Applied Linguistics Papers 27/4: 2023, 34–45

Corpus linguistics and generative AI tools in term extraction: a case of Kashubian – a low-resource language

Marek ŁUKASIK
Uniwersytet Pomorski w Słupsku/ Pomeranian University in Słupsk
E-mail: marek.lukasik@upsl.edu.pl,

Abstract: Electronic corpora have been an indispensable resource in a variety of language studies, including linguistics, lexicography or terminology. Provided that they are compiled in a systematic manner, such text collections can provide high quality data that can be readily used in a specific study or can be directly applied to a specific practical project. However, the creation of a usable corpus depends on the availability and the quality of source texts and the tools that are used for its processing. Another factor that often plays a significant role in successful *ad hoc* applications of corpora is their immediate accessibility. Recent developments in generative artificial intelligence (GenAI) have rendered the idea of instantaneous access to language data a feasible possibility. This paper discusses the results of a study into the feasibility of applying modern corpus and GenAI tools in the extraction of biological terminology in Kashubian, a regional language spoken in the north-central part of Poland (Kashubia). The overarching goal was to identify modern and effective tools that could be used by terminologists, lexicographers, translators, and teachers of Kashubian.

Keywords: artificial intelligence, corpus linguistics, Large Language Models, low-density language, Sketch Engine, minority language, terminology, terminography.

Introduction

Kashubian is a regional language spoken in Kashubia, a region in north central part of Poland, within the eastern boundaries of Pomerania region. It acquired its status as a regional language in 2005 by way of the Act of 6 January 2005 on National and Ethnic Minorities and on the Regional Language (2005 Act) [1]. According to the 2021 census data, the language is spoken by approx. 87,600 people[2], which renders it a lesser-used or a minority language. Woźniakiewicz emphasises that Kashubian belongs to endangered languages (also according to the UNESCO list of endangered language), mainly because of the continually decreasing number of speakers (Woźniakiewicz 2013: 13). Also, the language is constantly *in statu nascendi*, with its literary variety undergoing standardisation undertaken by the Kashubian Language Council, and more specifically by its Standardisation and Normalisation Committee. The process is essential from the educational perspective: since Kashubian exhibits some extensive multi-level differences, specifically observed in the varieties of the vernacular used in some (distant) parts of the region, it is

[1] There is no consensus as to the status of Kashubian as a language, with some linguists more inclined to define it as a dialect of Polish (see Jabłoński 2019: 49–50).
[2] The numbers vary significantly across the literature.

all the more so important to teach the standardised variety at schools (the language is taught is some institutions in the Kashubia region). Needless to say that also scientific texts written in Kashubian require terminological standardisation. However, before any form of standardisation is possible, it is necessary for the language to develop its specialised vocabulary. As a matter of fact, this problem concerns most of the expert fields, mainly because, historically, Kashubians used to deal with fishing and only a few other more traditional crafts. Despite the obstacles, Zieniukowa emphasises the need to introduce Kashubian to scientific communication in order to save the endangered language and to elevate its prestige (Zieniukowa 2015: 238).

Some standardisation efforts in the area of terminology are undertaken by the Kashubian Language Council, who publish the results of their recommendations in the *Bulletin of the Kashubian Language Council* (pol. *Biuletyn Rady Języka Kaszubskiego*; csb. *Biuletin Radzëznë Kaszëbsczégò Jãzëka*). The Council has published bilingual (Polish-Kashubian) lists of standardised terminology of mathematics, information technology, linguistics (grammar), mass media, agriculture, religion, film, music, theatre, geography, biology, and history, among others. The vocabulary lists are not comprehensive, yet still very useful in the educational context. For example, the terminology of biology covers approximately 430 terms (Biuletyn Rady Języka Kaszubskiego 2018: 47–60), which probably covers the core terminology set encountered by students in primary school education[3]. It is worth mentioning that despite the efforts undertaken by the Council, in a number of fields Kashubian terminology still exhibits orthographic/phonetic variations or is non-existent altogether.

Standardisation work in terminology often proceeds from existing vocabulary stock found in thematically-related texts. Therefore, text repositories and collections such as corpora are particularly useful in the extraction of term-candidates (UNESCO/ InfoTerm 2005: 10). These are later discussed upon by a terminological commission and either adopted as terms proper, adapted (e.g. orthographically) to the currently-binding language/commission rules (and then adopted), or discarded. According to M.T. Cabré

> [c]omputer-aided text analysis and the possibility of processing large amounts of information have changed the bases of terminology compilation, as well as how the appropriateness of terms is conceived, and the degree of human intervention in the whole work process (Cabré 1999: 163).

Until recently, the text corpus was one of the most reliable and methodologically-sound resources in the extraction of terms or term-candidates. The systematicity of the method surpassed the simplistic (and sometimes rather chaotic) search through the Web. However, the advent of widely available and ever better AI tools has changed the manner

[3] An initiative that deserves attention is a series of the *Kashubian Thematic Encyclopaedia* (*Kaszëbskô Encyklopediô Tematicznô*), whose individual volumes come with an encyclopaedic part (in Kashubian), a short glossary of terms with their definitions (in Kashubian) and a Kashubian-Polish list of terms. Unfortunately, only two volumes were published to date: Chemistry (2013) and Biology (2018). The bilingual term lists consist of 290 (plus names of elements in the periodic table added as an insert), and 571 terms, respectively (see Nacel 2013 and Nacel/ Jeliński 2018).

in which information is retrieved. Known as artificial intelligence-based information access systems, such as Microsoft's Bing/ChatGPT, Google's Bard and Meta's LLaMA, the tools "upend the traditional search engine mode of search input and output" (Shah 2023), making Web search results more individualised, tailored to the needs of the user, and under controlled conditions – possibly more systematic. The advantage of the Web as a resource lies in the fact that even in the case of such a low-resource language as Kashubian, the user should be able to access a considerable body of texts.

This paper focuses on the comparison of the effectiveness of term extraction with the application of two different methods: one based on text corpora and Sketch Engine corpus tools, and the other relying on an AI-powered chatbot – ChatGPT. In particular, the study focuses on the extraction of Kashubian terminology of biology from the Web, the rationale being that such searches are necessary in standardisation to analyse the current terminological stock in use and to attest the use of term-candidates. It is believed that the methods can also prove useful in the educational context, lexicography and translation.

1. Text corpora, corpus tools and GenAI

It has been sixty years since the first ever electronic corpus was used for linguistic investigation and lexicography. From that moment on, corpora have been widely used in a number of research fields and found numerous practical applications. Depending on the type of study/task, different types of corpora are used. However, facing lack of a specific, ready-to-use type of corpus, researchers compile their own resources. Such do-it-yourself (DIY) corpora can be created manually, by collecting specific texts in the electronic form, or semi-automatically, by employing automatic text harvesters, which download texts from the Web on the basis of keywords (often called seeds), or through URLs of specific websites or webpages (the process is generically known as web scraping). In the latter method, the time of the compilation procedure is significantly reduced (see Baroni et al. 2006). Corpus data may be used directly in quantitative and qualitative studies, or indirectly, requiring further analyses or expert decisions as to the validity of the data obtained. In either case, the quality of the output data will depend on the quality of texts comprising the corpus (Batini/ Scannapieco 2006: 161–167; cf. Brezina 2018: 262).

General corpora are created with a view of mirroring the entirety of texts produced in a specific language. The quality of the data is ensured by strictly following a specific corpus design, whose exact specifications conform to the notions of representativeness and balance (see McEnery/ Xiao/ Tono 2005: 13–21). Such repositories are useful in a number of different studies, however they also constitute a unique language 'vault', whose contents reflect language at a given particular stage of its development. Therefore, a corpus acts here as a specific attestation tool for all users seeking answers to questions about the actual language in use

The pressing necessity to create corpus resources is connected, *inter alia*, with the educational needs. Such needs become apparent in the case of low-density languages, such as Kashubian. D. Stanulewicz is one of the researchers who suggested a set of parameters of a corpus of Kashubian Language, which, in her view, should follow the design of the demo version of the Polish National Corpus (Stanulewicz 2015: 389). However, the choice of sources included in the first edition of the corpus would be limited to texts that are available in the digital form and available from digital libraries or are already in

the public domain (Pomierska/ Stanulewicz 2019: 165). The results of a preliminary study on the compiled test corpus were presented, *inter alia*, at the Symposium of the Kashubian Language Council on 25 August 2023. The corpus, however, is not widely available.

As a matter of fact, there is no universally available corpus of Kashubian, with the exception of the Kashubian Wikipedia Corpus[4] (approx. 126,000 tokens) available from the Corpora Collection Leipzig. The online resource allows the user to browse through the collection, however the query has to match the exact word-form (the corpus is not lemmatised). The output results include basic quantitative data (total number of occurrences, rank, and frequency class), illustrative sentences in which the query word or phrase occurs, and words that cooccur with the node (collocates). In this study, the corpus will not be analysed, mainly due to its limited scope and usability of the search tool.

In the current analysis, the tool that proved to be the most universal was Sketch Engine, a corpus manager and analysis platform. The platform offers manual as well as semi-automatic corpus compilation, which was in line with one of the prerequisites of the study, namely that the tool should provide the possibility of extracting terms from a text collection harvested (scrapped) from the Web. The main rationale behind this requirement was that both tools compared in this study should have similar functionality in this respect.

As of 2023, Sketch Engine comes with more than 700 ready-to-use (preloaded) corpora, representing more than 100 languages. However, there is no preloaded Kashubian corpus on the platform. The platform comes with a number of tools, such as Wordlist, N-grams, Concordance, Thesaurus, Keywords, Word Sketch, and Word Sketch Difference, among others (Kilgarriff et al. 2014). Yet, some of the tools are not active when a corpus is not lemmatised (lemmatisation is not supported for all languages) and/or there is no POS-annotation (see below). From the perspective of this study, the Keywords (and terminology extraction) tool is of particular importance, however the tool is only partially active for the reasons mentioned above. Only one-word Keywords and N-Grams tools are available in the Keywords tool.

With the advent of universally available Generative Artificial Intelligence, or GenAI (see Epstein et al. 2023), the toolkit available to language experts has considerably widened. The very nature of some GenAI tools, that is their reliance on large textual sets, or corpora, which are utilised in large language models (LLMs), and/ or the fact that the input (prompt) from the user is formulated in a natural language, makes the technology somewhat familiar, especially for those working with language engineering systems. However, GenAI tools excel in being versatile and simple to use. In this study, an AI-powered chatbot, namely ChatGPT (Chat Generative Pre-trained Transformer) using GPT-4 model, is used to perform – in a single prompt-instance – a fairly complex sequence of terminology-work-related tasks that would otherwise take a few dozen minutes, if not hours[5]. Yet, the efficiency of the tool depends on prompt quality, which in practical terms translate into the provision of a necessary context for an AI tool to execute a task (more on prompt engineering in Ziegler/ Berryman 2023).

[4] URL: https://corpora.uni-leipzig.de/en?corpusId=csb_wikipedia_2021 [Accessed: 20.11.2023].
[5] For other uses of ChatGPT see, for example Nowakowski et al. 2022; Sarrion 2023; Lew 2023.

3. Term extraction: Sketch Engine vs. ChatGPT

The aim of this study is to test the effectiveness of two different tools in the extraction of terms from Web texts representing a low-resource language, namely Kashubian. The two tools represent two distinct kinds of digital technology, which renders the analysis unique and require a different methodological approach to each tool.

In the case of Sketch Engine, the following steps were taken:

- selection of keywords ('seeds') for the web scraping tool (harvester),
- analysis of the quality of the URLs downloaded by the tool,
- generation of Keywords list (=term extraction procedure),
- further analyses using the Concordance tool, if necessary,
- quality test (against external resources).

In the case of ChatGPT (only GPT-4 model was used), the analysis proceeded in the following manner:

- drafting of an appropriate prompt,
- analysis of the output,
- prompt refinement,
- analysis of a refined output,
- quality test (against external resources).

In the following two sections, the exact procedure as well as the results obtained are presented.

3.1. Sketch Engine

1. With the prerequisite to use Web-based text for term extraction procedure, the seed-based web scraping method was applied. In this method, the harvester uses three distinct seeds (keywords), feeds a combination of the items to the search engine (Bing), and downloads URLs which should contain the lexical items. The following words were used as seeds: *biologiô* (eng. *biology*), *biôłtkò* (eng. *protein*), *mùtacjô* (eng. *mutation*), *metabolism* (eng. *metabolism*), and *kòmórka* (eng. *cell*). Interestingly enough, the corpus creator allows the user to select the language of the corpus, which in this case was Kashubian, however the setting is only used for corpus identification (management) purposes.

2. Sketch Engine does not feature a lemmatiser or a POS-tagger for Kashubian, and therefore the term extraction tool (Keywords) was only partially available (for one-word keywords and n-grams). Word Sketches were not available, either. Accordingly, the identification of terms proceeded in a manual manner.

3. In the initial analysis, with the default settings of the tool, the harvester returned 191 URLs, some of which did not even contain any text in Kashubian. A number of webpages were in Czech, Portuguese (.br domain) and Polish, the possible reason being that current search engines (including Bing) provide cross-lingual search as a result of the advancement and the implementation of natural language understanding (NLU) and semantic search algorithms. At this stage, no intervention was undertaken and the corpus compilation was allowed 'as is', producing a collection of 1,117,807 words.

4. In order for the Keyword tool to properly identify key lexical items, a reference corpus of Kashubian texts based on Wikipedia pages in Kashubian was created (325,391 words).

5. The results obtained from the Keyword tool yielded poor results, with no Kashubian terms related to biology in the first 100 words of the rank list, and only a few found among the first 500 words: *chòroba* (rank: 281; eng. *illness*); *òkò* (rank: 307; eng. *eye*); *oczë* (rank: 481; eng. *eyes*). A number of words in the Keyword list were general words in Kashubian (approx. 45%), but also in Polish (25%) and Portuguese (5%). The remaining 25% were abbreviations (see discussion below) and noise.

6. With the results not meeting the basic requirements of the study, a refined Web scraping procedure was implemented by (1) allowing the search engine to suggest 'more relevant' webpages, and (2) manually excluding URLs of websites that did not meet the language requirement.

7. In the refined search, the harvester returned 193 URLs, however, after manual quality check, a total of 124 URLs were selected for text scraping. The corpus created (1,124,099 words) was then analysed. The term extraction results were similar to those obtained in the previous attempt, and yielded the following results: *nosa* (rank: 209; eng: *nose*$_{\text{gen}}$) *noga* (rank: 194; eng. *leg*); *òkò* (rank: 266; eng. *eye*); *oczë* (rank: 423; eng. *eyes*); *rãką* (rank: 425; eng. *hand*$_{\text{abl}}$). The Keyword list was more 'Kashubian-focused', with more than 60% of words being Kashubian, 20% Polish, and the remaining 20% being abbreviations and noise.

8. A detailed analysis (=quality test) of the results obtained confirm that the terms identified do belong to the biological sciences. However, in the corpus, they did not occur in biological texts, but rather were part of dictionary definitions or general language texts. The Web scraping tool downloaded an electronic preprint of Eugeniusz Gołąbek's *Polish-Kashubian Dictionary* (cbs. *Słowôrz Pòlskò-kaszëbsczi*), and this is the reason why there were so many abbreviations (=dictionary labels) and general words in the list. Overall, in spite of rectification of the corpus compilation procedure, the results were unsatisfactory and therefore the method adopted cannot be considered viable in terminology work with low-resource languages. Also, the entire procedure proved time consuming (see below). Interestingly enough, the seed words were present in the frequency list (Wordlist tool), and with their inflected forms, they occurred with the following frequencies (the frequencies of all word-forms were added): *kòmórka* (AF=27), *biôłtkò* (AF=19), *mùtacjô* (AF=9), *biologiô* (AF=7), *metabòlizm* (AF=2). Also other more 'technical' terms were found in the list (e.g. *jednokòmórkòwi*, AF=2; *midzëkòmórkòwi* AF=1; *midzëkòmórkòwé*, AF=1, among others) however their overall frequency was rather low and they all were extracted from one resource: the above mentioned *Dictionary* by E. Gołąbek.

3.2 ChatGPT

1. For the comparability of the procedure, the AI model was explicitly instructed to use Web resources. Only GPT-4 model (available through a premium account) is capable of directly accessing the Web via Bing search engine (OpenAI GPT-4 2023). The following text was used as a prompt: "Download (from the Web) at least 25 terms in the Kashubian language related to biology. List the terms in a table where column 1 is for

terms in English (listed alphabetically), column 2 is for the Kashubian term and column 3 is for additional comments, mainly related to any orthographic variants of the Kashubian terms found."

2. Once the prompt was entered, the tool indicated the following Bing search queries and the websites visited: "Kashubian language biology term list", "Kashubian language biology terms list", "Kashubian language biology vocabulary list", "Kashubian language biology terms", "Kashubian biology vocabulary", "Kashubian biology terminology", "Kashubian biological terms list", and lexilogos.com, zawszepomorze.pl, omniglot.com, languagelist.org, en.wiktionary.org, en.wikipedia.org, cassubia-dictionary.com.

3. The exact output of the tool is provided below (asterisks and bold font highlights are mine – M.Ł.):

"I have started compiling a list of Kashubian terms related to biology, but I was unable to complete the list of 25 terms as per your request. Here is what I have managed to gather so far:

English Term	Kashubian Term	Comments
Animal	**zwiérz**	-
Beak	dzëba	-
Bulb	**bùlwa**	-
*Ear	czéwa	-
Fern	**roscëna**	-
*Grass	chróst	-
*Leaf	papla	-
*Mushroom	pãpk	-
Tail	ògón	-
Wing	skrzidło	-

This list includes a variety of terms related to different aspects of biology, such as animal body parts, plants, and fungi. However, due to time constraints, I couldn't complete the list to reach the requested number of 25 terms. Additionally, there were no specific orthographic variants noted for the terms I found.
To complete the list, you might consider visiting online resources like Wiktionary's Kashubian category or specialized Kashubian language websites."

4. Following the output provided, an additional prompt ("Please visit the websites for me and finish the list.") was fed into the chatbot, producing the following output:

"Here is the completed list of 25 biology-related terms in Kashubian, along with their English equivalents and comments where applicable:

English Term	Kashubian Term	Comments
Animal	**zwiérz**	-
Bark (of a tree)	kóra	-
Beak	dzëba	-
Bird	**ptôch**	-
Blood	krew	-
Bone	**gnôt**, kòsc	Two variants
Breast	piers	-
Dog	pies	-
Ear	ùchò	-
Egg	jaje, *jajkò, jajo	Three variants
Eye	òkò	-
Feather	pióro	-
Fish	rëba	-
Flower	kwiat	-
Fruit	**brzôd**	-
Grass	**trôwa**	-
Hair	włosë	-
Heart	serce	-
Leaf	**lëst**	-
Leg	noga	-
*Mouth	gãba	-
Nose	nos	-
Seed	**semiã**	-
Skin	skóra	-
Tail	ògón	-
Tree	drzewò	-
Wing	skrzidło	-

This list includes terms for various animals, plant parts, and body parts, reflecting a diverse range of biological vocabulary in Kashubian. The sources for this list are Wiktionary's categories on Kashubian animal body parts, plants, and a comprehensive Swadesh list for the language [Link 1][6] [Link 2][7] [Link 3][8]"

5. The output generated by ChatGPT seems to be coherent with the instructions, however an additional prompt was necessary to acquire a list of 25 terms related to biology. It is worth noticing that the user obtains a complete list of term-candidates or terms, which

[6] Kashubian terms for types or instances of body parts of animals (Wiktionary) (URL: https://tinyurl.com/czb5av82) [Accessed 25.11.2023].
[7] Kashubian terms for types of plant (Wiktionary) (URL: https://tinyurl.com/4kmbu78b) [Accessed 25.11.2023].
[8] Kashubian Swadesh list (Wiktionary). (URL: https://tinyurl.com/3hd6p37j [Accessed 25.11.2023].

may constitute a basis for further terminology work. This can be facilitated by the URLs included in the output. In the above tables, lexical items in bold font are the ones found in the official standardised list of biological terms drafted by the Kashubian Language Council (see Biuletyn Rady Języka Kaszubskiego 2018) or in the *Kashubian Thematic Encyclopaedia: Biology* (see Nacel/ Jeliński 2018). The remaining lexical items are valid Kashubian words, as verified against the *Polish-Kashubian Dictionary* by Eugeniusz Gołąbek (see Gołąbek 2012, 2013). The main inconsistencies concern the English equivalents: '*czéwa*' (eng. *gills*), '*chrósť*' (eng. *brush*), '*papla*' (eng. *poplar*), '*pãpk*' (eng. *type of mushroom*; lat. *Suillus luteus*), and *gãba* (eng. *face*, although the Kashubian word is rather colloquial). The only inconsistency found among Kashubian lexical items was one of the variants of the term '*egg*', namely '**jajkò*',[9] whose correct spelling is '*jôjkò*'.

4. Discussion

A quick overview of the term extraction results obtained from the two tools suggests that the AI-powered chatbot has an advantage over the corpus linguistics tool, especially if the researchers employ automatic or semi-automatic compilation methods in Sketch Engine. The advantage concerns both the number of terms or term-candidates generated, the form of data presentation, and the possibility of quick refinement of the output (through prompt engineering).

Another point in discussion is the time needed to complete the complex tasks. In the case of corpus tools, semi-manual or manual methods had to be employed in the case of: (a) the extraction of seeds from reliable sources; (b) the application of specific settings for Sketch Engine's harvester; (c) the verification of URLs, and, most importantly; (4) the analysis of Keyword lists and Wordlists in the search of terms or term candidates. Each complete cycle lasted up to one hour. Meanwhile, the generation of a list of term candidates by ChatGPT took less than a minute[10].

Nevertheless, corpus methods may still be helpful in the study of low-resource languages (see Gaeta 2022), but the scarcity of online resources calls for a manual text collection and manual verification, while data quality can be achieved through cyclical corpus compilation procedure (see Łukasik 2014: 79). Another issue is the lack of lemmatisation or POS-tagging, which additionally limits the usefulness of some tools. Accordingly, the first step would be to develop a lemmatiser and a POS-tagger for standard Kashubian[11].

There are also some inherent issues with ChatGPT, though. First, the terms or term-candidates obtained are rather general in nature and do not represent highly specialised concepts related to the field of biology (these could be found, albeit manually, in corpus frequency lists). Second, the tool did not crawl deep through the websites to find new lexical items (this has been verified by the author on the basis of Wiktionary entries). Third, there are some in-built limitations of the tool, one of which is the timeout problem (an automatic cancellation or cessation of a task after a predefined interval of time has passed), while another is the output character/ token limit. Therefore, a generation of longer lists of terms or

[9] The term in the spelling provided by ChatGPT was found to be part of the Swadesh list.
[10] This may not be consistent as the performance of the tool is highly dependent on current demand.
[11] See a discussion on the issues connected with POS tagging in inflected languages (Kobyliński/ Krasnowska-Kieraś 2018).

term candidates requires re-prompting, or 'reminding'/'encouraging' the chatbot to continue with the task. However, there is a limit here, since ChatGPT has a 'memory span', also known as a context window, counted in tokens (which are not necessarily words) beyond which the tool tends to 'forget' the instructions or the text it generated (see Ratner et al. 2023). One other issue is the default lack of personalisation: the chatbot does not learn a user's behaviour and has no access to prompts in other chat instances. Also, the output may be different even if the same prompt is used in the same chat (the user can also request the regeneration of the output). One method to address the issue is a careful customisation of the chatbot (available since July 2023 through the 'Create a GPT' option).

More importantly, GenAI models tend to hallucinate and make reasoning errors: "AI hallucination is a phenomenon wherein a large language model (LLM)—often a generative AI chatbot or computer vision tool—perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate" (IBM). Accordingly, from the scientific method perspective, all outputs require some form of information veracity test. This applies to all models, including GPT-4 one, which, admittedly, has the lowest, but non-zero hallucination rate (Wodecki 2023). Finally, unless instructed to do so, the tool usually does not provide references to the resources that were used to generate the output (it certainly does not happen when the model uses its own repository). Also, despite using Web resources, the chatbot still generates the output text, which may lead to factual mistakes and errors.

It needs to be emphasised that since the two tools are distant in terms of the technology used and the manner of producing output data, mainly on account of the impossibility to achieve the comparability and consistency of data, the present study lacks the necessary level of methodological rigour. Accordingly, the results should be viewed from a qualitative perspective.

5. Conclusion

One of the major challenges in any linguistic studies concerning low-resource languages is data scarcity, which in particular concerns written, digitised content available on the Web. Accordingly, any studies using automatic or semi-automatic Web scraping procedures may prove to be difficult or virtually impossible. This has been proven in the current study into the feasibility of terminology extraction from texts in Kashubian – a regional and a low-resource language. One solution could be a manual compilation of corpora, and another the automatic harvesting of pre-selected URLs of websites containing high-quality content (see Artetxe 2022).

It seems, however, that the task of term extraction task from texts representing low-resource languages may be better performed by GenAI tools, such as ChatGPT, compared to traditional corpus linguistics methods. With the considerable progress of AI tools seen over the last few years, it is highly probable that LLMs will soon be able to perform tasks similar to the ones discussed in this paper. It may well be that the models are already capable of doing it, the only necessary prerequisite being the effective application of prompt engineering techniques and customisation of ChatGPT.

This is probably the first study to compare two entirely distinct methods used for terminology extraction: corpus linguistics and GenAI. It is hoped that it will spark discussion of, and facilitate the work on, new approaches to terminology work. There is still place for

corpus linguistics (see P. Crosthwaite/ V. Baisa 2023), however the most promising direction is the development of new tools that will combine corpus linguistics with AI.

References

2005 Act = *The Act of 6 January 2005 on National and Ethnic Minorities and on the Regional Language*. (URL https://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU20050170141/O/D20050141.pdf). [Accessed 20.11.2023].

Artetxe, M. et al. (2022), *Does Corpus Quality Really Matter for Low-Resource Languages?* (in:) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 7383–7390.

Baroni, M. et al. (2006), *WebBootCaT: instant domain-specific corpora to support human translators* (in:) Proceedings of EAMT. 11th Annual Conference of the European Association for Machine Translation. Oslo, 247–252.

Batini, C./ M. Scannapieco (2006), *Data Quality: Concepts, Methodologies and Techniques*, Berlin-Heidelberg.

Brezina, V. (2018), *Statistics in Corpus Linguistics*, Cambridge.

Biuletyn Rady Języka Kaszubskiego (2018), *Uchwała Nr 2/RJK/18 z dnia 19.10.2018r. w sprawie stosowania słownictwa związanego z terminologią biologiczną*. Gdańsk, 47–60.

Cabré, M.T. (1999), *Terminology: theory, methods and applications*. Amsterdam/ Philadelphia.

Crosthwaite, P./ V. Baisa (2023), *Generative AI and the end of corpus-assisted data-driven learning? Not so fast!*, (in:) "Applied Corpus Linguistics" 3, 1–4.

Epstein, Z. (2023), *Art and the science of generative AI. Understanding shifts in creative work will help guide AI's impact on the media ecosystem*, (in:) "Science" 380(6650), 1110–1111.

Gaeta, L. et al. (2022), *Corpus linguistics for low-density varieties. Minority languages and corpus-based morphological investigations*, (in:) "Corpus" 23, 1–15.

Gołąbek, E. (2012/ 2013), *Słownik polsko-kaszubski*, Vol. A-K, L-O, P, R-Ś and preprints. Gdańsk.

IBM (n.d.), *What are AI hallucinations?* (URL https://www.ibm.com/topics/ai-hallucinations). [Accessed 20.11.2023].

Jabłoński, A. (2019), *Język kaszubski jako element narodowotwórczy*, (in:) „Postscriptum Polonistyczne" 1(23), 45–56.

Kobyliński, Ł./ K. Krasnowska-Kieraś (2018), *Part of speech tagging for Polish*, (in:) "Poznań Studies in Contemporary Linguistics" 55(2), 211–237.

Kilgarriff, A. et al. (2014), *The Sketch Engine: ten years on*, (in:) "Lexicography" 1, 7–36.

Leipzig Corpora Collection (2008), Kashubian Wikipedia corpus. (URL https://corpora.uni-leipzig.de/en?corpusId=csb_wikipedia_2021). [Accessed 24.11.2023].

Lew, R. (2023), *ChatGPT as a COBUILD lexicographer*, (in:) "Humanities and Social Sciences Communications" 10(704), 1–10.

Łukasik, M. (2014), *Compiling a Corpus for Terminographic Purposes*, (in:) "Komunikacja Specjalistyczna" 7, 71–83.

Nacel, J. (2013), *Chemiô òglowô i òrganicznô* (Kaszëbskô Encyklopediô Tematicznô). Gdańsk.

Nacel, J./ M. Jeliński (2018), *Biologiô: spòdlowô wiédzô* (Kaszëbskô Encyklopediô Tematicznô). Gdańsk.

Nowakowski, A. et al. (2022), *Adam Mickiewicz University at WMT 2022: NER-Assisted and Quality-Aware Neural Machine Translation*, (in:) Proceedings of the Seventh Conference on Machine Translation (WMT). Abu Dhabi. 326–334.

OpenAI GPT-4 (2003), *GPT-4*. (URL https://openai.com/research/gpt-4). [Accessed: 15.11.2023].

Pomierska, J./ D. Stanulewicz (2019), *Tworzenie korpusu języka kaszubskiego – dostępne źródła zdigitalizowanych tekstów*, (in:) „Acta Cassubiana" 21, 159–178.

Ratner, N. (2023), *Parallel Context Windows for Large Language Models*, (in:) "Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Volume 1: Long Papers", 6383–6402. (URL: https://aclanthology.org/2023.acl-long.352.pdf). [Accessed 20.11.2023].

Sarrion, E. (2023), *Exploring the Power of ChatGPT: Applications, Techniques, and Implications*. New York.

Shah, C. (2023), *AI information retrieval: A search engine researcher explains the promise and peril of letting ChatGPT and its cousins search the web for you*. (URL https://theconversation.com/ai-information-retrieval-a-search-engine-researcher-explains-the-promise-and-peril-of-letting-chatgpt-and-its-cousins-search-the-web-for-you-200875). [Accessed 20.11.2023].

Stanulewicz, D. (2015), *O wykorzystaniu korpusów w badaniach językoznawczych i o planowaniu korpusu języka kaszubskiego*, (in:) „Biuletyn Rady Języka Kaszubskiego", 375–390.

UNESCO/ InfoTerm (2005), *Guidelines for Terminology Policies Formulating and implementing terminology policy in language communities*. Paris.

Wodecki, B. (2023), *Leaderboard: OpenAI's GPT-4 Has Lowest Hallucination Rate*. (URL https://aibusiness.com/nlp/openai-s-gpt-4-surpasses-rivals-in-document-summary-accuracy). [Accessed 20.11.2023].

Woźniakiewicz, J. (2013), *Język kaszubski i friulijski. Zagrożenia i szanse. Porównanie stopnia zaawansowania procesów odchodzenia od języka*. Kraków.

Ziegler, A. / J. Berryman (2023), *A developer's guide to prompt engineering and LLMs.* (URL https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/). [Accessed 20.11.2023].

Zieniukowa, J. (2013), *Język kaszubski w tekstach naukowych*, (in:) M. Klinkosz/ Z. Lica (eds.), Językowy, literacki i kulturowy obraz Pomorza. Gdańsk, 19–28.

Zieniukowa, J. (2015), *Z badań nad językiem kaszubskich tekstów naukowych*, (in:) M. Klinkosz/ A. Lica/ Z. Lica (eds.), Językowy, literacki i kulturowy obraz Pomorza, T.2. Gdańsk, 228–238.