# Korpusomat.eu: A Multilingual Platform for Building and Analysing Linguistic Corpora

Karol Saputa(✉), Aleksandra Tomaszewska, Natalia Zawadzka-Paluektau, Witold Kieras, and Łukasz Kobyliński

Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warszawa, Poland
{k.saputa,a.tomaszewska,n.zawadzka-paluektau,wkieras, lkobylinski}@ipipan.waw.pl

**Abstract.** The paper presents Korpusomat, a new, free web-based platform for effortless building and analysing linguistic data sets (corpora). The aim of Korpusomat is to bridge the gap between corpus linguistics, which requires tools for corpus analysis based on various linguistic annotations, and modern multilingual machine learning-based approaches to text processing. A special focus is placed on multilinguality: the platform currently serves 29 languages, but more can be easily added per user request. We discuss the use of Korpusomat in multidisciplinary research, and present a case study located at the intersection between discourse analysis and migration studies, based on a corpus generated and queried in the application. This provides a general framework for using the platform for research based on automatically annotated corpora, and demonstrates the usefulness of Korpusomat for supporting domain researchers in using computational science in their fields.

AQ1

**Keywords:** natural language processing · corpus linguistics · corpus building · discourse studies · migration studies

## 1 Introduction

Computational technology has revolutionised linguistics by offering powerful tools for language analysis [7, p. 74]. They have progressed through several generations to address user demands and the growing need for larger corpora. First corpus platforms emerged in the late 1970 s with limited usage on mainframe computers in renowned institutions. Second-generation platforms were developed in the 1980 s, introducing concordance and frequency tools on IBM-compatible computers. The third generation facilitated the creation and processing of large corpora on personal computers, increasing accessibility [1]. Today, fourth-generation tools, like Korpusomat, provide advanced features and convenient access to larger data collections. They use multiple export options and features such as Corpus Query Language (CQL), frequency lists, concordance,


J. Mikyška et al. (Eds.): ICCS 2023, LNCS 14074, pp. 1–8, 2023.
https://doi.org/10.1007/978-3-031-36021-3_22

and many more, and can be integrated with other computational methods [3], facilitating the platform selection for a given project. They can be accessed via browsers and store corpora on servers, eliminating the need for local storage and allowing for the processing of larger corpora [1]. The ongoing development of corpus tools contributes to computer science and linguistics, enabling new insights and a deeper understanding of authentic language usage.

Korpusomat[1], a web application designed for building and analysing corpora based on user-provided texts, initially supported only Polish texts [4,5][2]. In this paper, we are presenting a new multilingual version of Korpusomat. Although recent advances in natural language processing have introduced a variety of multilingual high-level text processing frameworks, they still require programming skills and lack the analytical tools typically employed by corpus linguists. The aim of Korpusomat is to bridge this gap between corpus linguistics and modern multilingual machine learning-based approaches to text processing.

## 2 Korpusomat: Creating Universal Dependencies Corpora

Korpusomat [6][3] aims to address the limitations of corpus tools by offering advanced features, a user-friendly interface, and compatibility with many languages and corpora. It employs two high-level natural language processing libraries, spaCy and Stanza, and the Universal Dependencies (UD) framework [6]. It serves 29 languages and more may be added at users' request as long as text processing frameworks provide models for these languages. With its ability to process files in different formats and convert them to the required encoding, Korpusomat makes building and analysing corpora particularly effortless. Moreover, the tool's integration with the newspaper library allows researchers to add texts from websites, expanding the corpus beyond traditional file import. Korpusomat processes most text data formats[4]. It also allows its users to export data and download the processed texts with annotation layers.

AQ2

One of the most significant benefits of Korpusomat is the use of the Universal Dependencies (UD) framework. It provides cross-linguistically consistent treebank annotation for over 100 languages. It also enables researchers to easily transfer certain methods from one language to another. A functionality that is unique compared to other corpus tools is the visualization of dependency trees. It is interactive and includes the full utterance containing an example with query results. The results can be modified, for example, punctuation marks may be hidden or the tree may be automatically positioned.

[1] The version presented in this paper is available at https://korpusomat.eu.

[2] This older version is still maintained and available at https://korpusomat.pl.

[3] The Description of the Features, Including the Query Language, Is Available in the Documentation at https://korpusomat-eu.readthedocs.io/en/latest/.

[4] A list of possible formats is available in the documentation at http://tika.apache.org/1.17/formats.html.

## 3 Computational Architecture

The architecture of Korpusomat has a number of requirements regarding the processing of texts and their provision in a searchable, indexed form. In this section we describe the ideas behind the Korpusomat architecture important for the scaling of Korpusomat to the multilingual version. The overall architecture is shown in Fig. 1, with the main components being an API platform, a task queue, a search backend and the processing pipelines described below.

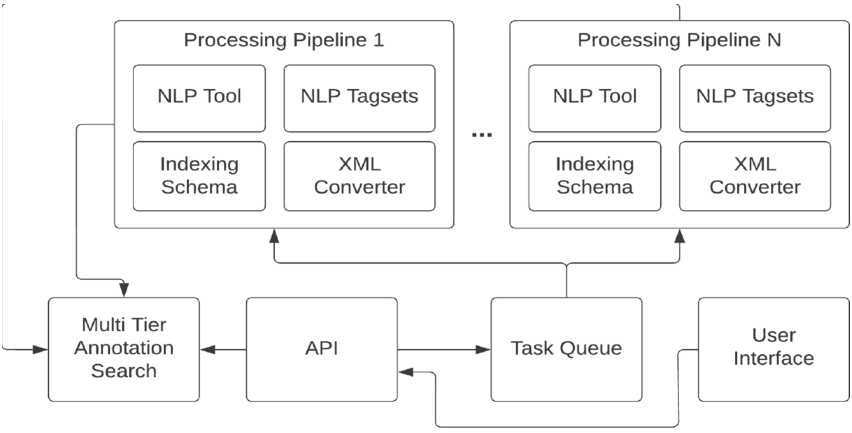


**Fig. 1.** Overview of the Korpusomat architecture components, showing the elements of the processing pipelines and their interaction with the search backend and the application.

**Table 1.** The Table shows the time (in minutes) required to build a corpus of three different sizes (in tokens), with the standard deviation in parentheses. The comparison includes two different NLP tools that we currently use.

| Corpus size | NLP tool | Build time |
|---|---|---|
| 147,580 | spaCy | 5 (0.57) |
| 1,475,800 | spaCy | 92 (4.2) |
| 14,758,000 | spaCy | 923 (8.6) |
| 147,563 | Stanza | 13 (1.5) |
| 1,475,630 | Stanza | 94 (5.6) |

### 3.1 Processing Pipelines

A processing pipeline (PP) employs a set of components used for adding a text to a corpus (Fig. 1). PPs are interchangeable and versioned thanks to the configurability of their elements: NLP tool used for tagging, extracted tagsets used

by the NLP tool, schema of indexing linguistic layers in the search engine, and XML converter of annotated text into indexing schema format.

Due to the multilingual nature of the Korpusomat platform, there is no single tool for all the languages served. We currently use two open source NLP libraries: Stanza and spaCy to provide us with linguistic annotations of texts. Each tool, spaCy and Stanza, can be packaged in a separate, interchangeable PP. It allows for the use of different tools, and the use of one tool for different contexts (e.g. in two different indexing schemes with different linguistic layers available for corpora with these schemes, although using the same NLP tool).

Upgrading the models in the NLP libraries, adding new languages, and other updates require versioning of PPs. Each of the PPs uses semantic versioning, which is linked to the version of the corpora. This ensures that each corpus has a consistent tagging of all its texts.

The specific tagset (e.g. set of named entity tags) extracted from the NLP tool is used for indexing schema generation and then attached to each corpus and used e.g. for the custom query builder of that corpus. An updated version of the PP can use a model with an extended tagset for one of the annotation layers, then there is a mechanism for extracting the tagset from the processing libraries for each language.

### 3.2 Index and Search

Korpusomat uses MTAS [2] as its corpora search backend. It is a Solr/Lucene search engine extended with the ability to search through annotation layers of texts using the CQL. The indexing scheme specifies the linguistic layers for the corpus, corresponding to the possible queries to the corpus. Each of the corpora has its own indexing scheme based on the chosen PP and its tagset. Each type of linguistic annotation, such as named entities or part-of-speech tags, is treated as a separate annotation layer.

Indexing schemas are automatically generated for the specific PP based on its tagset, and thus versioned to maintain compatibility with older corpora, while automatically allowing for new tags to be indexed, searched, and made visible in the application's query builder.

In addition to linguistic annotations, MTAS indexing includes document metadata, which can provide additional information about individual texts, such as author, publication date, publisher. This metadata can then be used to filter query results or to group results by metadata.

## 4 Computational Performance Benchmarks

In this section we discuss the performance of the platform based on two of its core features: creating corpora and querying corpora. We show how easily and efficiently Korpusomat can be used as an end-to-end research platform for building (in minutes) and searching (in seconds) corpora. The system is tested on a machine with 16 CPU threads, 47 GB of memory, and SSD storage.

We test the time of adding an ebook in English, containing 147 580 tokens, in three settings: 1, 10, and 100 copies. We measured the time for building the corpora (Table 1) and then querying them for each of these cases.

Query time was measured as the time taken to serve an http query via the API. We analyse four types of queries (respectively in Table 2): (1) all corpus tokens, (2) all corpus sentences, (3) all tokens that are adjectives, and all named entities that contain an adjective.

Queries were run on the three corpora of different sizes. Each query returned 1000 results and was repeated 3 times for 3 consecutive pages of matches. The results are shown in Table 2. We have observed a significantly lower value for the most restrictive query for the smallest corpus, where the second and third pages of results were empty due to the limited number of matches (319).

**Table 2.** The Table shows a benchmark of search time for different queries and different corpus sizes (in tokens). The results presented are the average time in seconds that the API took to return results, with the standard deviation in parentheses.

| **Corpus size** | **Response time for a query** | | | |
|---|---|---|---|---|
| | `[]` | `<s/>` | `[upos="ADJ"]` | `<ne/> containing [upos="ADJ"]` |
| 147,580 | 4.17 (0.056) | 7.88 (0.28) | 4.14 (0.11) | 0.86 (0.74) |
| 1,475,800 | 4.19 (0.069) | 7.58 (0.26) | 4.13 (0.11) | 4.8 (0.84) |
| 14,758,000 | 4.92 (0.21) | 8.35 (0.33) | 4.19 (0.12) | 5.01 (0.7) |

## 5 Applying Korpusomat to Multilingual Multidisciplinary Research: A Case Study

This section introduces a brief case study in order to showcase potential uses of Korpusomat for multilingual analysis. The study, which, in a slightly modified version, forms part of a larger project [9], is located at the intersection between migration research and corpus-assisted discourse analysis. Its aim is to contribute to research investigating discursive representations of migrations during the "European refugee crisis" by determining whether there is a mismatch between the amount of media attention received by displaced people of different geographical identities and the actual numbers of refugees who filed for asylum in a European Union country between 2015 and 2018. This is expected to help establish which groups of refugees were backgrounded in European media discourses, despite significant numbers of their actual populations, and which groups were foregrounded and, thus, potentially problematised (as suggested by earlier research [8]).

To this end, Eurostat's statistical data on first time asylum applicants in the EU between 2015 and 2018[5] were contrasted with the results of a corpus

[5] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Asylum_statistics&oldid=558844.

analysis conducted on three sets of newspaper reports on the "refugee crisis", published in Poland, Spain, and the United Kingdom during the same time period. The corpus contains more than six million tokens in total, with the British and Spanish subsets being more or less equal in size (approx. 2.5 million tokens), whereas the number of tokens in the Polish subcorpus amounts to just over 1.2 million.

As the first step in the analysis, we used Korpusomat's concordancer to determine how British, Polish, and Spanish press referred to the displaced people. Specifically, we verified and compared the frequencies of candidate terms (such as *refugee, immigrant, migrant, asylum-seeker*, etc.). The two most frequent items in each subcorpus (see Table 3) were then selected for the subsequent stage of the analysis, where we searched for adjectives that modify them. Crucially, since UD parsing circumvents differences between languages and tagsets (as discussed above), a single query (`[upos="ADJ" & head.lemma="X"]`, where X stands for the search term) could be used for all three language subcorpora, regardless of syntactic divergences (in English adjectives usually precede the nouns they modify, whereas for Spanish the inverse is true) and inconsistencies (Polish, in turn, has a relatively flexible word order which is, therefore, more difficult to predict).

**Table 3.** Top terms used with respect to the displaced people.

| Corpus | Term | Abs. Freq. | Rel. Freq. |
|---|---|---|---|
| Poland | *uchodźca* | 9,604 | 7,316 |
| | *imigrant* | 3,272 | 249 |
| Spain | *refugiado* | 10,959 | 4,374 |
| | *inmigrante* | 6,075 | 2,424 |
| United Kingdom | *migrant* | 10,467 | 3,930 |
| | *refugee* | 9,695 | 3,640 |

As a result, we obtained lists of adjectives most frequently modifying the search terms (the cut-off point was set at the frequency per million of over 3). We then selected only adjectives denoting geographical identities from each of the six lists, and compared them to Eurostat's data on first time asylum applicants in the EU[6]. These results are presented in Table 4: the nationalities which are foregrounded in the corpus, compared to the "real-life" data, are marked in bold.

The results show that refugees and, to a lesser extent, (im)migrants were often discussed in terms of their geographical identities. At the same time,

[6] The top countries of origin of asylum applicants (over 30 thousand applications) were: Syria, Afghanistan, Iraq, Iran, Pakistan, Nigeria, Albania, Eritrea, Russia, Bangladesh, Guinea, Sudan, Turkey, Ukraine, Somalia, Georgia, Ivory Coast, Gambia, Venezuela, Algeria, Mali, Morocco, and Senegal.

**Table 4.** Top adjectives modifying the search terms.

| **Poland** | | **Spain** | | **UK** | |
|---|---|---|---|---|---|
| uchodźca | imigrant | refugiado | inmigrante | refugee | migrant |
| syryjski | arabski | sirio | subsahariano | Syrian | African |
| **polski** | **polski** | afgano | sirio | Afghan | Syrian |
| **czeczeński** | syryjski | **palestino** | **africano** | Sudanese | Afghan |
| afgański | afrykański | rohingyo | **europeo** | Iraqi | Sudanese |
| **palestyński** | | iraquí | marroquí | Iranian | Eritrean |
| | | eritreo | **magrebí** | Pakistani | Iraqi |
| | | **español** | | **Palestinian** | **Mediterranean** |
| | | somalí | | **African** | |
| | | | | Somali | |

there was a significant divergence between the purported nationalities of people labeled refugees versus immigrants/ migrants, especially in the Polish and Spanish corpora, where only one adjective appears on both lists (*syryjski/ sirio* ['Syrian']). In general, refugees tended to be more strongly associated with Syria and Afghanistan (the two top nationalities of first time asylum seekers, according to the Eurostat data), whereas (im)migrants were more likely to be described using more delegitimising vague geographical descriptions, such as *African/ afrykański/ africano, arabski* ['Arab'], or *subsahariano* ['Subsaharan'], despite the fact that among the top ten most numerous groups of asylum-seekers only two were of African origin. At the same time, the newspapers analysed had a tendency to background the specific African nationalities (Nigerians, who filed more than 140 thousand applications for asylum, are a case in point). This may suggest that they were, instead, referred to using the more general modifiers, which demonstrates a conflation of different identities into one. The examined newspapers also tended to foreground nationals of countries which remain in geographical or historical proximity to the countries where they were published (for example, Spanish newspapers paid special attention to Moroccans). On the other hand, migrants crossing the EU borders legally (such as, for instance, Albanians or Russians) were considered to be less newsworthy than those who resorted to irregular means of entry.

While this analysis can only provide a cursory look at the discursive phenomenon in question, it nevertheless points towards some prominent trends in European media representations of migrations during the so-called "refugee crisis". Above all, however, it demonstrates the usefulness of Korpusomat, especially for multilingual studies, where issues of replicability and comparability are of particular concern. Korpusomat is an important step towards reducing the impact of these issues by outflanking (as much as possible) differences between languages and tagsets and, thus, facilitating both the analytical process and the cross-linguistic comparison of results. This might be particularly helpful to

users with little to no technical expertise as well as limited experience with linguistic analysis, such as, for instance, scholars from other disciplines wishing to introduce aspects of corpus linguistics into their research.

## 6 Conclusions and Future Work

Korpusomat is a free platform for creating, processing, and analysing user-created text corpora in a variety of languages. It is scalable and modular, allowing for future user-base growth, and for including new potential processing pipelines. In the future, Korpusomat is also expected to support constituency parsing, parallel corpora, automatic speech recognition systems, and analytical tools such as keywords and terminology extraction, as well as quantitative measures of vocabulary richness.

**Acknowledgements.** This work was supported by the European Regional Development Fund as a part of 2014–2020 Smart Growth Operational Programme, CLARIN- Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19 and by the project co-financed by the Minister of Education and Science under the agreement 2022/WK/09.

## References

1. Anthony, L.: A critical look at software tools in corpus linguistics. Linguist. Res. **30**, 141–161 (2013)
2. Brouwer, M., Brugman, H., Kemps-Snijders, M.: MTAS: a solr/lucene based multi tier annotation search solution (2017)
3. Dunn, J.: Natural language processing for corpus linguistics. Elements in Corpus Linguistics, Cambridge University Press (2022). https://doi.org/10.1017/9781009070447
4. Kiera, W., Kobyliski, L.: Korpusomat-stan obecny i przyszo projektu. Jzyk Polski **CI**(2), 49–58 (2021)
5. Kiera, W., Kobyliski, L., Ogrodniczuk, M.: Korpusomat – a tool for creating searchable morphosyntactically tagged corpora. Comput. Methods Sci. Technol. **24**(1) (2018). https://doi.org/10.12921/cmst.2018.0000005
6. Kiera, W., Saputa, K., ukasz Kobyliski, Tuora, R.: Korpusomat.eu - user guide (Polish) (2022). https://korpusomat-eu.readthedocs.io/pl/latest/
7. McEnery, T., Brezina, V., Gablasova, D., Banerjee, J.: Corpus linguistics, learner corpora, and SLA: employing technology to analyze language use. Ann. Rev. Appl. Linguist. **39**, 74–92 (2019). https://doi.org/10.1017/S0267190519000096
8. Taylor, C.: Investigating the representation of migrants in the UK and Italian press: a cross-linguistic corpus-assisted discourse analysis. Int. J. Corpus Linguist. **19**(3), 368–400 (2014)
9. Zawadzka-Paluektau, N.: The “European refugee crisis" through a media lens. A cross-national study into press representations of displaced people combining corpus linguistics and argumentation analysis (PhD dissertation). University of Warsaw, University of Seville (forthcoming)