https://doi.org/10.25312/j.10126

Karolina Jankowska https://orcid.org/0000-0002-5023-8952
Adam Mickiewicz University Poznań
e-mail: karolina.jankowska@wsjo.pl

Mikołaj Pieniowski https://orcid.org/0000-0002-9613-6134
Adam Mickiewicz University Poznań
e-mail: mpieniowski@gmail.com

Domain linguistics resources for discovering criminal activities in Polish texts[1]

Domenowe zasoby lingwistyczne do wykrywania aktywności przestępczych w tekstach polskich

Abstract

This article considers the process of obtaining text data and the methodology of creating text corpora as well as the selection and the definition of individual lexical units in order to create a lexicon of crime vocabulary in Polish. The language material was developed and used in order to create an IT system supporting Polish uniformed services in searching for crimes committed or planned on the Internet. The crime categories considered were the following: smuggling and trafficking of drugs, cigarettes, alcohol, vehicles and machinery, weapons and explosives, trafficking in human goods and organs, trafficking and falsification of documents, sexual crimes and paedophilia. As a result of the work, a collection of over three thousand words and phrases was created. Additionally, a linguistic dataset of 3337 full texts from online sources was collected. The lexicon has been adapted to the requirements of computer processing for the needs of three system modules: Definition, Context, and Translator. The linguistic material was collected from various types of anonymous forums,

advertising sites online, where there is no content control, moderation and administration. The linguistic material has been tested and implemented in the *AISearcher* Border Guard System[1]

Keywords: domain lexicon creation, text corpora creation, linguistic resources, web-based crime detection

Streszczenie

Praca przedstawia proces pozyskiwania danych tekstowych i metodologię tworzenia korpusów leksykalnych, a także selekcję i definicję poszczególnych jednostek leksykalnych w celu stworzenia leksykonu słownictwa kryminalnego w języku polskim. Materiał językowy został opracowany i wykorzystany w celu stworzenia systemu informatycznego wspomagającego polskie służby mundurowe w poszukiwaniu przestępstw popełnionych lub planowanych w Internecie. Rozpatrywane kategorie przestępstw to: przemyt i handel narkotykami, papierosami, alkoholem, pojazdami i maszynami, bronią i materiałami wybuchowymi, handel ludzkimi dobrami i narządami, handel i fałszowanie dokumentów, przestępstwa seksualne oraz pedofilia. W wyniku prac stworzono zbiór ponad 3000 słów i fraz. Dodatkowo zebrano zbiór danych lingwistycznych składający się z 3337 pełnych tekstów ze źródeł internetowych. Leksykon dostosowano do wymogów przetwarzania komputerowego na potrzeby trzech modułów systemu: definicja, kontekst i tłumacz. Materiał językowy zebrano z różnego rodzaju anonimowych forów, witryn ogłoszeniowych online, gdzie nie ma kontroli, moderacji i administrowania treścią. Materiał językowy został przetestowany i wdrożony w systemie Straży Granicznej AISearcher.

Słowa kluczowe: tworzenie leksykonów domenowych, tworzenie korpusów tekstowych, zasoby językowe, wykrywanie przestępstw w internecie

Introduction

The Internet is a platform where people perform more and more everyday activities, such as private and work-related communication, data transfer, official matters, booking services and shopping, or barter goods exchange. Many security features and functions are used to identify the authors of entries and their location, but despite this, the Internet allows people to remain anonymous. This, in turn, enables illegal activities such as trafficking in illegal goods, distribution of illegal materials (e.g. child pornography) and other illegal activity. The Internet can function as a communication platform for those interested in taking part in this type of transactions and activities, which has long become the object of interest to services such as the Police, Border Guard and others. On the Internet, the main means of communication is natural, written language. Most advertisements, entries contain written content and possibly additional material in the form of photos, audio or video recordings. Due to the number of websites, users and the amount of content for potential analysis, attempts were made to automate the process of searching and detecting specific text content.

This article presents the process of collecting, normalizing and annotating the linguistic materials that were used to create a system for searching and identifying criminal activities on the Internet. These materials constitute tests of criminal advertisements, lexicons of domain (crime) vocabulary and others.

[1] Advanced analysis of Internet resources supporting the detection of criminal groups (AI Searcher) no. DOB-BIO9 /19/01/2018

State of the art

Computational lexicography deals with the study and modelling of the automatic acquisition of lexical units from collections of texts, the construction of lexicons on the basis of corpus, automatic extraction of syntactic and semantic information, creating, extending and maintaining machine-readable dictionaries (Van Eynde, Gibbon, 2000). There are many ready-made, generally available language sources in various languages, i.e. resources provided by the European Language Resources Association (ELRA), the Linguistic Data Consortium (LDC) or the ELSNET group. However, these resources are mainly in English and dedicated to the general language.

There are many language resources, both written and spoken, in Polish. In recent years, very high-quality, annotated language corpora and tools for their analysis have been developed. One of the examples of the Polish language corpus is National Corpus of Polish (Narodowy Korpus Języka Polskiego, n.d.) containing over 1 billion words, of which a 300-million word subcorpus has been carefully balanced, and a manually-annotated 1-million corpus has been released under an open license. The corpus offers two tools for searching and analysing – PELCA and Poliqarp (Pęzik, 2012).

There are also other text corpuses in Polish, such as: (1) PICLE corpus – the Polish sub-corpus of the International Corpus of Learner English (ICLE) (n.d.)[2], (2) IMPACT – ground-truth data for selected Polish historical documents from PIONIER Digital Libraries Federation (*Results of the IMPACT project*, n.d.), (3) KPWr – a collection of documents annotated with syntactic chunks, proper names, semantic relations, anaphora and word senses (*KPWr*, n.d.), (4) Polish Corpus of Suicide Notes (Polski Korpus Listów Pożegnalnych, n.d.), (5) Polish Wikipedia Corpus (Polish Wikipedia Corpus, n.d.), (6) gpwEcono – a corpus of stock market reports with manual word sense annotation (gpwEcono, n.d.), (7) plWikiEcono – a corpus of Polish Wikipedia articles from the domain of economy (plWikiEcono, n.d.b), (8) Polish Coreference Corpus – a corpus of Polish coreference relations, created as part of the CORE project (Polish Coreference Corpus / Korpus zależności referencyjnychn, n.d.), (9) Microcorpus of Synesthetic Metaphors (Polski Korpus Metafor Synestezyjnych SYNAMET, n.d.).

An example of a scientific group working on tools and language resources for Polish is the Polish consortium partner within the Common Language Resources and Technology Infrastructure (CLARIN) initiative (CLARIN-PL, n.d.). CLARIN-PL provides software tools for the use of existing digital archives and corpora. It also creates, stores and shares new resources for natural language processing. The software enables working on raw text form the Internet such as articles, blogs and other documents. The basic language processing tasks that can be performed with CLARIN software are: automatic text summarisation, search for entity names, morphological and syntactic analysis. CLARIN-PL offers resources, such as Spokes – conversation data search engine, ChronoPress – test version of the Chronological Corps of Polish Press Texts (1945–1954), Parallel – Polish-English parallel corps search engine, Words of the day – a list of words from media discourse

[2] http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.html [accessed: 3.09.2025].

the frequency of use in seven selected daily newspapers on a given day is clearly higher compared to the frequency of these words in the last twelve months, Walenty – valence dictionary of the Polish language. Such task are very useful in, for example, discourse studies.

The wordnet for the Polish language developed within the CLARIN-PL project is Słowosieć 3.0 (Słowosieć, n.d.). This presents lexical system of the Polish language. Słowosieć 3.0 is a set of lexical items that have specific semantic relations based on their meaning, e.g. hyponyms, meronyms, synonyms and antonyms. One lexical item gets its meaning by referring to another lexical item within the system. The structure of the Słowosieć 3.0 addresses the requirements of automatic text analysis which is crucial for artificial intelligence studies in Polish. Currently, Słowosieć 3.0 consists of 178.000 verbs, nouns, adjectives and adverbs. It contains nearly 259.000 unique meanings and 600.000 relations (Maziarz et al., 2016).

Another example of a group of scientists working in the field of language technologies and dedicated resources is the Group of Language Technologies of the Wrocław University of Technology (GTJ PWr) (Apache2 Ubuntu Default Page, n.d.). GTJ PWr works on the development and implementation of natural language computer processing tools, including the creation and expansion of applications and language resources for the Polish language. The GTJ PWr website presents the following linguistic resources: (1) CEN – economic content corpus containing 797 documents from the Polish edition of Wikipedia (CEN, n.d.), (2) KPWr – Polish Language Corpus of Wrocław University of Science and Technology – corpus text files annotated at various levels, e.g. phrases, spatial expressions, temporal expressions, keywords, situations (KPWr, n.d.), (3) KPWr lemma – corpus, hand-lemmatised Polish noun and adjective phrases (Corpus of manually lemmatised Polish noun and adjective phrases, n.d.), (4) list of distributive semantic similarity (Lista dystrybucyjnego podobieństwa semantycznego, n.d.), (5) frequency lists generated on the basis of large text corpora (*Lista frekwencyjna*, n.d.), (6) NELexicon – a list of proper names noting over 1.4 million unique onims and NELexicon2 being an extended version containing over 2.3 million unique onomastic words (NELexicon, n.d.).

The most extensive language resources practically tested in the Polish LVCR system is the Polish Lexical SpeechLabs Database created at Adam Mickiewicz University in Poznan (Demenko, 2015), which includes 3 lexicons containing: common words, proper names and special application words, (several million noted units) and resources text from various fields (tens of thousands).

Several groups of scientists work in this field of science and provide qualitative, useful materials and tools useful in text processing. However, for language analysis technology research and development in a specific thematic category, specialized linguistic resources are required, which are usually not available and must be developed specifically for the project. For the project described in this article, text analysis to identify criminal content, no existing material was found to meet the requirements of specific application of language information processing for detecting criminal activities in web texts.

Methodology

Texts corpora

The first step towards the creation of a crime-related text corpus to serve as the basis for lexicon creation was the analysis of requirements, which the texts were to meet (Gibbon, Moore, Winski, 1997; Demenko, 2015). The main criteria for the inclusion of texts in the corpus were their language, richness in domain-specific lexis, representativeness and authenticity, as well as an adequate structure enabling the extraction of entities belonging to many diverse categories of entities and crimes. Nine categories of entities were identified for the entity extraction purposes (Demenko et al., 2022). Table 1 lists each of the categories together with brief descriptions.

Table 1. Named entity categories

Category	Description
Location	Includes all terms indicating the location of the event described in the text under analysis.
Object – person	Includes entities referring to persons
Object – thing	Includes entities referring to items
Action	Includes all actions related to criminal activities
Organization	Includes the names of organizations of all kinds
Identifier	Includes all units enabling the identification of the author of the text.
Time	Includes temporal expressions including, but not limited to, dates, times, durations, etc.
Measure	Includes terms referring to the volume or physical size of specific items.
Description	Includes all kinds of descriptions, comments and explanations.

Source: author's own work

In order to meet all the criteria mentioned above, it was crucial to identify the types and then specific addresses of websites, which could potentially contain authentic texts of advertisements for illegal activities. For this purpose, usually simple, two-part queries were entered into the Google search engine. The queries consisted of an action verb, mainly *I'll sell* or *I'll buy*, followed by the name of an illegal item. In the case of any of the search results being deemed relevant, the query was re-entered into the internal search engine of the website. Using this approach, three Polish websites containing ample relevant linguistic material were identified. These included two advertising websites, namely Oglaszamy24h.pl (n.d.) and TOPOgłoszenia (n.d.) as well as an online shop offering illegal drugs and so-called research chemicals, namely Dopalacze-sklep (n.d.).

In order to obtain more texts abundant in domain-specific lexis, attempt was made to identify relevant sources of such data in the TOR (The Onion Router) network. The TOR network enables the user to hide their activity in the Internet (Tor Browser, n.d.). Given this characteristic of the TOR network, it is often used for illegal activities, such as on-

line drugs, medicine or even child pornography trafficking. Various TOR sites serve as auction sites dedicated to trafficking illegal goods. The sites are not reachable through any search engines and their URLs constitute random strings of letters and numbers, each ending with the "onion" domain (Krauz, 2017). Those factors generated the need for the use of lists of working TOR links published in Clearnet. Using a listing published by ITcontent (ITcontent, n.d.), it was possible to identify five relevant TOR websites, which later served as a source of linguistic data. The sites identified included two Polish forums, namely *Cebulka* and *Darknet* and three international auction sites, namely *Apollon Market*, *Dream Market* and *White House Market*. TOR sites, with their clearly defined target audience, enabled detailed filtering of content by, for example, product category or country of origin, which made it much easier to identify relevant texts. The categories of products or advertisements were often corresponding to the aforementioned categories of crimes relevant in the context of preparing the lexicon. Except for *Cebulka,* each of the sites required registration in order to view their content in its entirety, as it is with most of illegal activity related TOR network websites (Mider, 2019). Moreover, the sites were also protected by challenge-response tests, which prevent bots from entering the sites and browsing their contents. This fact rendered automatic data extraction inefficient when compared to manual approach in the case of TOR sites.

In the case of Clearnet data sources, the linguistic material was obtained automatically using a proprietary web scraper adjusted individually to each of the sites. The programming language of choice for the web scraper development was Python (python, n.d.). It was also necessary to select modules responsible for handling HTTP requests, regular expressions, and parsing the HTML code. The tools selected for this purpose include Requests: HTTP for Humans (n.d.), Beautiful Soup (*Beautiful Soup Documentation*, n.d.) as well as the *re* (*re – Regular expression operations*, n.d.) module enabling the use of regular expressions. Besides capturing the texts, the script was also removing duplicates on an ongoing basis. The results were saved in *txt* files, where each text was separated by an outstanding symbol and later in the *csv* format as well, to facilitate manual analysis of the results.

Alongside the web scraper, two additional, simple programs for semi-automatic analysis of the obtained results were developed. Both were developed using Python programming language and the NLTK (Natural Language Processing Toolkit, n.d.) library. The first one was responsible for generating frequency lists with stop words excluded. The second program enabled quickly generating concordances based on the previously obtained corpus for given lexical units.

Creating lexicons

First, the types of crimes the project administrator was interested in were defined. The following crime categories were defined: smuggling, trafficking, production of drugs, trafficking of human beings and organs, smuggling and trafficking of cigarettes, smuggling and trafficking of alcohol, pornographic industry and paedophilia, forgery and trafficking of documents, smuggling and trafficking of arms and explosives, smuggling and

illicit trade in passenger vehicles, trucks, agricultural and construction machinery. Due to the wide range of criminal categories and time and personnel constraints, prioritization of crimes of interest was introduced. It was decided that crimes related to the trafficking and smuggling of drugs, alcohol and cigarettes will be of the highest priority. Secondly, sex crimes and human trafficking. Illicit trafficking in vehicles, machinery, weapons and explosives remained in third place. This decision was made by contractors and recipients (end users) of the project and was based on arguments such as the frequency of crimes committed, the ability to detect them online and, consequently, the extensive linguistic resources. Secondly, the modules and technical methods for which the linguistic material was to be used were identified. Within the project, a criminal language translation module was developed in Polish-Belarusian, Polish-Russian and Polish-Ukrainian. The AI Searcher context module extends user queries, uses synonyms of a keyword or phrase to search the Internet, based on a lexicon of jargon vocabulary and general equivalents. The definition module uses lexicons to recognize specific text elements - Named Entity Recognition (NER). Therefore, for the purposes of the project, the following requirements for lexicons were defined: (1) to contain vocabulary relating to trackable online crimes, (2) to contain vocabulary relating to the illegal trafficking and smuggling of drugs and drugs, cigarettes, alcohol, vehicles, agricultural and construction machinery, weapons and explosives, trafficking in human goods and organs, sexual crimes and paedophilia, trafficking and falsification of documents, (3) crimes related to the trafficking and smuggling of drugs, cigarettes, alcohol, sexual crimes and paedophilia were included in the most important categories, (4) the lexicon must be computer and AI Searcher modules readable, (5) the lexicon must be open-ended, allowing the addition and modification of its content.

The procedure included the following steps: (1) creating frequency lists from all existing lexical units occurring in the text corpora created within the project, (2) expert linguistic analysis of lexical units and phrases, (3) context, semantic and pragmatic analysis, (4) compilation of collected data with existing sources (Demenko, 2015). In order to create a dictionary and analyse the Polish lexica related to drugs, the existing dictionaries and texts from the Internet forums and other Internet sources were collected. The dictionaries considered are mainly collections for the police, support centres for troubled youth and their parents, forensics and various social campaigns.

In order to collect vocabulary characteristic of trafficking (forced prostitution, slave labour), the porn industry and paedophilia, linguistic material was collected from various online sources such as sites with pornographic material, anonymous forums with user contributions containing pornographic and paedophilic material etc. For smuggling and illegal trade of alcohol and cigarettes, proper names and names of types of articles were collected from generally accessible websites of online stores. The Register of Explosives was used for the category of illegal possession and trafficking of weapons and explosives intended for civilian use, names of weapons from online stores. For the category of trade in passenger vehicles, trucks, agricultural and construction machinery, dictionary synonyms for keywords and brand names from intermediary portals were used. To complete the lexicon, synonyms of keywords available in general dictionaries of the Polish language and jargon dictionaries were used.

Domain linguistic resources

Text corpora

As mentioned, texts were stored in both *txt* and *csv* format files. In the case of txt files, each text was separated by an outstanding symbol, which facilitated automatic processing of the corpus. The csv format facilitated manual analysis of the corpus contents. The corpus was divided into two subcorpora corresponding with the crime categories of their contents. The division created *Alcohol/Tobacco* and *Drugs/Medicine* subcorpora.

For validation purposes, another subcorpus consisting of 450 full texts was isolated from the general dataset in the *txt* version in order to be annotated subsequently. The subcorpus dedicated to evaluation was then divided into a total of 45 files, each containing 10 texts. Each of the files was given an identifier. Section "Annotation" describes the process during which the evaluation subcorpus was annotated.

Annotation

After completing the pre-processing of the evaluation subcorpus explained in this section, the process of annotating the linguistic data can be divided into several pivotal steps: (1) Preparation of an adequate training for the annotators, (2) Preparation of a workspace for the annotators and selecting a method of work progress tracking (3) Tracking the work progress and staying in constant touch with the annotators.

A total of nine annotators were engaged to work with the collected linguistic material. Within the first step, which was to prepare the training for the annotators, two documents containing annotation instructions were produced. One of them very briefly presented a list of categories of entities to be tagged in the texts, together with the method of tagging and additional assumptions, such as cases as nested tags, etc. The entity categories used for annotation differ slightly from those shown in Table 1. Table 2 presents the annotation categories together with the method of entities tagging.

Table. 2. Categories and their corresponding tags used within the annotation process.

Category	Tags
Identifier	*<I></I>*
Identifier – Signature	*<SIG></SIG>*
Identifier – Contact	*<K></K>*
Object – Person	*<O></O>*
Object – Thing	*<P></P>*
Action	*<A></A>*
Organization	*<ORG></ORG>*
Location	*<L></L>*
Time	*<T></T>*
Measure	*<M></M>*
Description	*<D></D>*

Source: author's own work

As shown in the table above, the aforementioned expansion of the categories from nine to eleven is due to the division of the "Identifier" category into two subcategories – Signature (all the entities that constitute the first name, surname, or pseudonym of the author of the text) and Contact (a category that includes all the elements of the text that enable contact with its author, i.e. contact forms, email addresses, telephone numbers, etc.). The aim of dividing the Identifier category was to achieve higher accuracy.

The other document describes in great detail each of the entity categories giving sets of examples for each one of them. A dedicated group of 10 advanced students in linguistics in Microsoft Teams was created, where both annotation manuals were made available to the annotators, and an implementation meeting was organized and held, during which both annotation manuals were discussed, some practical examples of annotating criminal texts were presented, and a discussion in the Q&A format was held, during which the most troublesome issues and all kinds of questions from annotators about the work were clarified. In addition, the platform served as a communicator enabling constant contact between annotators and the lead annotator, which increased the efficiency of the entire process.

Within the second step, whose aim was to provide the annotators with an adequate workspace, a cloud space was allocated in which the repository of pre-processed texts was stored. The repository comprised the 45 bundles of texts, each having an individual identifier. The cloud location contained two folders, namely *Data* and *Instructions* as well as one file, namely *Annotators' IDs*. As the file name suggests, each annotator was assigned an identification number to facilitate the allocation of texts and to improve communication. As the name of the last listed file suggests, each annotator was assigned an identification number in order to facilitate the allocation of texts and to improve communication. The file *Annotators' IDs* contained information on who exactly receives what ID. All IDs were complying with the simple structure of "A<No.>" (e.g., A1, A2, A3 etc.). The *Instruction* folder contained the two aforementioned documents describing the guidelines for crime texts annotation. The *Data* folder contained multiple subfolders and a file named *Text Allocation* representing the allocations of texts for each annotator. Simultaneously, the *Text Allocation* spreadsheet served as a simple work progress tracking tool. Each sheet of the spreadsheet was named after the ID of a given annotator and contained a list of the files, which were to be tagged by them. Alongside the filenames, there were two columns, which the annotators were supposed to fill after the completion of each text bundle. The first one of the columns contained simple checkboxes to be marked once a given bundle was fully tagged and the second one containing the space for entering the time it took to complete the work on the given bundle. Nine of the subfolders located in the *Data* folder were the designated spaces for placing completed (tagged) text bundles. The final subfolder in this location – *Texts* – served as the raw linguistic data bundles repository.

Lexicons

As a result of automatic and manual analysis and the compilation of a dictionary created from collections of texts with the existing resources, a new domain lexicon in Polish was created. It is divided into categories of crimes (trackable crimes on the Internet related

to smuggling and illegal trade in drugs, alcohol, cigarettes, cars and machines, weapons and explosives, organs, document forgery, human trafficking, procurement and prostitution, pornography and paedophilia) and named entities categories (object, action, description, measure). There are more named entities categories defined (see table 1), however they are not included in the lexicon of criminal vocabulary, but a general language vocabulary. That is why these lexical items are not included in the lexicon and statistics. The lexicon is saved as an *xls* file, where one sheet corresponds to one crime or named entity category. The drugs category is the most extensive due to the volume of linguistic material and the priority given by the end users of the system.

Evaluation

Text corpora statistics

The process of searching and collecting criminal texts resulted in the creation of a corpus containing a total of 3337 texts corresponding to a total of four categories of crime, namely trafficking of drugs, medicine, alcohol and cigarettes. This number breaks down into 2261 full texts corresponding to drugs and medicine trafficking and 1076 texts corresponding to tobacco and alcohol trafficking.

The entire endeavour of evaluation subcorpus annotation resulted in obtaining a set of 9464 annotated text fragments and an additional subset, whose elements amounted to 855, which were annotated by two independent annotators. The annotated dataset served as was used for automatic evaluation of the name entity recognition algorithm being developed in parallel, however what is more important from the point of view of creating a domain-specific lexicon, the tagged linguistic material could facilitate the search for more domain-specific lexis, as it could be browsed by marked categories of entities.

Lexicon's statistics

The entire lexicon contains 3135 lexical units. The table below presents the number of individual lexical units and phrases in categories and their percentage in relation to the entire lexicon.

Table 3**.** Lexicon's statistics

Named entity categories	**Units**	**Percent**
object	2225	70%
action	284	10%
description	566	18%
measure	60	2%
Crime categories		
drugs	706	23%
alcohol	850	27%
cigarettes	255	8%

Named entity categories	Units	Percent
cars and machines	997	32%
weapons and explosives	68	2%
organs, document forgery	40	1%
human trafficking	72	2%
sexual offences	147	5%

Source: author's own work

The lack of balance in the number of vocabulary in individual categories results from the aforementioned prioritization of crime categories, the degree of vocabulary development in a given field as well as their popularity and availability of linguistic material. In the case of drugs, verbal language is the highest, as there are many forums where users talk about their experiences anonymously. For each drug there is a common word, a chemical compound, more or less known abbreviations and terms. In the case of pharmaceutical drugs used as psychoactive substances, intoxicating drugs come with trade names, usually in several variants. On the other hand, the category related to the trade in organs and documents contains the fewest terms due to the fact that they are usually general language words and their few variants.

The lexicon has the characteristics of a valuable linguistic resource such as: (1) up-to-date – it contains terms currently used by Internet users in specific thematic areas, (2) extensive – it includes vocabulary and phrases from general language, jargon (cryptolect), (3) computer-readable – its content is easy to read. automatic processing, (4) universal – it can be used for various scientific and practical purposes (e.g. as support for the work of the police, social welfare, caretakers of troubled youth, etc.).

Conclusion

Creating domain resources is a difficult task, especially in the case of criminal vocabulary, i.e. secret language, cryptolect. Getting to the right materials is difficult and understanding the content requires getting to the context, which is usually non-obvious and difficult to understand. In such situations, the only solution was to analyse a given linguistic unit in many use cases and contexts.

In the course of compiling language materials and searching available glossaries and other materials, it was noticeable that this vocabulary was changing quite dynamically. This phenomenon is related to the need to create new terms so that they are understandable only to a specific group of people. When vocabulary becomes more popular, its meaning becomes clear to people outside the group, it ceases to function. In the case of drugs, it is also important that new psychoactive substances with different chemical and commercial names are constantly emerging, which is also reflected in the language. Similarly, but to a lesser extent, this applies to the trade names of alcohol, cigarettes, vehicles, etc.

Due to these issues, we conclude that such reference books should be supplemented and updated on a regular basis.

Searching the Internet for linguistic data relevant to the task of composing a domain corpus described in this article has proven that the Internet is an exceptionally rich source of linguistic data and is abundant in authentic linguistic samples crowded with domain-specific lexis of various kinds. Even in the case of the desired language field being closely related to illegal activities, there are ample sources capable of providing expansive collections of data. Mention of the Internet includes both Clearnet and the TOR network, which has provided much valuable data within the task. The vast majority Clearnet sites provide the possibility to gather, normalize and finally process their contents automatically due to the lack or simplicity of anti-bot security measures. The majority of them also do not require any form of registration, which makes the task of developing a suitable web scraper easier.

The results of the corpus creation shed some light on which categories of crimes are the most prevalent among the Internet's auction and advertising sites. We were able to either automatically, as in the case of Clearnet sources, or manually, as in the case of TOR network sources, obtain over three thousand full texts belonging to one of the following categories of crime – trafficking of drugs, medicine, alcohol and tobacco. Locating Polish texts related to forgery, car trafficking or firearms trafficking were much more rare and their number was insufficient for composing an adequate subcorpora. The hardest to locate in both Clearnet and the TOR network were Polish texts related to sex trafficking, human cargo trafficking and paedophilia. On the other hand, the most popular category of crimes among the located linguistic data sources were illegal drugs and medicine requiring prescription. Even on the regular Internet websites such as Oglaszamy24h.pl, more texts were found related to illegal drugs than those offering illegal alcohol or tobacco.

The lexicon is currently being supplemented by a group of specialists in the field of international crime on the Internet from the Polish Border Guard, using the designed AISearcher system. The lexicon is expected to have an increasingly practical dimension due to the continuous collection of text by border guards.

Acknowledgements

The research has been conducted within the project *Advanced analysis of Internet resources supporting the detection of criminal groups (AI Searcher)* financed by the National Center for Research and Development (no. DOB-BIO9 / 19/01/2018).

References

Demenko G. (2015), *Korpusowe badania języka mówionego*, Poznań: Akademicka Oficyna Wydawnicza EXIT.

Demenko G., Skórzewski P., Kuczmarski T., Pieniowski M. (2022), *Linguistic Information Extraction from Text-based Web to Discover Criminal Activity*, s.l.: unpublished manuscript.

Eynde Van F., Gibbon D. (2000), *Processing, Lexicon Development for Speech and Language*, Berlin: Springer.

Gibbon D., Moore R., Winski R. (1997), *Handbook of standards and resources for spoken language systems*, Berlin: Walter de Gruyter.

Krauz A. (2017), *Mroczna strona Internetu – tor niebezpieczna forma cybertechnologii*, „Dydaktyka informatyki", nr 12, pp. 63–74.

Maziarz M., Piasecki M., Rudnicka E., Szpakowicz S., Kędzia P. (2016), *plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource*, [in:] Y. Matsumoto, R. Prasad (eds.), *26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, Osaka: The COLING 2016 Organizing Committee, pp. 2259–2268.

Mider D. (2019), *Czarny i czerwony rynek w sieci The Onion Router – analiza funkcjonowania darkmarketów*, "Przegląd Bezpieczeństwa Wewnętrznego", nr 29, pp. 154–190.

Pęzik P. (2012), *Wyszukiwarka PELCRA dla danych NKJP*, [in:] A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk, *Narodowy Korpus Języka Polskiego*, Warszawa: PWN, pp. 253–273.

Websites

Apache2 Ubuntu Default Page (n.d.), http://www.nlp.pwr.wroc.pl/ [accessed: 23.02.2026].

Beautiful Soup Documentation (n.d.), https://www.crummy.com/software/BeautifulSoup/bs4/doc/ [accessed: 3.03.2026].

CEN (n.d.), Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej, http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/cen [accessed: 3.09.2025].

CLARIN-PL (n.d.), https://clarin-pl.eu/ [accessed: 3.03.2026].

Corpus of manually lemmatised Polish noun and adjective phrases (n.d.), (n.d.), Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej, http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/kpwr-lemma [accessed: 3.09.2025].

Dopalacze-sklep (n.d.), https://dopalacze-sklep.org/ [accessed: 3.09.2025].

gpwEcono (n.d.), https://zil.ipipan.waw.pl/gpwEcono [accessed: 23.02.2026].

ITcontent (n.d.), https://itcontent.eu/ [accessed: 23.02.2026].

KPWr (n.d.), Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej, http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/kpwr [accessed: 3.09.2025].

Lista dystrybucyjnego podobieństwa semantycznego (n.d.), Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej, http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/lista-podobienstwa [accessed: 3.09.2025].

Lista frekwencyjna (n.d.), Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej, http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/lista-frekwencyjna [accessed: 3.09.2025].

Narodowy Korpus Języka Polskiego (n.d.), http://nkjp.pl/ [accessed: 23.02.2026].

Natural Language Toolkit (n.d.), https://www.nltk.org/ [accessed: 23.02.2026].

NELexicon (n.d.), Grupa Technologii Językowych G4.19 Politechniki Wrocławskiej, http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/nelexicon [accessed: 3.09.2025].

Ogłaszamy24h.pl (n.d.), https://oglaszamy24h.pl/ [accessed: 23.02.2026].

plWikiEcono (n.d.), http://zil.ipipan.waw.pl/plWikiEcono [accessed: 23.02.2026].

Polish Coreference Corpus / Korpus zależności referencyjnych (n.d.), http://zil.ipipan.waw.pl/PolishCoreferenceCorpus [accessed: 23.02.2026].

Polish Wikipedia Corpus (n.d.), http://clip.ipipan.waw.pl/PolishWikipediaCorpus [accessed: 23.02.2026].

Polski Korpus Listów Pożegnalnych (n.d.), http://www.pcsn.uni.wroc.pl/ [accessed: 23.02.2026].

Polski Korpus Metafor Synestezyjnych SYNAMET (n.d.), http://synamet.polon.uw.edu.pl/ [accessed: 23.02.2026].

python (n.d.), https://www.python.org/ [accessed: 23.02.2026].

re – Regular expression operations (n.d.), https://docs.python.org/3/library/re.html [accessed: 23.02.2026].

Requests: HTTP for Humans™ (n.d.), https://docs.python-requests.org/en/latest/ [accessed: 23.02.2026].

Results of the IMPACT project (n.d.), Digital Libraries and Knowledge Platforms Department, http://dl.psnc.pl/activities/projekty/impact/results/ [accessed: 3.09.2025].

Słowosieć (n.d.), http://plwordnet.pwr.wroc.pl/wordnet/ [accessed: 23.02.2026].

TOPOgłoszenia (n.d.), https://top-ogloszenia.net/ [accessed: 23.02.2026].

Tor Browser (n.d.), https://www.torproject.org/ [accessed: 23.02.2026].