The Corpus of Contemporary Polish — a New Reference Corpus with Rich Syntactic Annotations

Witold Kieraś, Małgorzata Marciniak, Marcin Woliński, Katarzyna Krasnowska-Kieraś, Marek Łaziński
Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
{w.kieras, m.marciniak, m.wolinski, k.krasnowska-kieras, m.lazinski}@ipipan.waw.pl

Abstract

In the paper, we describe the Corpus of Contemporary Polish (KWJP) and its rich syntactic annotation. The corpus covers a wide range of text originally published between 2011 and 2020. Although it carries on the idea of providing up-to-date reference corpora of Polish initiated by the National Corpus of Polish (NKJP) project, the principles underlying its development are not the same. In this article, we outline the different choices that affect corpora content and give an explanation for them. The article focuses mainly on the description of annotation layers in KWJP which are generated with a neural network based tool specially developed for this purpose. We describe in details syntactic structure annotation, which is represented by hybrid trees combining information typical to constituency and dependency trees. Finally, we provide several examples showing how annotation with hybrid trees facilitates querying and effective searching for information in the corpus.

Keywords: reference corpus, Polish, multi-layer annotation, hybrid trees, parsebank

1. Introduction

The history of electronically developed linguistic corpora of Polish goes back to the late 1960s when a small 500K words corpus of samples from texts published between 1963 and 1967 was prepared for the purpose of gathering frequency data. In the 1990s and early 2000s a number of much larger corpora of Polish were developed by various institutions, but since its publication in 2012 the National Corpus of Polish (*Narodowy Korpus Języka Polskiego*, NKJP, Przepiórkowski et al. 2012) became the most widely used reference corpus of contemporary Polish. NKJP contains texts from various genres published since the beginning of the 20th century but the vast majority of the data covers the two decades of 1990s and 2000s. Since its release the content of NKJP has never been updated and the consortium developing the corpus became inactive, hence the need for a newer corpus arose.

In this paper, we present the Corpus of Contemporary Polish (*Korpus Współczesnego Języka Polskiego*, KWJP, Marciniak et al. 2023)[1] — a corpus originally published in December 2023, covering the period of 2011–2020 and aimed at continuing the effort of developing and updating a reference corpus of Polish enriched with various linguistic annotations. It is also envisioned as the first milestone towards a corpus resource updated in a 5-years cycle.

The KWJP corpus is available exclusively through a web-based search engine, as the vast majority of the texts included in the corpus are subject to copyrights of the publishing houses who donated them. Access to the search engine is unrestricted, and there is no limit on the number of query results for registered users. Registration is free of charge and does not require providing any additional data other than an email address.

[1] https://kwjp.pl

2. Corpus Contents

Although KWJP was conceived as a next decade counterpart to the discontinued National Corpus of Polish, it differs from its predecessor in several key aspects. A primary distinction in the development of KWJP lies in its scope: while NKJP aimed to represent written Polish from the early 20th century onwards, KWJP focuses exclusively on a segment of contemporary texts limited to a single decade. Furthermore, KWJP was developed within a significantly altered corpus infrastructure, which has grown and evolved considerably since the release of NKJP in 2012.

2.1. Data Selection

The KWJP corpus contains a narrower range of text types than NKJP. We made the decision to collect edited texts only, which resulted in the exclusion of spoken and internet texts from the corpus (these types of resources account for approximately 15% of NKJP). This decision was based on the distinct nature of these language varieties, necessitating the development of independent resources and processing pipelines. Therefore, strictly internet-based

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11585–11592
11–16 May 2026.

texts – forums, blogs, social media posts, etc. – were not included in KWJP. While acknowledging the diversity and dynamic nature of internet genres and the necessity of covering them in corpus data, we have decided to tackle this task in a separate project falling beyond the scope of this article. Some of the internet genres are also covered by the MoncoPL monitoring corpus (Pęzik, 2020).

NKJP includes spoken language represented by a relatively small subcorpus of transcribed conversations and a larger subcorpus of 'quasi-spoken' texts, primarily comprising records of proceedings from the Sejm, Senate, and parliamentary committees. Since 2011, an extensive, specialised Corpus of Parliamentary Discourse (Ogrodniczuk, 2018) has been developed, collecting transcriptions of parliamentary proceedings from both houses of the Polish Parliament dating back to 1919. Additionally, a spoken corpus[2] containing contemporary spontaneous conversations from YouTube channels was created in 2023 at the Institute of Polish Language of the Polish Academy of Science. To avoid redundancy with those existing resources, parliamentary and coversational data was not included in KWJP.

Given the temporal scope of KWJP – limited to texts from the preceding decade – data acquisition was conducted exclusively in electronic format. However, this did not eliminate the need for substantial manual labour. While some texts were available as directly convertible electronic files, others required semi-automatic conversion with manual correction (for books) or complete manual processing (for periodicals). The corpus includes only full texts of books originally written in Polish and first published between 2011 and 2020, although non-first editions were also admitted. In such instances, both the date of the first edition and the edition actually included in the corpus are recorded in the metadata. For periodicals, complete issues were prioritised, although data availability did not always permit this (affecting a small number of titles). Selected issues for the balanced corpus were chosen through stratified random sampling within each year of a given periodical's archive, maximising topical diversity. During processing, footers, announcements, advertisements, photo and illustration captions, very short notes, and translated texts (when identifiable) were excluded.

Regional periodicals were selected to ensure representation from all regions of the country, resulting in the inclusion of 130 different titles from 44 cities. We did not include periodicals printed outside of Poland in the corpus.

[2] https://korpusmowy.ijppan.pl

2.2. Text Classifications

The balanced corpus consists almost exclusively of books and periodical texts. Similar to NKJP, each text is accompanied by metadata including labels for text type and distribution channel. However, the repertoire of these labels is slightly different.

Firstly, text type distinction has been significantly simplified, and consists of three values: *fiction*, *non-fiction*, and *journalism*. *Fiction* primarily comprises literary books (novels and short story collections) representing various genres, and (to a limited extent) literary periodicals publishing predominantly short stories. This text type corresponds fairly well to NKJP's *fiction* label. The *non-fiction* type encompasses a broad range of non-fictional texts: journalist books, diaries, biographies, travel guides, popular science and scholarly texts, and official documents, which were gathered under several distinct classification labels in NKJP: *non-fiction*, *scientific and didactic texts*, *official documents*, *guide books*. Unlike NKJP, thematic magazines, such as economic, sports, popular science or beekeeping periodicals, also fall under the *non-fiction* category. Finally, the *journalism* genre encompasses exclusively typical news and public affairs press (mainly daily and weekly newspapers). These three general text types account for 30% (*fiction*), 35% (*non-fiction*), and 35% (*journalism*) of KWJP's balanced corpus.

Secondly, distribution channel classification was generally limited to two primary types: *books* and *press*. The press channel was further subdivided into *daily newspapers*, *weekly magazines*, *monthly periodicals*, and *other* publications. Only a small subset of texts (0.3%) was assigned the *internet* value for the distribution channel, representing a sample of court rulings from various instances – a type of official document. In all other cases, classification into *book* and *press* channels followed library classification schemes, specifically ISBNs (for books) and ISSNs (for press), even when the publication existed solely in electronic format. *Books* comprise approximately 55% of the balanced corpus, while the *press* accounts for slightly over 45% (*daily newspapers* representing just under 19%, *weekly magazines* approximately 12%, *monthly periodicals* 9.5%, and *other* press formats 5.5%). Table 1 summarises the corpus compostion.

Furthermore, books within the corpus are subject to an additional topical classification, consisting of brief descriptive labels indicating the literary genre (e.g., *crime fiction*, *science fiction*, *interview*, *biography*) or subject domain (e.g., *history*, *sociology*, *art*, *vegetarianism*). However, the values assigned within this topical classification are not a closed set, hence it should be considered only as partial and auxiliary.

by genre	
fiction	30%
non-fiction	35%
journalism	35%
by distribution channel	
books	55%
press	45%
daily newspapers	19%
weekly magazines	12%
monthly periodicals	9.5%
other	5.5%

Table 1: Text proportions in the balanced corpus.

2.3. Unbalanced Corpus

In addition to the balanced corpus, following the NKJP scheme, we also provide an unbalanced (so called 'opportunistic') corpus, comprising the majority of the data gathered during the course of the KWJP project. Although corpora of such type should not be trusted as sources of statistical information for quantitative studies of any kind, they may nevertheless be useful, e.g., in research on rare or new vocabulary or prove helpful for lexicographers searching for examples of the use of specific words and phrases. For these purposes, corpus users also frequently seek the largest possible corpora, prioritising their size over balanced composition.

The vast majority of this collection consists of national daily and weekly newspapers. For practical reasons, this corpus has been divided into five parts based on the criteria of genre and distribution channel (see Table 2). The first part is a corpus of fiction and non-fiction (ca. 200M tokens), comprising primarily thematic and literary periodicals, and to a much lesser extent, books. The rest of the data consists of press journalistic texts, divided into four subcorpora: the largest corpus of national daily newspapers (ca. 600M tokens), a corpus of regional daily newspapers (ca. 290M tokens), a corpus of weekly magazines (ca. 200M tokens), and other periodicals (35M tokens). The total size of the unbalanced corpus exceeds 1.4B tokens, making it over fourteen times larger than the balanced corpus. Unlike in the case of NKJP, the balanced and unbalanced corpora are disjoint; however, it is possible to search across both of them simultaneously.

subcorpus	tokens
fiction and non-fiction	200M
national daily newspapers	600M
regional daily newspapers	290M
weekly magazines	200M
other periodicals	35M

Table 2: Composition of the unbalanced corpus.

2.4. Subcorpus of Random Samples

A freely-available KWJP½M corpus was constructed to facilitate research and educational applications. This corpus comprises short text samples ranging from 40 to 60 words (up to a full sentence), randomly selected from a balanced source corpus to ensure proportional representation of its books and periodicals. KWJP½M comprises approximately 500K tokens and is provided without linguistic annotation (neither manual nor automatic), accompanied only by metadata. The corpus is distributed as JSON files containing complete source metadata and the extracted text fragments, and is suitable for applications not requiring full-text access.[3]

3. Corpus Annotation

The entirety of texts comprising the corpus (both balanced and opportunistic collections) are provided with several layers of automatically created linguistic annotations. KWJP generally follows the annotation scheme of NKJP (Przepiórkowski et al., 2012) with respect to segmentation, lemmatisation, morphosyntactic tagging and named entity marking. The rules for segmentation (tokenisation) are followed directly. As for morphosyntactic tagging, the tagset used has been made coherent with that of the latest version of the Polish morphological analyser Morfeusz (Kieraś and Woliński, 2017; Woliński, 2014). The modifications introduce several more fine-grained distinctions and concern the gender system, the treatment of so called collective numerals, nonstandard forms of adjectives, some noninflected classes and alien words. Some changes have also been introduced in the rules for lemmatising multi word proper names. The new tagset is entirely backward-compatible with the one used in NKJP. A detailed description and rationale behind the changes can be found in an article by Kieraś et al. (2021).[4]

[3] https://github.com/ipipan/kwjp100-varia

[4] NKJP annotated according to these rules is available at https://nkjp.nlp.ipipan.waw.pl/.

morphosyntax	
tokenisation (F1)	**99.83%**
tags (F1)	**97.99%**
lemmas (F1)	**98.48%**
dependency structures	
unlabelled edges (UAS)	**96.29%**
labelled edges (LAS)	**91.04%**
constituency structures	
unlabelled constituents (F1)	**97.80%**
labelled constituents (F1)	**97.77%**
with marked centres (F1)	**97.25%**

Table 3: Hydra's annotation quality for Polish.

Compared to NKJP, KWJP introduces two new annotation layers: named entities and syntactic structures. Annotation of named entities strictly follows the schema used in the one-million-word manually annotated subcorpus of NKJP called NKJP1M (Przepiórkowski et al., 2012). This annotation was never extended to the full NKJP corpus, while the automatic NE layer of KWJP covers the whole resource. The syntactic layer is entirely new with respect to NKJP, making KWJP the first large parsebank of Polish. We describe the syntactic annotation in further detail in the next section.

Most annotation layers in KWJP are generated with a tool named Hydra (Krasnowska-Kieraś and Woliński, 2023, 2024), the only exception being named entity recognition (NER) performed using PolDeepNer2.[5] As a neural network based tool, Hydra uses the Polish BERT model HerBERT (Mroczkowski et al., 2021) as a context-sensitive encoder. On top of it, a set of classification layers generates outputs used to construct lemmas, morphosyntactic tags, and elements of hybrid syntactic trees.

Models for segmentation, tagging, lemmatisation, and NER were trained and evaluated on the manually annotated NKJP1M following the split introduced by Wiącek et al. (2024). Hybrid trees that served as training material for the syntactic module (with 10% used for validation and 10% held out for evaluation) were created by merging constituency trees of Składnica (Woliński and Hajnicz, 2021; Woliński, 2019) and dependency trees of Polish Dependency Bank (Wróblewska, 2014). The measures of quality for the resulting annotations as shown in Table 3 are on state-of-the-art level for Polish.

[5] https://github.com/CLARIN-PL/PolDeepNer2

3.1. Hybrid Syntactic Trees

In KWJP, syntactic structures are represented using hybrid trees combining information typical to constituency and dependency trees. An example of such a structure is shown in Fig. 1. At first glance, this is a constituency tree. The leaves of the tree correspond to inflected forms characterised by their text token, lemma, and morphosyntactic tag. The internal nodes of the tree represent constituents of various levels: S – sentence, NP – nominal phrase, VP – verbal phrase, AdjP – adjectival phrase, N – noun form, V – verb form, etc.

What is not typical for a plain constituency tree is that syntactic centres are explicit in this visualisation. The central constituent is placed directly below its parent in the tree. In the example, the central constituent of the entire utterance (marked with the symbol ROOT) is the sentence S, whose centre is the verbal phrase, and whose centre, in turn, is the conjunction.[6]

The non-central branches of the tree are labeled with dependency relations, which allows to read out the dependency structure. In the tree in Fig. 1, a subj label is placed on the edge connecting the node S, located above the conjunction *i* 'and', with the node NP over the noun *Jaskrawość* 'brightness'. This indicates that there is a dependency edge labelled subj joining the conjunction *i* with the token *Jaskrawość*. The dependency relations can also be seen as joining constituents (phrases) and not words. In the running example, the label subj tells us that the nominal phrase NP yielding *Jaskrawość naszego świata* plays the role of the subject for the sentence S (or for the coordinated verbal phrase *brzydzi i odrzuca* 'disgusts and repulses').

Users accustomed to standard dependency structures can choose an alternative display, as shown in Fig. 2.

3.2. Indexing Syntactic Structures for Search

KWJP is made available via MTAS corpus query tool (Brouwer et al., 2017), which provides a reasonable compromise between search speed and the power of the query language. The engine uses a dialect of Corpus Query Language (CQL), which is unfortunately not tree oriented. To provide partial access to the syntactic annotation we use the engine's ability to assign arbitrary features to tokens and to index ranges of tokens.

The following features (illustrated with corpus queries for their particular values) are indexed for

[6] We adopt a surface-syntactic paradigm and therefore assume that conjunctions are centres of coordinated structures, prepositions are centres for prepositional-nominal phrases, and complementisers are centres for subordinate clauses.

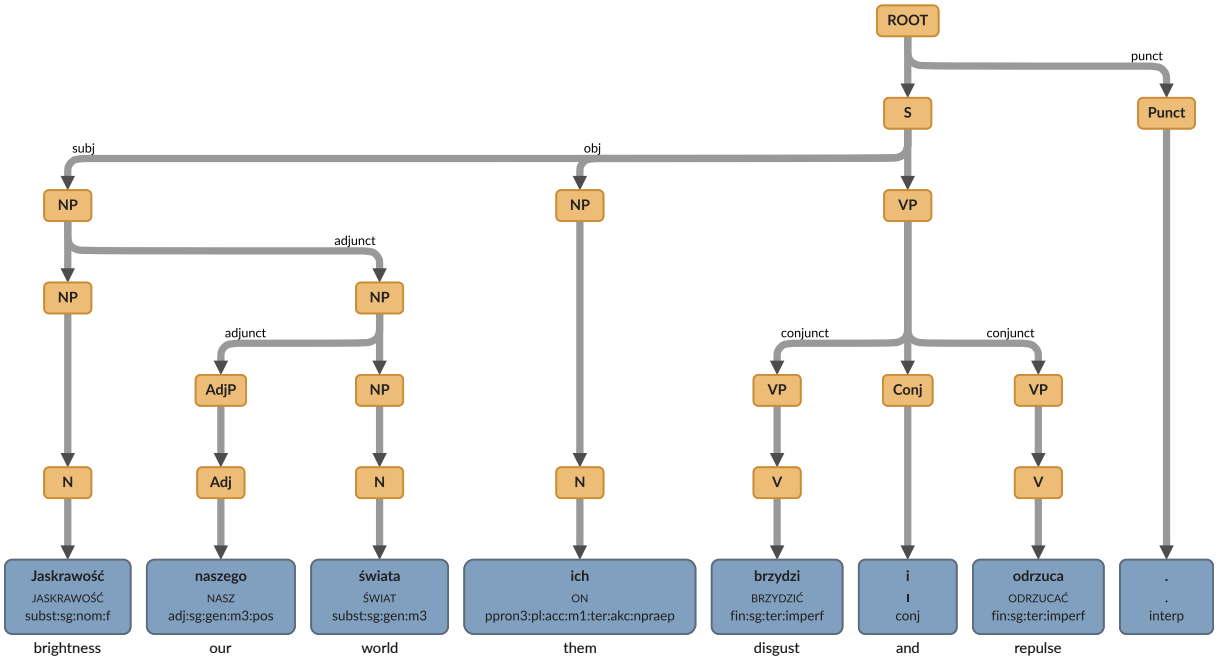

Figure 1: Hybrid syntactic tree for the sentence *Jaskrawość naszego świata ich brzydzi i odrzuca.* 'The brightness of our world disgusts and repulses them.' (from *Opowiadania bizarne* by Olga Tokarczuk, 2018)

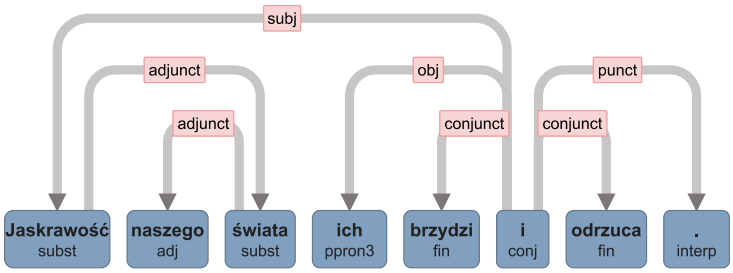

Figure 2: Dependency tree for the sentence of Fig. 1

all tokens in the corpus:

- `[deprel="subj"]` — the label of the dependency edge leading to the given token,
- `[head.lemma="wskocz"]` — the lemma of the immediate parent of the searched token,
- `[head.pos="prep"]` — the grammatical class of the parent,
- `[head.feat="acc"]` — one of the grammatical category values of the parent,
- `[head.position="left"]` — the position (left or right) of the parent in the linear order,
- `[head.distance="5"]` — the distance (measured in tokens) of the parent from the matching token.

The above features combined with the morphosyntactic ones for the current token allow to describe in all detail a single dependency edge in a tree.

Indexing also includes phrases from the constituency tree characterised with their type. For efficiency reasons, only maximal phrases of a given type are indexed, so in the case of the tree of Fig 1 the NP *Jaskrawość naszego świata* will be available for query, while its constituent NPs *Jaskrawość* and *naszego świata* will not. The query for a constituent of a given type has the form `<c="NP" />`.

A similar form of query is also used for named entites. For example the query:

```
<ne="persName" />
```

will match all named entities of type `persName` (person names).

In addition to syntactic information accessible through corpus queries, for each sentence in the corpus the user may display an interactive syntactic tree in both variants — pure dependency (as shown in Fig. 2) and hybrid (cf. Fig. 1).

4. Querying the Corpus

In this section, we provide several examples showing how annotation with hybrid trees facilitates querying and effective searching for information in the corpus.[7]

Let us assume that we would like to examine the contexts in which the word *teoria* 'theory' appears in our corpus. If we run the following query:

```
[head.lemma="teoria" &
    deprel="adjunct"]
```

we obtain all tokens modifying (more precisely: the heads of modifying phrases) the word *teoria* 'theory' in all its grammatical forms. The query returns tokens bearing different parts of speech, e.g., in the phrase *odczytanymi zupełnie na opak teoriami Darwina* 'completely misinterpreted (lit. ‹read backwards›) theories of Darwin' it selects the past participle *odczytanymi* 'read' and the noun *Darwina*. If we are only interested in adjectival modifiers, the following question restricts the results accordingly:

```
[head.lemma="teoria" &
    deprel="adjunct" & pos="adj"]
```

Note that the query also finds adjectives which are not directly adjacent to their syntactic head. In the phrase *zgodną z duchem czasu teorię* (lit. ‹consistent with the zeitgeist theory›) 'a theory consistent with the zeitgeist' it selects *zgodną* 'consistent' and in the phrase *teoria drastycznie sprzeczna* (lit. ‹theory drastically contradictory›) 'a drastically contradictory theory' it selects *sprzeczną* 'contradictory'.

If we want to search the corpus for phrases (not only tokens as in the previous examples) we can use the expression `<c/>` to match any constituent or limit the search to specific constituent types: all nominal phrases can be selected by `<c="NP" />`. These very general queries match vast amounts of corpus material and can be quite time-inefficient when executed on large resources like KWJP. For more interesting results, the search can be further restricted with a query referring to phrase components. This can be achieved with the operator `containing`. However, it should be noted that these components can be matched at any level of the structure. For instance, the following query:

```
((<c="NP" /> containing
    [head.lemma="teoria" &
        deprel="adjunct"])
```

finds all noun phrases containing subphrases that meet the condition described in the first example of this section, including, e.g., *najdoskonalszą literacko ilustrację marksistowskiej teorii* 'the most perfect literary illustration of the Marxist theory' built around the noun *ilustracją* 'illustration' with *teorii* 'theory' nested as a modifier. To remove this redundancy and to find only phrases being actual modifiers of the word *teoria* 'theory' we can prohibit this token from occurring in the matched phrases with the negation operator:

```
((<c="NP" /> containing
    [head.lemma="teoria" &
        deprel="adjunct"])
    !containing [lemma="teoria"])
```

From the text passage *teorie gospodarcze wzrostu przyspieszonego* 'economic theories of accelerated growth', the above query selects only the phrase *wzrostu przyspieszonego* 'accelerated growth'.

Suppose now we want to examine whose theories are mentioned in the corpus. The `within` operator allows to select all named entities (`ne`) of `persName` (person name) type that occur inside noun phrases that modify the word 'theory' (this includes named entities constituting the whole NP):

```
<ne="persName" /> within
((<c="NP" /> containing
    [head.lemma="teoria" &
        deprel="adjunct"])
    !containing [lemma="teoria"])
```

From the excerpt *z popularnymi wówczas teoriami systemu światowego Immanuela Wallersteina, Andre Gundera Franka i Samira Amina* 'with the then popular world system theories of Immanuel Wallerstein, Andre Gunder Frank and Samir Amin' the above query selects three names (in the genitive case): *Immanuela Wallersteina*, *Andre Gundera Franka*, and *Samira Amina*.

As another example let us consider a certain behaviour of subjects in Polish: they are often not mandatorily present in the sentence. Polish verbal forms carry the information about person, number, and (for some forms) gender. This makes some subjects dispensable, in particular those expressed with semantically light pronouns. Nonetheless, we can find personal pronouns in subject position in KWJP:[8]

```
[pos="ppron(12|3)" & deprel="subj"]
```

Obviously, subjects of verbs in the first and second person are typically pronouns, because regular nouns generally cannot appear at this position. However, the results of the following search are not empty:

```
[deprel="subj" & !pos="ppron12"
   &  head.feat="(pri|sec)"]
```

The corpus interface allows to group results, revealing that the most common non-pronoun is the word *wszyscy* 'all' (2450 occurrences, in the meaning

[7] All examples in this section refer to the balanced corpus.

[8] In the KWJP tagset first and second person pronouns belong to the grammatical class `ppron12`, and third person pronouns are marked as `ppron3`.

'you all' or 'we all'); *oba* 'both' (713 occ.; 'we/you both'); *i* 'and' heading coordinated phrases like *ja i wielu moich znajomych* 'myself and my many friends' (497 occ.) etc.

In order to examine actual sentences without an explicit subject, we can execute the following query:

```
<c="S"/> !containing [deprel="subj"]
```

yielding 3,970,145 matches. If we are only interested in complete sentences, we can additionally require that the results contain the dependency root:

```
(<c="S"/>
    !containing [deprel="subj"])
    containing [deprel="root"]
```

It limits the results to 1,675,882 complete sentences with omitted subject.

5. Further Developments

From its very beginning the KWJP project was aimed not only at building a new standalone corpus but also at creating a team and infrastructure for the long-term development of the resource. Since the release of the first version the team planned to extend the original 2011–2020 corpus in 5-year cycles similar to the SYN corpora series developed by the Czech National Corpus team (Křen et al., 2016). Currently, work is ongoing to collect texts originally published in the 2021–2025 period. So far approximately one thousand books and a number of various periodicals have been acquired from the publishers. Only a fraction of this data will be selected for the 2021–25 balanced corpus, whereas the remaining texts will enrich the opportunistic unbalanced corpus. The estimated minimal size of the 2021–25 update is 50 million tokens divided in the same proportions as the original KWJP 2011–20 corpus. The update is expected to be released in mid-2026.

Apart from extending the corpus, technical efforts aimed at enhancing the searching efficiency and expressive power of query language are also underway. The primary goal is to extend the syntactic annotation indexed in the search engine to allow more flexible queries involving longer sequences of dependency tree nodes. However, another long-term goal is to deploy an independent syntax-oriented search engine to allow corpus users to take advantage of full syntactic information (both dependency and constituency) contained in the hybrid syntactic trees, thus enabling truly recursive corpus queries.

Finally, the team is working on enhancing the usability of corpus search by leveraging Large Language Model solutions. This includes both the automatic construction of CQL queries based on natural language input and the managing and grouping of search results based on vector embeddings of concordances.

6. Acknowledgements

This article was written as a result of project no. POIR.04.02.00 00-D006/20 — Digital Research Infrastructure for Arts and Humanities DARIAH-PL, co-financed in the amount of PLN 99,800,000 from European Funds — European Regional Development Fund (ERDF) under Action 4.2 of the Smart Growth Operational Programme (SGOP), whose beneficiary is ICHB PAN PCSS, together with 15 partners of the DARIAH-PL consortium, including IPI PAN. The objective of the project is to build Dariah.lab — a research infrastructure for digital humanities. The infrastructure has been placed on the Polish Research Infrastructure Map in the area of social sciences and humanities.

7. Bibliographical References

Matthijs Brouwer, H. Brugman, and M. Kemps-Snijders. 2017. MTAS: A Solr/Lucene based Multi Tier Annotation Search solution. In *Selected papers from the CLARIN Annual Conference 2016. Linköping Electronic Conference Proceedings*.

Witold Kieraś and Marcin Woliński. 2017. Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Polski*, XCVII(1):75–83.

Witold Kieraś, Marcin Woliński, and Bartłomiej Nitoń. 2021. Nowe wielowarstwowe znakowanie lingwistyczne zrównoważonego Narodowego Korpusu Języka Polskiego. *Język Polski*, CI(2):59–70.

Katarzyna Krasnowska-Kieraś and Marcin Woliński. 2023. Constituency parsing with spines and attachments. In *Computational Science – ICCS 2023. 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part I*, volume 14073 of *Lecture Notes in Computer Science*, pages 191–205, Cham, Switzerland. Springer Nature Switzerland.

Katarzyna Krasnowska-Kieraś and Marcin Woliński. 2024. Parsing headed constituencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12633–12643, Turin, Italy. ELRA and ICCL.

Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská,

Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička, and Adrian Jan Zasina. 2016. SYN2015: Representative corpus of contemporary written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2522–2528, Portorož, Slovenia. European Language Resources Association (ELRA).

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Maciej Ogrodniczuk. 2018. Polish Parliamentary Corpus. In *Proceedings of the LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, pages 15–19, Paris. European Language Resources Association (ELRA).

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Piotr Pęzik. 2020. Budowa i zastosowania korpusu monitorującego MoncoPL. *Forum Lingwistyczne*, 7(7):133–150.

Martyna Wiącek, Piotr Rybak, Łukasz Pszenny, and Alina Wróblewska. 2024. NLPre: A revised approach towards language-centric benchmarking of natural language preprocessing systems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12271–12287, Turin, Italy. ELRA and ICCL.

Marcin Woliński. 2014. Morfeusz reloaded. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. European Language Resources Association (ELRA).

Marcin Woliński. 2019. *Automatyczna analiza składnikowa języka polskiego*. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw.

Marcin Woliński and Elżbieta Hajnicz. 2021. Składnica: a constituency treebank of Polish harmonised with the Walenty valency dictionary. *Language Resources and Evaluation*, 55:209–239.

Alina Wróblewska. 2014. *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

8. Language Resource References

Marciniak, M. and Kieraś, W. and Bojałkowska, K. and Borkowski, P. and Borys, M. and Eźlakowski, W. and Guz, W. and Kobyliński, Ł. and Komosińska, D. and Krasnowska-Kieraś, K. and Łaziński, M. and Miernecka, M. and Nitoń, B. and Ogrodniczuk, M. and Rudolf, M. and Tomaszewska, A. and Woliński, M. and Wołoszyn, J. and Wójtowicz, B. and Wróblewska, A. and Zawadzka-Paluektau, N. 2023. *Korpus Współczesnego Języka Polskiego*. Instytut Podstaw Informatyki PAN, Warszawa. PID https://kwjp.pl.