Lang Resources & Evaluation (2022) 56:309–332
https://doi.org/10.1007/s10579-021-09549-1

PROJECT NOTES

The Electronic Corpus of 17th- and 18th-century Polish Texts

Włodzimierz Gruszczyński[1] · Dorota Adamiec[1] · Renata Bronikowska[1] · Witold Kieraś[2] · Emanuel Modrzejewski[1] · Aleksandra Wieczorek[1] · Marcin Woliński[2]

Accepted: 15 June 2021 / Published online: 18 September 2021

Abstract The paper describes the process of building the electronic corpus of 17th- and 18th-century Polish texts, a relatively large, balanced, structurally and morphologically annotated resource of the Middle Polish language, available for searching at https://www.korba.edu.pl. The corpus consists of samples extracted from over seven hundred texts written and published between 1601 and 1772, summing up to a total size of 13.5 million tokens which makes it one of the largest historical corpora for a Slavic language.

Keywords Historical corpora · Corpus construction · Corpus annotation · Middle Polish

1 Introduction

This article presents a 13.5-million-token corpus of Polish texts covering the period between 1601 and 1772.[1] In the 17th and 18th centuries, several important grammatical categories of the Polish language developed and others disappeared. This period also brought a number of new lexical borrowings (especially from Latin, German, French, and Turkish). There was an evolution of style and syntax.

[1] The epochs before the year 1601 are covered by other Polish historical corpora (cf. Section 2). As for the end date, we refer to the tradition of considering the year 1772, when the Polish statehood began to collapse, as a symbolic date of the beginning of the New Polish period (cf. Długosz-Kurczabowa and Dubisz, 2001, p. 56), although of course it is not a real turning point in the development of the Polish language.

✉ Aleksandra Wieczorek
aleksandra.wieczorek@ijp.pan.pl

[1] Institute of Polish Language, Polish Academy of Sciences, Cracow, Poland

[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

For these reasons, texts from this period are an important source for research on the history of the Polish language. They can also be useful in the field of history, culture, literature, history of science and others. The informal name for the corpus, which will be used throughout this article, is *KorBa*—an abbreviation of Polish *Korpus Barokowy* ('Baroque Corpus').[2]

The corpus has been available for online search since 2018[3] at https://korba.edu.pl/ via MTAS (Multi Tier Annotation Search) search engine (Brouwer et al., 2017). We also provide (under the Creative Commons Attribution Share Alike licence) the source files of the manually annotated 500-thousand-token subcorpus which contains 200-word samples of texts included in the corpus (https://www.korba.edu.pl/download). As for the source files of the entire corpus, we will provide those texts that are not copyrighted.[4] On the KorBa website there will also be links to scans of old prints available in digital libraries underpinning texts in the corpus.

KorBa is the first relatively large corpus of old Polish texts and the only morphosyntactically annotated (including lemmatization)[5] online corpus of pre-19th-century texts of such size in the Slavic world. It contains diverse texts, both in their transliterated and transcribed form.[6] The metadata, structural markup,[7] and morphosyntactic annotation enable a variety of queries, filtering of results, and locating them within the source down to the page number.

The article is divided into seven parts. The first two sections are introductory in nature. Section 3 is dedicated to the source material within the corpus. It presents the main principles behind the selection of the texts, their diversity in time and location of origin, genres, and topics, the percentage shares of various types of texts in the corpus, as well as the metadata used in their description. Section 4 discusses two methods of text rendering (transliteration and transcription) and individual layers of text marking (structural and language markup, morphosyntactic annotation). Section 5 describes the creation of the corpus, from the conversion of transliterated texts into a transcribed form through a morphological analysis to tagging (disambiguation).[8] It also presents the tools used in that process, mostly based on existing solutions adapted to the specific features of our historical corpus.

[2] The corpus covers a period mostly dominated in Polish literature by the Baroque style (cf. Hernas, 2002, p. 20).

[3] The first version of the corpus was created as part of the project carried out in the years 2013–2018. The second stage of the project, which has been running since 2019, is expected to end in 2023. The volume of the corpus will increase, and it will, among others, contain texts from the period 1773–1800, previously not included. Work is also underway to integrate the corpus with other electronic resources.

[4] Copyright laws in Poland prohibit making full texts available for publication for 70 years after the author's or the editor's death. It also applies to modernized editions of old texts.

[5] The term 'morphosyntactic annotation' applies to the process of lemmatization and inflection marking (whether manual or automatic one), as well as its result (for more details see Sect. 4.4).

[6] Historical texts often contain some characters not used in the contemporary language. Thus, the electronic rendering of the original is already considered a transliteration (see Sect. 4.1). The transcription brings the texts closer to the contemporary spelling (see Sect. 4.2).

[7] By 'structural markup' we mean marking of the division of the text into pages, chapters, etc., as well as marking of parts outside the main text, such as cover page elements, footnotes, marginal notes, etc.

[8] We adhere to the following use of the terms 'morphological analysis' and 'tagging': Morphological analysis assigns all possible grammatical interpretations to a textual word without paying attention to the

Section 6 presents an example of searching the corpus via MTAS search engine. Section 7 contains conclusions and outlines further developments planned for the corpus.

2 Related work

The first historical corpora were created for English (Kroch et al., 2004; Rayson et al., 2007). More have appeared since, e.g. for German (Scheible et al., 2011) or Swedish (Borin et al., 2012). The relatively large corpora for global languages, such as English—Early English Books Online (ca. 755 million words, 16th–17th c., cf. EEBO[9]) and Corpus of Historical American English (ca. 400 million words, 19th–20th century, cf. COHA[10])—or Spanish (over 100 million words, cf. CdEGH[11]) are particularly notable in comparison to most historical corpora. The more comprehensive list of historical corpora and related literature can be found, e.g. at https://www.clarin.eu/content/historical-corpora.

Morphosyntactic annotation is an essential feature for the corpora of Slavic languages (most of which are highly inflective). In the Slavic world, morphosyntactically annotated historical corpora exist for Russian and Slovenian. In fact, Russian has several historical corpora. The oldest texts include an annotated corpus of birch bark manuscripts (11th–15th c.) of around 19.5 thousand words[12] and an annotated corpus of 11th–14th c. Russian texts of c. 570 thousand words. Besides these, there is also an unannotated corpus of 14th–17th c. texts, containing more than 8 million words. Russian texts from the 18th century onwards are included in the Russian National Corpus (RNC[13]) containing 7.2 million words for the 1700s and 7.9 million words for the 1800s (cf. Mishina & Pichkhadze, 2015; Dobrushina et al., 2015; Sichinava, 2016). The automatically annotated Slovenian corpus of texts from 1584 to 1918 (a vast majority written after 1850) numbers 15 million tokens, whereas the manually annotated subcorpus—300 thousand tokens (Erjavec, 2015).

Other historical corpora of Slavic languages are not morphosyntactically annotated. Major projects include the historical subcorpus of the Czech National Corpus (CNC[14]) from 14th to 20th c. of over 4 million tokens (Kučera & Stluka,

Footnote 8 continued
context of its use. Morphosyntactic tagging selects the interpretation that is correct in a given context from those provided by a morphological analysis.

[9] https://www.english-corpora.org/eebo/

[10] https://www.english-corpora.org/coha/

[11] Corpus del Español: Genre/Historical. https://www.corpusdelespanol.org/hist-gen/

[12] The size of the corpora in question is listed as the number of words or the number of tokens, depending on the type of information provided in publications regarding the given corpus. The figures on the sizes of the Russian National Corpus subcorpora are quoted from https://ruscorpora.ru/new/corpora-stat.html, as of December 22, 2020.

[13] http://www.ruscorpora.ru

[14] https://www.korpus.cz/

2011; Kučera et al., 2015), the other two Czech historical corpora (StaTB[15], till the end of 15th c., 5.9 million tokens, and StrTB[16], 16th–18th c., 930 thousand tokens) and three historical subcorpora of the Slovak National Corpus (864–1843, containing 2.1 million tokens, 1843–1954, containing 24 million tokens, and the Historický korpus slovenčiny, a subcorpus of the Slovak National Corpus, hereafter SNC[17], containing 917 thousand tokens from the 15th–18th c.; Garabík & Kajanová, 2015). The work on automatic annotation of the StaTB is currently ongoing (Jínova et al., 2014). All the above-mentioned Slavic corpora are available for searching via search engines (some of them after registration); the Slovene corpus is also available for downloading.

The first historical corpus of Polish is a corpus of texts from the years 1572–1756 created by the IMPACT project (Bień, 2014). It contains 1.6 million tokens and comprises DjVu format scans linked to transliterated texts. There is also a small but very diverse corpus of Polish texts from the 19th century (Derwojedowa, 2020), half of which has been manually marked in its inflectional layer (Kieraś & Woliński, 2018). Besides KorBa, the Corpus of Polish up to 1500 (Korpus Polszczyzny do 1500 roku[18]; Deptuchowa et al., 2020) and the Corpus of 16th-century Polish (Korpus Polszczyzny XVI wieku[19]; Opaliński & Potoniec, 2020) are currently in development as well. In the future the authors plan to integrate the above-mentioned Polish historical corpora, as well as the contemporary National Corpus of Polish (Narodowy Korpus Języka Polskiego, hereafter NKJP[20]) under one common search engine (Król et al., 2019).

3 Texts

According to the literature, there are far more gaps in the textual coverage of historical corpora in comparison to modern language ones (cf. Kytö, 2011: p. 430). The reasons for these shortages can be divided into two groups: one of them refers to the limited access to the literary production of a given period, the other one—to the limited knowledge about the language of a given epoch conveyed by the collected texts.

Regarding the first type of constraints, many texts have been lost due to different catastrophes that have taken place since their creation. This is of particular importance in the case of Polish experiences of long-lasting wars and occupations combined with the destruction and seizure of national cultural resources.

[15] Staročeská textová banka [online]. Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka. Version 1.1.15. http://vokabular.ujc.cas.cz/banka.aspx?idz=STB

[16] Středněčeská textová banka [online]. Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka. Version 1.1.15. http://vokabular.ujc.cas.cz/banka.aspx?idz=SDTB

[17] https://korpus.sk; historical corpora: https://korpus.sk/old1.html, https://korpus.sk/old2.html, https://korpus.sk/hks.html

[18] https://ijp.pan.pl/nauka-ibadania/projekty/projekty-realizowane/baza-leksykalna-sredniowiecznej-polszczyzny-do-1500-rokufleksja/

[19] http://spxvi.edu.pl/korpus/

[20] http://nkjp.pl/

Additionally, numerous texts are available only in a manuscript form, which makes it difficult to change them into the editable version (although with the development of the HTR technology it is becoming much easier now to compile them in a corpus).

The limitations of the other type result either from an underrepresentation of particular types of texts in historical corpora (e.g. texts written by women or people from lower social strata) or an uncertainty about the authorship of a given text or its fragments—it applies, e.g. to later copies and editions, which can be significantly changed in comparison with the original. For example, the 19th-century editions of Polish old texts are characterized by editors' significant interference in the original texts in line with the 19th-century tendency to modernize their spelling and inflection.

For all these reasons, accomplishing the fundamental objectives of language corpora, balance and representativeness, is far more difficult in a historical corpus than in a corpus of a modern language. As for the period covered by KorBa, the best preserved works tend to be literary, since those, unlike more utilitarian texts, were frequently re-released. Therefore, it was difficult to limit the share of literary texts to less than twenty percent of the corpus, as recommended for contemporary corpora (cf. Przepiórkowski et al., 2012, p. 34). Likewise, achieving another goal in the creation of corpora, which is to reflect what types of texts are (or were) most read in a given society (this was followed, e.g. in BNC[21], cf. Nelson, 2010, p. 58, and in NKJP, cf. Przepiórkowski et al., 2012, p. 27–30), was hampered by the fact that our knowledge in this regard for that historical period is vanishingly small and gained entirely in an indirect manner. For example, we can assume that the texts published several times were popular and read by a larger number of recipients.

3.1 Selection and classification

The corpus includes four types of sources: old prints, manuscripts, 19th-century editions, and modern editions. Original texts (manuscripts and old prints) combine to account for 64% of the corpus. As many important works have not survived, it seemed preferable to include a later edition, even as imperfect as those from the 19th century, rather than omit them entirely.

Texts have been selected for the corpus with the aim of maintaining a diversity of periods, places, types, genres,[22] and subjects. The time range of 172 years covered by the corpus has been divided into four completely arbitrary periods—1601–1650, 1651–1700, 1701–1750, and 1751–1772. Table 1 presents their representation within the corpus. The largest number of tokens comes from the first half of the 19th century, as that period saw many large and important texts that remained popular throughout Polish Baroque. The relative under-representation of the first half of the 18th century results from the political and cultural crisis persisting in Poland at the time, which led to a decline in publishing.

[21] British National Corpus. http://www.natcorp.ox.ac.uk/

[22] According to some similarities in form, style, or subject matter, the texts are divided into eleven major groups called here text types and further subdivided into genres.

Table 1 Chronological representation of texts

Period	Fraction of the corpus
1601–1650	38.4%
1651–1700	29.2%
1701–1750	16.3%
1751–1772	16.1%

In qualifying texts for the corpus, the authors also maintained diversity by their area of origin. Historically, the Polish-Lithuanian Commonwealth of the time can be divided into the following regions[23]: Lesser Poland, Mazovia, Greater Poland, Ruthenian Lands, Grand Duchy of Lithuania, Livonia, Silesia (which, despite being outside the borders of the Commonwealth, witnessed a relatively widespread use of Polish and a number of publications in the language). The texts gathered in the corpus were assigned to those regions based on the place of publication, or, if the original edition was unavailable, the place of writing (if known). Polish texts published abroad at the time (e.g. in Leipzig) constitute a separate class. The geographical distribution of corpus texts is shown in Fig. 1. It is, first and foremost, a product of the activity of leading publishing centres.

The corpus texts were classified into eleven types (including four literary, six non-literary, and the Bible,[24] cf. Figure 2). The classification of literary types is consistent with that adopted by literary studies and the whole division corresponds—as far as possible—to that applied to modern texts in NKJP (Przepiórkowski et al., 2012, pp. 15 and 33). However, it was not possible to avoid the differences altogether, as they result from a different structure of the literary production and readership at periods separated by three hundred years. Our corpus, for obvious reasons, does not contain spoken or internet texts. The share of press texts is smaller than in NKJP, as the Polish press was only being created at that time. Literary texts account for 23.4% of the corpus, while non-literary texts for 74.2%, with the remaining 2.4% being the Bible.

As for a more detailed classification, the types are divided into genres (cf. Table 2 below).[25]

3.2 Metadata

The corpus is enriched with metadata—various information about every text, allowing the user to filter search results. It includes bibliographic data and other information described in Sect. 3.1.

[23] This division is based on Zofia Florczak's book on the participation of individual regions of Poland in shaping Polish literature of the sixteenth century (Florczak, 1967). Her map, however, does not include Livonia and Silesia, featured here.

[24] Particular fragments of the Bible belong to various literary and non-literary types and genres; thus, it appeared best to place them in a separate group.

[25] As the second edition of the project is ongoing and the new texts are still being added to the corpus, the list of genres will be enlarged.

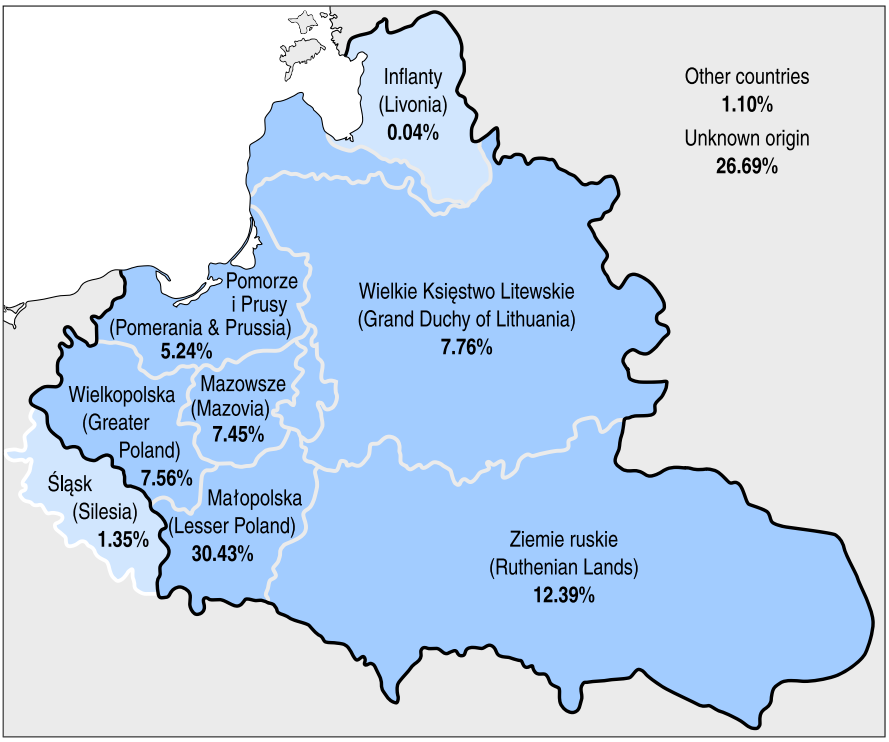

Fig. 1 Geographical distribution of texts in the corpus displayed on the map of the Commonwealth after the Union of Lublin of 1569

We have included the following metadata for each text: unique identifier, title, author, translator (for translated texts), date and place of publication, printing house and area of origin. Of course, not all the information is available for every text; some works are marked as anonymous, with an unknown place of publication, or with an unknown or approximate date of publication. Editions from the 19th century or later are appropriately marked and provided with the bibliographic data of the modern version. Metadata allows the user to, for example, search for texts from a given time frame, author, or region. Filtering for places can be used for tracking dialectal diversity in texts, while narrowing down searches to time frames allows the user to observe linguistic developments over particular periods.

All texts have also been appended with data on their stylistics and genre. They are marked for mode of representation of speech (rhymed, non-rhymed, mixed texts), type of text, genre, subject matter, and whether they are humorous or not. The latter category includes various satirical texts and is meant to allow for research into a humorous, even idiolectal, usage of the language. The division into rhymed and non-rhymed texts may be helpful for research as well, since the use of linguistic means in poetry tends to be subordinated to rhymes and rhythm.

Assigning a text to a genre unequivocally was frequently problematic; describing the subject matter would occasionally prove even harder. Only for some types of

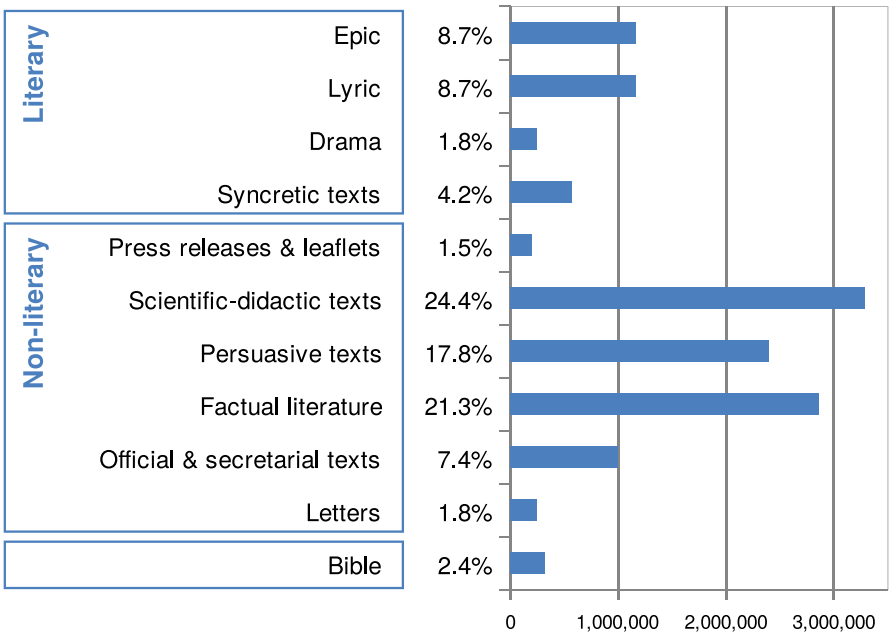

Fig. 2 Types of texts

works the matter was clear (e.g. for various scientific texts—astronomy, biology, physics, mathematics, etc., for parliamentary acts—politics and law, for sermons—religion). It was possible to choose more than one genre or subject matter for a given work. This was most typically justified in the collections of poems, which may include songs, epic poems, satires, hagiographies, etc. Regarding the subject matter, in some cases, like press releases, none was chosen, since they cover many different topics. We are aware of some research problems this may cause; nevertheless, it is the only solution if we assign subjects to the whole texts.

4 Electronic representation of texts

4.1 Transliteration layer

For this corpus, the texts have been transliterated according to the principles based on the editorial rules for historical Polish (Górski et al., 1955). Original spelling of editions and manuscripts from the 17th and 18th centuries was preserved, with the only change being the standardisation of diacritics for a given function (e.g. the letter *ż* is always written with a dot, even though originals occasionally use the forms *ž* or *z̄*). Ligatures are decomposed into separate letters (e.g. *ß* as *sz*[26]). Other features of original spelling incompatible with modern rules were preserved, e.g.

[26] Only for texts in Polish. *ß* was preserved in foreign language fragments.

Table 2 Types and genres in the corpus

Types	Genres
Epic	Epic poems, fables, hagiographies, parables and specula (mirrors), romances
Lyric	Carols and folk carols, emblems, epigrams, epithalamia, epitaphs, laments, odes, panegyrics, psalms, riddles, songs, sonnets
Drama	Comedies, dialogues, nativity plays, tragedies
Syncretic texts	Pastorals, satires
Scientific-didactic texts	Calendars, culinary recipes, encyclopedias and compendiums, guidebooks, guides, herbaria, instructions, lectures, phrasebooks, textbooks, treatises
Persuasive texts	Dedications, sermons, political speeches, proverbs, speeches for various occasions, writings on political and social topics, writings on religious topics
Factual literature	Accounts of events, chronicles, descriptions of journeys, geographical descriptions, memoirs, rolls of arms
Official & secretarial texts	City laws, contracts, documents of regional parliaments, inventories, judicial records, official letters, parliamentary bills, prenuptial agreements, privileges and charters, Sejm journals, Sejm texts, testaments
Press releases and leaflets	–
Letters	–
Bible	–

using letters ***ś***, ***ź***, ***ć*** before the letter ***i***, the digraph ***cż*** instead of ***cz***, the original use of letters ***y*** and ***i***, and acute accents over ***a*** and ***e*** (for examples, see Sect. 4.2). Original spacing and capitalisation were also preserved. Any abbreviations were left as per the original. In the texts obtained from 19th-, 20th- and 21st-century sources, spelling was recorded as for those editions.

4.2 Transcription layer

Texts in historical corpora tend to undergo a form of normalisation. It usually consists of modifying the original text to make its reception easier for modern audiences and more accessible for automatic text processing tools. The degree of intervention varies greatly—from standardisation of spelling to applying modern inflection or even lexis.[27] The decision on the extent of normalisation depends on the specific conditions of the given language and the goals of the authors of the corpus.

In KorBa, the general principle was to subject only spelling to normalisation (hereafter referred to as 'transcription'), with historical inflectional endings or lexis unchanged. This decision is fundamental for further automatic text processing, as it requires the extant tools to be adjusted for the state of Polish inflection in the 17th and 18th centuries.

The transliterated texts were subjected to automatic transcription (see Sect. 5.1). The main goal of this process was to conflate various spellings of a given wordform. This made automatic morphosyntactic annotation easier and more consistent during corpus creation. It also allows the user to search for specific forms without having to account for spelling variants. In principle, it has been decided that the transcribed text should, in the spelling layer, be as similar to a modern Polish text as possible. Therefore, the starting point was the current letter set (32 characters, including 9 with diacritics). The use of diacritics has been altered in line with the modern orthography (e.g. *gora* → *góra* 'mountain', *ćicho* → *cicho* 'silently', *rzeczy* → *rzeczy* 'things'). In particular, the letters *á* and *é*, which are no longer in use, were changed into their modern forms (e.g. *álbo* → *albo* 'or', *téj* → *tej* 'that'). Letters *q*, *x*, and *v*, not used in the Polish alphabet, were replaced with their phonetic equivalents of *k*, *ks*, *u*, or *w* (e.g. *reliquie* → *relikwie* 'relics', *taxa* → *taksa* 'pay rate', *vbić* → *ubić* 'slaughter,' *vino* → *wino* 'wine'). Letters *y* and *i*, where used in ways incongruous with modern norm, were replaced with *i* or *j*, e.g. *y* → *i* 'and', *iedna* → *jedna* 'one', *mieysce* → *miejsce* 'place'. Any words written down as pronounced (but not as mandated by the modern spelling standard) were updated to their modern form, e.g. *poniewasz* → *ponieważ* 'since'. Spelling different from the modern one was kept only in cases where dialect influences were suspected (*zwirz*, modern standard Polish *zwierz* 'beast') and in some special cases (e.g. the letter *q* was kept in the currently defunct word *Tlaquaciow* 'exotic species of animal' as it

[27] Such a far-reaching standardization has been carried out in the Slovenian corpus (cf. Erjavec, 2015: 13–14). However, the standardised layer was used only for automatic annotation, whereas the users are only provided with the transliteration layer.

was impossible to determine how that form would function under modern spelling rules).

4.3 Document structure and language markup

The processing of transliterated texts includes marking up the structure of the source document, identifying foreign-language fragments, and morphosyntactic annotation of every token (for the description of the morphosyntactic layer, see Sect. 4.4).

Thanks to a structural markup, the user gains information about such elements as the identifier of a page that a given token is on. This allows for a precise location of the searched expressions in the source, facilitating the use of quotes from the corpus in academic and lexicographic work. One can also relatively easily find the relevant fragment in the original copy.

Other elements of the document structure are also marked, providing the user with more complete knowledge of the context for the queried expressions. The marked elements include:

- Fragments not being part of the original text, i.e. general editor's additions from later editions (19th–21st-century), as well as commentaries introduced by transliterators, such as any signs of doubt regarding the form of the word;
- Passages omitted in transliteration, such as extended foreign-language passages, mathematical equations, etc.;
- Additional tags allowing one to place the fragment in the broader structure of the text, such as tags of the title page and its elements (e.g. the name of the printing house), tags for fragments before the main texts (e.g. dedications), tags for appendices of the main text (e.g. marginal notes), etc.

Language markup consists of assigning information on the specific foreign language used for every non-Polish token. This was necessary, first and foremost, due to the large amount of Latin inserts in 17th- and 18th-century Polish texts. Aside from Latin, the following languages are represented in the corpus: Arabic, Czech, French, German, Greek, Hebrew, Hungarian, Italian, Lithuanian, and Spanish. In a few cases, entire language (sub)families were marked with the same tag. These include the Scandinavian family, Turkic-Tatar languages, Southern Slavic languages, and East Slavic languages. This solution was applied to those languages that were still at the early stages of their development in the 17th and 18th centuries and would thus be difficult to distinguish from others in the same language family.

4.4 Morphological layer (tagset)

The distinguishing trait of modern corpora is the detailed linguistic annotation of all tokens in the text, consisting of their basic forms (lemmata) and linguistic categories assigned to them. In inflectional languages such as Polish, each token is not only POS-tagged, but also characterized by a set of tags specifying the values of its grammatical categories. These categories include inflectional categories (such as a gender of an adjective) and categories which are not inflectional for a given lexeme,

but have some syntactic functions (e.g. a gender of a noun, a case of a preposition). That is why we use the term 'morphosyntactic annotation'.

KorBa, much like NKJP, bases its POS classification on the idea of 'flexeme'—a term narrower than 'lexeme' (Bień & Saloni, 1982). While traditionally defined lexemes may include forms assigned to diverse grammatical categories,[28] flexemes consist only of forms that can be characterised through the use of the same grammatical categories. The flexeme sets noted in NKJP and KorBa, despite being similar, are not identical: firstly, the KorBa tagset includes flexemes which existed in 17th- and 18th-century Polish and are now either completely gone or have only survived in a relict form; secondly, some functions of individual units within the linguistic system were reflected more precisely than in NKJP.

A good illustration of the former case is the 'adjective in non-complex inflection' flexeme of the KorBa, which includes the so-called short forms of the adjective, today surviving only in masculine nominative singular of a handful of adjectives (e.g. *zdrów* 'healthy', *gotów* 'ready') and some ossified expressions (e.g. *z bliska* 'up close', *po polsku* 'in Polish'), but used far more broadly in the 17th and 18th centuries. An example of a more detailed description would be splitting off the future forms of the verb BYĆ 'to be' as markers of the future tense in compound constructions with an infinitive or the l-participle (respectively *będ-ę* 'be-1SG.FUT' *czyta-ć* 'read-INF' or *będ-ę* 'be-1SG.FUT' *czyta-ł* 'read-M' ('I will read')) into a separate flexeme. This style of annotation makes it easier to search the corpus for future forms of verbs, which may be useful, for example, in lexicography.

Further differences of a similar nature between NKJP and KorBa can be seen in grammatical categories. On the one hand, the value sets of some categories were expanded with the ones that existed in the Middle Polish period, such as the dual value ('du') in the number category. On the other hand, the repertoire of grammatical categories and their values was changed. For instance, a new value for the aspect category was added—the biaspectual ('biasp'). It was assigned to verbs which can be perfective or imperfective depending on the context (e.g. ABDYKOWAĆ 'to abdicate') and ones where aspect is impossible to determine due to lack of diagnostic forms in the corpus.

The last example also shows the most extensive change within the set of grammatical categories in comparison to the NKJP, i.e. the introduction of tags that allow for reporting ambiguous tokens or 17th–18th-century forms unknown to modern users. The best illustration of such tokens and the procedures of tagging adopted to reflect their ambiguity is the differentiation within the masculine gender. In general terms, KorBa operates on the principle of assigning gender to wordforms and determining it with the degree of precision afforded by the context in which a given form is found. Thus, we assign the so-called generalized masculine value ('m') to most masculine forms where there is no variation in the endings (e.g. nom. sing.). Two other values, 'masculine animate 1' ('manim1') and 'masculine animate

[28] For example, we can distinguish within a verb lexeme, among others, the finite forms of the verb (assigned to the categories of number, person, etc.), the infinitive form (which cannot be characterised through any of the traditional categories), and, in some frameworks, the gerund (with its case and number categories).

2' ('manim2'),[29] are assigned only to these forms where the endings allow to distinguish either 'animate 1' or 'animate 2' from generalized masculine gender. Therefore, the form *tygrys-owie* 'tiger-NOM.PL.PERS' shall be characterised as 'manim1', while the form *tygrys-y* 'tiger-NOM.PL.ANIM'– as 'm'.

The full list of grammatical classes and categories alongside the values assigned to them can be found in the corpus user manual (Gruszczyński & Bronikowska, 2018), available at the corpus website https://korba.edu.pl (item "Instruction").

5 Stages in the compilation of the corpus and tools

Morphosyntactic annotation of the corpus was performed through a combination of tools. The transliterated text of the original was subjected to transcription (see Sect. 5.1). Subsequently, a morphological analyser was applied, interpreting the possible inflectional forms of every transcribed token (see Sect. 5.2). The contextual selection of a single interpretation for a given token was done by a tagger. As is usual in corpus development, a part of it was disambiguated and verified manually (see Sect. 5.3). That subcorpus demonstrates the intended 'ideal' tagging (excluding errors by human annotators) and also serves to train the automatic tools (see Sect. 5.4).

5.1 Transcription

The preliminary stage of processing the texts consisted of their transliteration, as well as marking up their structure (including identification of fragments in foreign languages). The texts prepared in this way were subjected to transcription (standardisation). For this, it was decided to use an existing tool developed for the transcription of Polish historical texts within the IMPACT project (Bień, 2014). The tool uses a set of rewrite rules based on regular expressions. It was decided to maintain two separate sets of rules—one for original editions and the other for 19th-century ones. Both sets of rules were extended while annotating the manual subcorpus, on the basis of the feedback given by the annotators. Unfortunately, each of them increased to over 3000 rules and became hard to maintain. Despite its simplicity the tool proved to be useful both as a support for human annotators during the creation of gold standard data, as well as for automatic transcription of the full corpus.

5.2 Morphological analysis

The automatic inflectional analysis of various forms of lexemes present in the corpus texts was performed by a morphological analyser named Korbeusz. It is a

[29] Nowadays, we would use the terms 'masculine personal' and 'masculine animate', but in the 17th and 18th centuries the category of personality was not stabilized yet, therefore we have decided to change this denomination.

modified version of a tool named Morfeusz 2 (Woliński, 2014) developed for analysing forms functioning in Modern Polish.

Morfeusz requires a list of inflectional forms—words and their interpretations. The basic source of such data for modern Polish is the *Grammatical Dictionary of Polish* (*Słownik gramatyczny języka polskiego*, hereafter SGJP; Saloni et al. 2015), but historical Polish requires an additional list of forms that do not appear in modern texts. The source of such data could be another dictionary or an effective procedure for modifying ('ageing') the SGJP data. Both methods were used in the creation of Korbeusz, although a great majority of the data was produced through the latter, in part because the core of the SGJP data consists of the entries from the *Dictionary of Polish* (*Słownik języka polskiego*) edited by W. Doroszewski (SJPDor; Doroszewski 1950–1969), which includes lexical material going back to the last quarter of the 18th century and, therefore, it contains a large amount of old or obsolete vocabulary that was still in general use in the 17th and 18th centuries.

5.2.1 Modified SGJP data

SGJP data was first adapted for the tagset of KorBa by assigning tags consistent with the KorBa tagset to inflectional forms generated through the SGJP model. For example, this involved modifying the gender system in line with the one adopted by KorBa (see Sect. 4.4).

Moreover, some forms in the SGJP were used to generate certain historical regular inflectional forms, e.g. the first and second person imperative duals of a verb (e.g. *pisz-wa* 'write-1DU.IMP', *pisz-ta* 'write-2DU.IMP') were created by adding the *-wa* and *-ta* endings to the second person singular imperative (*pisz* 'write[2SG.IMP]'). The historical forms were created without exceptions and for entire lexeme classes. This means that in many cases they are surplus, for example because they represent dual forms of verbs that did not exist in the 17th and 18th centuries or were extremely rare. This is not a problem from the point of view of inflectional analysis, since none of the above-mentioned forms are systematically homonymous to others, and, therefore, the surplus forms will not result in an incorrect analysis of other lexemes.

It was also necessary to remove some forms from the SGJP dataset that could not appear in 17th- or 18th-century texts and whose presence in this dataset could lead to an erroneous interpretation of other words, such as the vocabulary describing elements of modern reality.

5.2.2 Inflectional data from e-SXVII and its expansion

The second supplementary inflectional data source fed into Korbeusz is the inflectional information from the *Electronic Dictionary of 17th- and 18th-century Polish* (Gruszczyński[30], hereafter e-SXVII), which is currently under development. It has to be emphasised that the dictionary notes only the forms attested in the

[30] Gruszczyński, W. (ed.). *Electronic Dictionary of 17th- and 18th-century Polish* (*Elektroniczny słownik języka polskiego XVII i XVIII wieku*). https://sxvii.pl/

dictionary's canon texts and so the inflectional paradigms in e-SXVII are almost always incomplete. Consequently, the inflectional data of e-SXVII consists of only around 84 thousand inflectional forms in a dictionary of 44 thousand entries. This data was converted to KorBa's tagset and added to Korbeusz's data.

During the conversion, the dataset was augmented with some homonymous or regularly derived forms that had been previously unrecognised by the e-SXVII. These include, for example, the dative and locative singular forms of feminine nouns ending in *-a* created from homonymous nominative and accusative dual forms (e.g. *żabi-e* 'frog-NOM.DU'), and the superlative forms of adjectives created from the comparative forms by the addition of prefixes *naj-* and *na-* (*ładni-ejszy* 'pretty-CMPR' → *naj-ładni-ejszy*, *na-ładni-ejszy* 'SUPL-pretty-SUPL'). Eventually, e-SXVII data produced a total of almost 100 thousand inflectional forms for the Korbeusz dataset.

That data was then subjected to automatic partial reconstruction of the most productive inflectional paradigms (Kieraś et al., 2017). As a result of this procedure, other 160 thousand forms were generated, a large majority of them being correct, but occasionally only postulated. This data was also added to Korbeusz's set of inflectional forms.

5.2.3 Segmentation rules

Aside from the set of inflectional forms, the inflectional analyser also requires a set of segmentation rules. They allow for the analysis of words that consist of more than one token. The Korbeusz segmentation ruleset has been notably modified in comparison to the ruleset for modern Polish analyser so as to include non-standard (incorrect under modern language norm) spelling. For example, in the 17th and 18th centuries, the particle *nie* 'no' could be spelled together with verb forms (e.g. *nie-frasowa-ć* 'NEG-worry-INF' ('not to worry')).[31]

5.3 Manually annotated subcorpus

Morphosyntactic interpretations produced by an automated morphological analyser were disambiguated, verified, and completed by a human annotator. This procedure was performed on a part of the corpus of ca. 500 thousand tokens. The work was performed in the Anotatornia 2 system developed within the Chronofleks project (Woliński et al., 2017).[32] The system takes into consideration the particular challenges involved in annotating a historical corpus, including the existence of transliterated and transcribed parallel versions of the text and the necessity of preserving information about their original pagination.

Anotatornia 2 functions as a web application, allowing a group of annotators to work over the parts of the corpus assigned to them. It is assumed that the utility is

[31] Some segmentation problems resulting from the differences between Middle Polish and the modern language are discussed in: Gruszczyński et al., 2020, p. 43.

[32] Anotatornia 2 is an open source software. The code may be found at: http://zil.ipipan.waw.pl/Anotatornia2/.

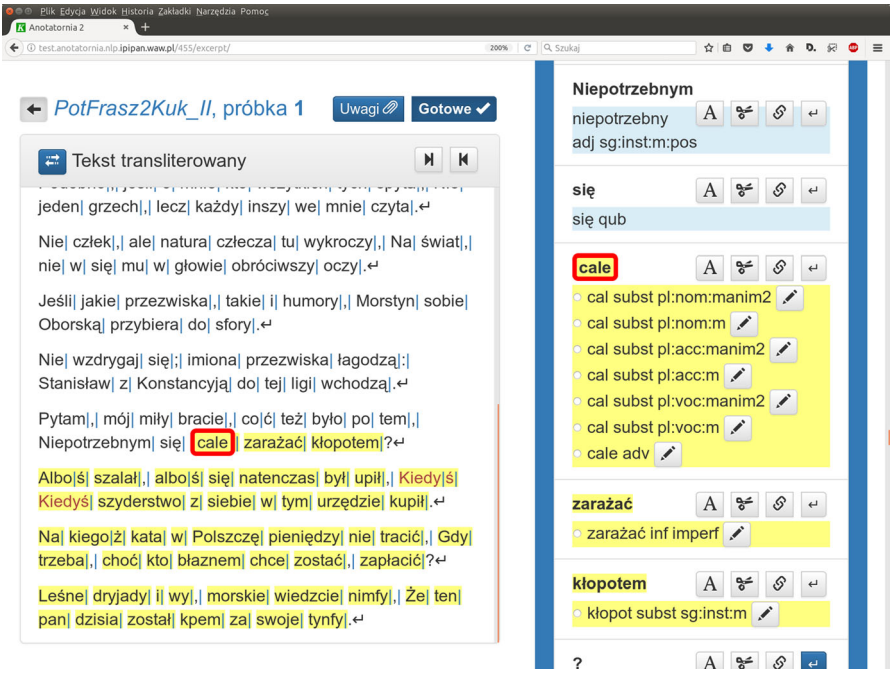

Fig. 3 Annotator interface in Anotatornia 2

fed a text that has been initially processed by an inflectional analyser with the appropriate dictionary. The system users' tasks include: verification and completion of inflectional tags supplied by the analyser; disambiguation of analyses; correction of transcription; and correction of sentence divisions. The annotator interface is shown in Fig. 3. The left part of the screen displays the corpus sample; tokens that still require the annotator's attention are highlighted. The most important job of the annotator—disambiguating inflectional interpretations—consists of selecting one interpretation from the list displayed on the right of the screen. The buttons allow them to modify the transcription, to change token and sentence boundaries, and introduce an interpretation that was not anticipated by the automatic analyser.

The work proceeded in an analogous manner to the tagging of the NKJP: every sample was processed independently by two annotators whose answers were then compared automatically. Any case of divergence was flagged and the sample was shown to the users once more, asking them to verify their answers. Any conflicts that remained would be decided by an adjudicator specialised in maintaining a coherent tagging of the corpus. This process ensures a high quality of the tagging, but it is labour-intensive. Eventually, it became necessary to have a part of the corpus processed by a single annotator. Table 3 compares the sizes of the parts of the corpus annotated in both ways (excluding tokens representing punctuation, foreign language inserts, and other elements not subject to inflectional interpretation). It shows that the frequency of corrections introduced by annotators were similar in both cases (slightly lower in the part tagged by only one person):

Table 3 Manually annotated subcorpora

	Corpus tagged by two annotators	Corpus tagged by one annotator
Number of tokens	355,725	141,939
Transcription corrections	2.42%	2.26%
Segmentation corrections	1.30%	1.31%
Added inflectional interpretation	7.03%	6.01%

Table 4 Annotator agreement

	As a percentage of all tokens
Inter-annotator agreement	91.25%
Cohen's Kappa for parts of speech	0.962
Cohen's Kappa for full tags	0.927
Tokens with conflict	8.75%
One annotator's answer was chosen	7.46%
Adjudicator's own answer	1.30%
Adjudicator's interventions without conflict	0.33%

Table 4 presents information about agreements and conflicts in the part of the corpus annotated by two annotators and corrected by an adjudicator. The two annotators agreed in 91.25% of cases, which may be considered a very high percentage for texts of such difficulty. The values of Cohen's $\varkappa$ in Table 4 were computed only for the tags since the assignment of lemmas and of transcriptions is not a choice from a closed set of labels. It is probably interesting to note that the tagset consists of about two thousand distinct tags, so probability of an agreement by chance is anyway very low in this task.

Conflicts between annotators' decisions appeared in 8.75% of tokens. The adjudicator approved a solution proposed by one of the annotators for 7.46% of tokens (i.e. 85% of all conflicts) and declared both proposals incorrect in the remaining 15% of differences. The adjudicator was requested to check only tokens with conflicts, nonetheless, in 0.33% of tokens the adjudicator changed the answer of the annotators even though they both agreed on it. We may assume that these changes were triggered by conflicts in some neighbouring tokens.

5.4 Morphosyntactic tagging

Syncretism and homonymy are typical for both historical and contemporary Polish, as well as many other fusional languages. However, in KorBa a less typical problem of ambiguous segmentation arises and needs to be addressed. It is marginal in contemporary Polish, but becomes a significant problem in the 17th- and 18th-century language.

Table 5 Accuracy of taggers in tenfold cross-validation

	Korba dataset (%)	NKJP dataset (%)
Concraft	88.8	92.4
Toygger	92.2	95.3

Consider for example the verb lexeme DAĆ 'to give', with a future tense third person singular form *da*. Attaching common particles *-ć* (emphatic particle) or *-li* (question marker) to this form results in constructions *da·ć* 'give-3SG.FUT'· 'EMPH' and *da·li* 'give-3SG.FUT'· 'Q', homonymous with actual inflectional forms of the DAĆ paradigm: *da-ć* 'give-INF' and *da-l-i* 'give-PST-3PL'. Thus, each of the words *dać* and *dali* can be interpreted either as one token (*dać*, *dali*) or as two consecutive tokens (*da·ć*, *da·li*). This homonymy is accidental and can be disambiguated only in context, but it applies to a long series of verbal forms and causes systematic ambiguity. The same applies to historical masculine or neuter instrumental adjectival forms such as *różn-em* 'different-INS.SG' (today only *różnym*), which are systematically homonymous with the alternative segmentation *różne·m* 'different-NOM.SG'· 'be-1SG.PRS' ('I am different'), where the form *różne*, a nominative or accusative form of the same lexeme, appears together with the *-m* suffix functioning as an agglutinative form of BYĆ 'to be'. This group of homonyms is even larger than the former one.

Two stochastic taggers were used to automatically annotate the KorBa corpus data. Both were trained on the manually annotated subcorpus described above. The first was Concraft 2 (Waszczuk, Kieraś & Woliński, 2018), a tagger based on conditional random fields which was specifically adapted to cope with the problem of ambiguous segmentation. Concraft builds three separate statistical models aimed at the division into sentences, disambiguation of ambiguous segmentation and ambiguous morphosyntactic tags, and attempts at guessing morphosyntactic tags for unknown tokens. The other tagger, Toygger (Krasnowska-Kieraś, 2017), based on Bi-LSTM neural networks, performs only the latter task, i.e. morphosyntactic disambiguation; for that reason, it uses data previously segmented and disambiguated on the segment level by Concraft, but assigns its own morphosyntactic tags to it. Additionally, both taggers guess, i.e. assign statistically likely tags to tokens unknown to the morphological analyser. Text segmentation in both annotations is fully aligned, the taggers only assign morphological tags based on their own statistical models.

It was expected that both taggers would achieve lower benchmark results than in the case of contemporary Polish. The 17th- and 18th-century corpus covers a much larger timespan than modern NKJP, the language is much more diverse and less standardised than nowadays. Furthermore, the 17th- and 18th-century training dataset is significantly smaller than the manually annotated subcorpus of NKJP (ca. 1.2 million tokens), which obviously impairs the taggers' statistical models. Table 5 presents the results of tenfold cross validation of the taggers on Korba and on contemporary data of NKJP. The measure used is accuracy counted per token. The

```
<seg xml:id="morph_2.174-seg" corresp="ann_segmentation.xml#segm_2.174-seg">
  <fs type="morph">
    <f name="orth">
          <string>Książęcia</string>
    </f>
    <f name="translit">
          <string>Xiążęćia</string>
    </f>
    <f name="interps">
        <fs xml:id="morph_2.174.1-lex" type="lex">
            <f name="base">
              <string>książę</string>
            </f>
            <f name="ctag">
              <symbol value="subst"/>
            </f>
            <f name="msd">
            <vAlt>
                <symbol xml:id="morph_2.174.1.1-msd" value="sg:gen:m"/>
                <symbol xml:id="morph_2.174.1.2-msd" value="sg:gen:n"/>
                <symbol xml:id="morph_2.174.1.3-msd" value="sg:acc:m"/>
                <symbol xml:id="morph_2.174.1.4-msd" value="sg:acc:n"/>
            </vAlt>
            </f>
        </fs>
    </f>
    <f name="disamb">
        <fs type="tool_report">
            <f name="choice" fVal="#morph_2.174.1.2-msd"/>
            <f name="interpretation">
              <string>książę:subst:sg:gen:n</string>
            </f>
        </fs>
    </f>
  </fs>
</seg>
```

Fig. 4 Fragment of the morphosyntactic layer of KorBa in XML encoding

tagging results for historical dataset can be considered moderately good, as the morphological disambiguation accuracy of each tagger is about 4 pp. lower than in the case of NKJP dataset. Despite Concraft's noticeably worse tagging accuracy, it was decided that both morphological annotations will be available in KorBa as separate layers accessible from the corpus query language. The users can decide which annotation they deem more reliable in their research and can even require concordance (or divergence) between the taggers both on POS and on specific values of grammatical categories. Such a constraint should increase the precision of the query, but may impair its recall. In some research, however, this could be a useful feature.

5.5 XML encoding of the corpus

One of the design goals of KorBa was to remain as compatible with the contemporary National Corpus of Polish as possible. For that reason, KorBa uses the XML encoding designed for NKJP with minor changes. This encoding is an instance of TEI P5 guidelines using a stand-off annotation (Przepiórkowski et al., 2012). KorBa includes three of NKJP's layers of annotation: the text structure (keeping the text in the transliterated form and structural tags), segmentation layer

(describing division of the text into tokens), and morphosyntax layer (providing morphosyntactic interpretation for each token). Unlike in the Slovene corpus, the transcription is treated as part of annotation of transliterated tokens (and not a variant of the text, cf. Erjavec, 2015, p. 765). Thus, the transcription belongs to the morphosyntactic layer of the corpus (ann_morphosyntax.xml). A fragment of such a file describing a single token *Xiążęćia* 'prince' is shown in Fig. 4. The transliterated/original form of the token is available as the value of the feature 'translit' belonging to the feature structure 'morph' describing the token. The transcribed/modernized form is available as the value of feature 'orth' (as in NKJP). The rest of the structure shown follows exactly the NKJP pattern: all possible morphosyntactic interpretations given by the morphological analyser are included and one of them is marked as correct for the context with the 'disamb' feature.

6 An example of a corpus search

As a practical example, take tracing an inflectional phenomenon through the corpus. Specifically, we shall focus on the plural locative noun ending *-ech*, as opposed to *-ach*, which used to be a feminine suffix, but later spread to all genders. The *-ech* ending is present in other Western Slavic languages, such as Czech. While in modern Polish the *-ech* suffix survives only in three proper names—Węgry 'Hungary'; Włochy 'Italy'; and Niemcy 'Germany' (*Węgrz-ech* 'Hungary-LOC', *Włosz-ech* 'Italy-LOC', and *Niemczech* 'Germany-LOC', respectively)—in historical Polish, it appeared in a much larger group of lexemes, both common and proper nouns. It is possible to trace the regression of those forms in our corpus.

A corpus query returning all the instances of the phenomenon needs to restrict the search to a single token based on its modernised form (orth = ".*ech"), belonging to a particular part of speech (pos = "subst"), and having grammatical features of number and case (number = "pl" & case = "loc"). These can be shortened into a query based on the complete form of the tag: tag = "subst:pl:loc:.*". Such a concatenated query would yield 3139 results. To minimize the number of false positives from automatic tagging errors, the user may add another term, selecting only the matches where both taggers agreed on the morphosyntactic tag (tag_c = "subst:pl:loc:.*").

The results can be exported to a CSV or XLS file for further processing in spreadsheets or scripting languages such as Python or the R programming language for statistical analysis. The exported file contains not only the returned token (or tokens), left and right contexts, and morphological tags from both taggers, but also the complete metadata for each match. Figure 5 presents a plot based on the query results described above. The matches were grouped by decade. The plot presents a number of unique nouns that were used at least once with the *-ech* ending in plural locative form in the given decade. It demonstrates a clear and constant regression of the *-ech* suffix from nearly one hundred lexemes at the beginning of the 17th century to less than 20 in the 18th century.

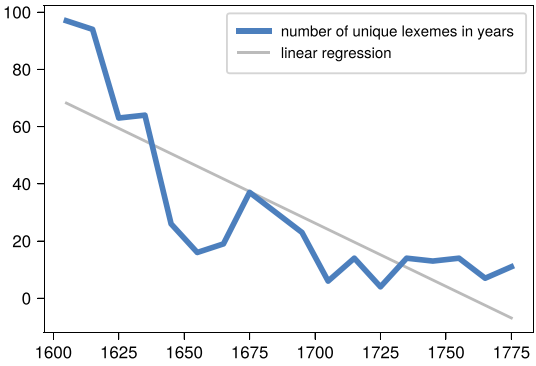

Fig. 5 A plot illustrating the regression of the historical plural locative suffix *-ech* in nouns based on data provided by a query of the corpus

7 Conclusions

The electronic corpus of 17th- and 18th-century Polish texts in the form presented in this article was made available in 2018. It was and is the first such a large corpus of historical Polish featuring morphosyntactic annotation. This work shows that a very detailed annotation schema of the National Corpus of Polish can be successfully adapted to historical Polish. We hope that the corpus will allow language historians to verify the knowledge of the Polish language of the 17th and 18th centuries by providing a much broader material than that previously available. It is important for this kind of research that the corpus can be searched by means of a CQL-based search tool.

An important feature of the corpus is that each token has its dual representation—transliterated and transcribed ones. The former allows the users to study old wordforms, while the latter makes searching the corpus easier.

As for NLP resources, our contributions include a publicly available manually annotated subcorpus of half a million wordforms which can be used to train various NLP tools and a comprehensive morphological dictionary as well as a tagger adapted to our annotation schema.

Since 2019, the work on the corpus has continued as part of a new project. The corpus is being expanded both through increasing its volume of texts for the previously used time frame (1601–1772) and through extending its chronological coverage into the years 1773–1800. The corpus is planned to contain 25 million tokens in total. New tools are also under development: a transcriber and a tagger. The new version of the corpus will be transcribed using a machine learning approach trained on the manually verified transcription layer of the above-mentioned gold standard subcorpus. Initial experiments also show that using BERT-based neural networks is possible to improve the tagging accuracy for Korba.

The project also includes plans for the integration of various Polish linguistic resources for the 17th and 18th century (Ogrodniczuk & Gruszczyński, 2019). These include, aside from the electronic corpus of 17th- and 18th-century Polish, the following: the *Electronic Dictionary of 17th- and 18th-century Polish*, the paper records of that dictionary, and the Digital Library of Polish and Poland-related

Ephemeral Prints from the 16th, 17th and 18th Centuries (Cyfrowa Biblioteka Druków Ulotnych Polskich i Polski Dotyczących z XVI, XVII i XVIII Wieku[33]).

Acknowledgements We would like to thank the anonymous reviewers for their careful reading of our manuscript and their suggestions and comments.

Funding This work was supported by the National Programme for the Development of Humanities (NPRH) [0036/NPRH2/H11/81/2012, 0413/NPRH7/H11/86/2018].

Data availability https://www.korba.edu.pl

Code availability http://chronofleks.nlp.ipipan.waw.pl/, http://zil.ipipan.waw.pl/Anotatornia2/

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Bień, J. S. (2014). The IMPACT project Polish Ground-Truth texts as a DjVu corpus. *Cognitive Studies|Études Cognitives*, (14), 75–84. https://ispan.waw.pl/journals/index.php/cs-ec/article/view/cs.2014.008/174. Accessed 19 March 2021.

Bień, J. S. & Saloni, Z. (1982). Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna). *Prace Filologiczne*, XXXI, 31–45. http://bc.klf.uw.edu.pl/63/. Accessed 19 March 2021.

Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey: European Language Resources Association (ELRA), pp. 474–478. http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf. Accessed 19 March 2021.

Brouwer, M., Brugman, H., & Kemps-Snijders M. (2017). MTAS: A Solr/Lucene based multi tier annotation search solution. In Borin, L. (Ed.), *Selected papers from the CLARIN Annual Conference 2016 (Aix-en-Provence, 26–28 October 2016). Linköping Electronic Conference Proceedings 136*, 19–37. http://www.ep.liu.se/ecp/136/002/ecp17136002.pdf. Accessed 19 March 2021.

Deptuchowa, E., Jasińska, K., Klapper, M. & Kołodziej, D. (2020). O projekcie Korpusu Polszczyzny do 1500 roku, *Poradnik Językowy*, 8, 7–16. https://doi.org/10.33896/PorJ.2020.8.1. Accessed 19 March 2021.

Derwojedowa, M. (2020). Mikrokorpus Gronowy Polszczyzny 1830–1918, *Poradnik Językowy*, 8, 52–65. https://doi.org/10.33896/PorJ.2020.8.4. Accessed 19 March 2021.

[33] https://cbdu.ijp.pan.pl/

Długosz-Kurczabowa, K. & Dubisz, S. (2001). *Gramatyka historyczna języka polskiego*. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.

Dobrushina, Y. R., Kravetskiy, A. G., & Polyakov, P. Y. (2015) – Добрушина, Е. Р., Кравецкий, А. Г. & Поляков, А. Е. (2015). Корпус и частотный грамматический корпусный словарь церковнославянского языка в составе НКРЯ. In *Труды Института русского языка им. В. В. Виноградова,* 6, 116–141.

Doroszewski, W. (Ed.) (1950–1969). *Słownik języka polskiego* (Vol. 1–11). http://doroszewski.pwn.pl/. Accessed 19 March 2021.

Erjavec, T. (2015). The IMP Historical Slovene Language Resources. In *Language Resources and Evaluation*, 49(3), 753–75. https://doi.org/10.1007/s10579-015-9294-7. Accessed 19 March 2021.

Florczak, Z. (1967). *Udział regionów w kształtowaniu się piśmiennictwa polskiego XVI wieku*. Wrocław-Warszawa-Kraków: Zakład Narodowy imienia Ossolińskich Wydawnictwo Polskiej Akademii Nauk.

Garabík, R. & Kajanová, M. (2015). Digitalizácia a anotácia Prameňov k dejinám slovenčiny. In Balleková, K., Múcsková, G. & Králik, Ľ. (Eds.), *Prirodzený vývin jazyka a jazykové kontakty*. Bratislava: Veda, pp. 577–583.

Górski, K., Kuraszkiewicz, W., Pepłowski, F., Saski, S., Taszycki, W., Urbańczyk, S., Wierczyński, S. & Woronczak J. (1955). *Zasady wydawania tekstów staropolskich. Projekt*. Wrocław: Zakład im. Ossolińskich, Wydawnictwo PAN.

Gruszczyński, W., Adamiec, D., Bronikowska, R. & Wieczorek, A. (2020). Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. – problemy teoretyczne i warsztatowe, *Poradnik Językowy*, 8, 32–51. https://doi.org/10.33896/PorJ.2020.8.3. Accessed 19 March 2021.

Gruszczyński, W., & Bronikowska, R. (2018). Electronic corpus of 17th- and 18th-century Polish texts search engine user manual. https://korba.edu.pl/. Accessed 19 March 2021.

Hernas, Cz. (2002). *Barok*. Warszawa: Wydawnictwo Naukowe PWN.

Jínová P., Lehečka B. & Oliva K. (2014). Describing Old Czech Declension Patterns for Automatic Text Analysis. *Mundo Eslavo*, 13, 7–17. http://mundoeslavo.com/index.php/meslav/article/view/161/144. Accessed 24 Oct 2019.

Kieraś, W., Komosińska, D., Modrzejewski, E. & Woliński, M. (2017). Morphosyntactic Annotation of Historical Texts. The Making of the Baroque Corpus of Polish. In Ekštein, K., Matoušek, V. (Eds.), *Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science*, vol. 10415. Springer, Cham, pp. 308–316. https://doi.org/10.1007/978-3-319-64206-2_35. Accessed 19 March 2021.

Kieraś, W., & Woliński, M. (2018). Manually annotated corpus of Polish texts published between 1830 and 1918. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahar, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France: European Language Resources Association (ELRA), pp. 3854–3859. http://www.lrec-conf.org/proceedings/lrec2018/pdf/675.pdf. Accessed 19 March 2021.

Krasnowska-Kieraś, K. (2017). Morphosyntactic disambiguation for Polish with bi-LSTM neural networks. In Z. Vetulani & P. Paroubek (Eds.), *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland, pp. 367–371.

Kroch, A., Santorini, B., & Delfs, L. (2004). The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3. https://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-3/index.html. Accessed 19 March 2021.

Król, M., Derwojedowa, M., Górski, R. L., Gruszczyński, W., Opaliński, K. W., Potoniec, P., Woliński, M., Kieraś, W., & Eder, M. (2019). Narodowy Korpus Diachroniczny Polszczyzny. *Projekt, Język Polski,* XCIX Is. 1, 92–101.

Kučera, K., Řehořková, A., & Stluka, M. (2015). DIAKORP: Diachronní korpus, verze 6 z 18. 12. 2015. Ústav Českého národního korpusu FF UK, Praha.

Kučera, K., & Stluka, M. (2011). DIAKORP: Diachronní korpus, verze 5 z 21. 2. 2011. Ústav Českého národního korpusu FF UK, Praha.

Kytö, M. (2011). Corpora and historical linguistics. I*n Revista Brasileira de Linguistica Aplicada*, 11(2): 417–457. https://doi.org/10.1590/S1984-63982011000200007. Accessed 19 March 2021.

Mishina, Y. A., & Pichkhadze, A. A. (2015) – Мишина, Е. А., Пичхадзе, А. А. (2015). Древнерусский подкорпус Национального корпуса русского языка. In *Труды Института русского языка*

им. В. В. Виноградова. VI. *Национальный корпус русского языка: 10 лет проекту*. Москва, pp. 99–115.

Nelson, M. (2010). Building a written corpus: what are the basis? In O'Keeffe, A. and McCarthy, M. (Eds.), *The Routledge Handbook of Corpus Linguistics*. London-New York, pp. 53–65.

Ogrodniczuk, M., & Gruszczyński, W. (2019). Connecting Data for Digital Libraries: The Library, the Dictionary and the Corpus. In Jatowt, A., Maeda, A. and Syn, S. (Eds.), *Digital Libraries at the Crossroads of Digital Information for the Future. ICADL 2019. Lecture Notes in Computer Science*, vol. 11853. Springer, Cham, pp. 125–138. https://doi.org/10.1007/978-3-030-34058-2_13. Accessed 19 March 2021.

Opaliński, K., & Potoniec, P. (2020). Korpus Polszczyzny XVI wieku, *Poradnik Językowy*, 8, 17–31. https://doi.org/10.33896/PorJ.2020.8.2. Accessed 19 March 2021.

Przepiórkowski, A., Bańko, M., Górski, R. L., & Lewandowska-Tomaszczyk, B. (Eds.) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN. http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf. Accessed 19 March 2021.

Rayson, P., Archer, D., Baron, A., Culpeper, J. & Smith, N. (2007). Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English Corpora. In *Corpus Linguistics Conference (CL2007)*, University of Birmingham, Birmingham, UK, pp. 1–14. http://ucrel.lancs.ac.uk/publications/CL2007/paper/192_Paper.pdf. Accessed 19 March 2021.

Saloni, Z., Woliński, M., Wołosz, R., Gruszczyński, W., & Skowrońska, D. (2015). *Słownik gramatyczny języka polskiego* (3rd ed.) online. Warszawa. http://sgjp.pl. Accessed 19 Mar 2021.

Scheible, S., Whitt, R. J., Durrell, M. & Bennett, P. (2011). A Gold Standard Corpus of Early Modern German. In *Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics*, Portland, Oregon, USA, pp. 124–128. http://www.aclweb.org/anthology/W11-0415. Accessed 19 March 2021.

Sichinava, D. V. (2016) – Сичинава, Д. В. (2016). Старорусские/среднерусские тексты в НКРЯ: от экстенсивной коллекции к корпусу. In Baranovas, V., Timčenko, T. (Eds.), *El'Manuscript–2016. Rašytinis palikimas ir skaitmeninės technologijos: VI tarptautinė mokslinė konferencija*, Vilnius, 2016 m. rugpjūčio 22–28 d. Pranešimai. Lietuvos mokslo taryba, Vilnius, Izhevsk, pp. 208–210.

Waszczuk, J., Kieraś, W. & Woliński, M. (2018). Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields. In P. Sojka, A. Horák, I. Kopeček, K Pala (Eds.), *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11–14, 2018, Proceedings, number 11107 in Lecture Notes in Artificial Intelligence*, Springer-Verlag, pp. 188–196.

Woliński, M. (2014). Morfeusz reloaded. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavík, Iceland: European Language Resources Association (ELRA), pp. 1106–1111. http://www.lrec-conf.org/proceedings/lrec2014/pdf/768_Paper.pdf. Accessed 19 March 2021.

Woliński, M., Kieraś, W., & Komosińska, D. (2017). Anotatornia 2—An Annotation Tool Geared towards Historical Corpora. In Z. Vetulani, & P. Paroubek (Eds.), *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp. 158–162.