corpora?, in: *International Journal of Lexicography,* 9(3), pp. 238–264.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Tomizuka, Masayoshi (2000). Mechatronics: From The 20th to 21st Century, in: *IFAC Proceedings Volumes,* 33(26), pp. 1-10.

Trojar, Mitja (2017). Wüster's View of Terminology, in: *Slovenski jezik – Slovene Linguistic Studies,* 11, pp. 55–85.

Tummers, Jose / Kris, Heylen / Geeraerts, Dirk (2005). Usage-based approaches in cognitive linguistics: a technical state of the art, in: *Corpus Linguistics and Linguistic Theory*, 1(2), pp. 225–261.

Vrgoč, Dalibor (2021). *Terminološki aspekti stvaranja hrvatskoga vojnoga nazivlja*. Zagreb: Faculty of Humanities and Social Sciences, University of Zagreb. [Doctoral Thesis].

Vrgoč, Dalibor (2023). Frontal Attack or Retrograde: Croatian Military Terminology Confronting Anglo-American Terminological Influx, in: *Collegium antropologicum,* 47(2), pp. 91–99.

Webster, Jonathan J. (Ed.) (2005). *Computational and Quantitative Studies. Volume 6 in the Collected Works of M.A.K. Halliday*. London / New York: Continuum.

Wüster, Eugen (1968). *The Machine Tool. An Interlingual Dictionary of Basic Concepts comprising An Alphabetical Dictionary and A Classified Vocabulary with Definitions and Illustrations*. London: Technical Press.

Wüster, Eugen (1974). Die allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften, in: *Linguistics*, 119(1), pp. 61–106.

Wüster, Eugen ([3]1991). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*, Bonn: Romanistischer Verlag [first edition 1979].

Yaskawa Global (2025). Available online: https://www.yaskawa-global.com/ (accessed on: 16 June 2025).

Zanettin, Federico (1998). Bilingual Comparable Corpora and the Training of Translators, in: *Meta,* 43(4), pp. 616–630.

Žic Fuchs, Milena (1991). *Znanje o jeziku i znanje o svijetu*. Zagreb: SOL.

Sastavljanje korpusa za izlučivanje naziva s ciljem socioterminološkoga usustavljivanja mehatroničkoga nazivlja

U ovome radu istražuje se metodološki okvir za sastavljanje specijaliziranoga korpusa za izlučivanje naziva s ciljem optimizacije socioterminološkoga usustavljivanja mehatroničkoga nazivlja. Primarni je cilj ovoga istraživanja razviti empirijski utemeljene strategije sastavljanja korpusa kako bi se povećala pouzdanost i učinkovitost izlučivanja naziva iz korpusa. Pošli smo od pretpostavke da pristup vođen korpusom, kada se primijeni isključivo na didaktički potkorpus (udžbenike i priručnike iz područja mehatronike), omogućava preciznije izlučivanje potencijalnih mehatroničkih naziva od pristupa koji uključuje akademski potkorpus (znanstvene članke iz područja mehatronike). S ciljem testiranja spomenute hipoteze, sastavljena su dva potkorpusa mehatroničkih tekstova na engleskome jeziku (didaktički i akademski). Nazivi su izlučeni iz oba potkorpusa mehatroničkih tekstova, te su uspoređeni s rezultatima koji su izlučeni pomoću drugih funkcija u jezičnokorpusnome alatu Sketch Engine. Komparativna statistička analiza pokazuje da didaktički potkorpus omogućava izlučivanje većega broja potencijalnih mehatroničkih naziva, dok akademski potkorpus uvodi veću količinu šuma. Stoga, u radu se zastupa stav da je manji, uravnoteženi didaktički korpus prikladniji za izlučivanje naziva u okviru socioterminološkoga usustavljivanja mehatroničkoga nazivlja od većega korpusa koji uključuje i akademski potkorpus. Doprinos opisane metodologije ogleda se u većoj učinkovitosti kod socioterminološkoga usustavljivanja mehatroničkoga nazivlja, a može se primijeniti i na druga interdisciplinarna područja.

Ključne riječi: pristup vođen korpusom, mehatronika, socioterminologija, izlučivanje naziva, usustavljivanje nazivlja