Diachronic Corpus of Bulgarian

Corpus Compilation

Initially, 54 texts were extracted from the Bulgarian National Corpus, amounting to 754 814 words of running text covering periods up to 1950 (for the periods 1951-1990 and 1990-2021 no available texts were found in the Bulgarian National Corpus with open access).

Additionally, to cover the periods of 1951-1990 and 1990-2021 texts were collected automatically from the following sources: Liternet, Slovo, Project Gutenberg, Leteraturen svyat, Kultura, etc.

11 new texts were added amounting to 341 926 words.

All text units are supplied with extensive metadata following the metadata conventions of the Bulgarian National Corpus.

Description

Time periods: 1851-1880; 1881-1910; 1911-1930; 1931-1950; 1951-1990; 1991-2021.

Domains: fiction, news, science. The selection of domains was based on observations on the coverage of the domains across time periods. Administrative and other types of texts are rare in the earlier periods and are thus not included in the Diachronic Corpus.

Period	Number of texts	Number of words	Number of authors	Domains
1850-1880	5	154 886	4	Fiction, News, Science
1881-1910	10	252 426	7	Fiction, News, Science
1911-1930	24	180 241	10	Fiction, News, Science
1931-1950	15	167 261	5	Fiction, News, Science
1951-1990	5	195 500	5	Fiction
1991-2021	6	146 426	6	Fiction
TOTAL	65	1 096 740	37	Fiction, News, Science

Access

The following subcorpora are available for download:

- Corpus from the period of 1850 – 1880: ➥ Corpus download (with metadata)
- Corpus from the period of 1881 – 1910: ➥ Corpus download (with metadata)
- Corpus from the period of 1911 – 1930: ➥ Corpus download (with metadata)

- Corpus from the period of 1931 – 1950: ➥ Corpus download (with metadata)
- Corpus from the period of 1951 – 1990: ➥ Corpus download (with metadata)
- Corpus from the period of 1991 – 2021: ➥ Corpus download (with metadata)

Frequency analysis

General data

Period	#tokens	#words	#sents	#clauses	#nouns	#verbs	#adj	#adv
1850-1880	154886	137656	6499	15550	38088	21055	17072	7059
1881-1910	252426	215243	12655	29963	57533	36832	21293	12584
1911-1930	180241	168087	8689	20849	46590	26533	17671	9836
1931-1950	167261	149937	7736	17795	41590	22271	17155	8854
1951-1990	195500	132885	9015	20402	30467	25070	11072	10071
1991-2021	146426	133989	9375	19083	38727	23578	12031	7358

Statistical analysis shows relatively similar distribution of words, word-to-lemma ratio, etc. across time periods. Similar is also the complexity of sentence structure in terms of number of classes in the sentence (average of 2.0 to 2.4 clauses per sentence). Significant divergence is observed in the length of the sentence – 21.2 in the earliest period (1850-1880) down to 14.7 and 14.3 in the latest ones (1951-1990 and 1991-2021). Further analysis is needed to confirm whether this is due to text selection.

Observations on the top 100 most frequent word senses from each time period shows: (a) A total of 222 word senses appear in top 100 of at least one time period. (b) 44 senses appear in all time periods and further 13 appear in 5 out of the 6 periods – these cover frequent words of general meaning such as godina (year),

myasto (place), pat (road), mislya (believe), balgarski (Bulgarian), rabota (work), imam (have), zhivot (life), etc.

The number of words (word forms and lemmas) and sentences in the 6 time periods of the Diachronic corpus of Bulgarian

Period	#words per sent	#clauses per sent	#unique wordforms per 1000 words	#unique lemmas per 1000 words
1850-1880	21.18	2.39	161.57	92.97
1881-1910	17.01	2.37	174.28	100.13
1911-1930	19.34	2.40	176.24	98.84
1931-1950	19.38	2.30	185.84	106.09
1951-1990	14.74	2.26	200.29	109.08
1991-2021	14.29	2.04	198.37	111.03

Semantic Analysis

We have considered the filtered senses extracted after automatic semantic disambiguation of the texts in the corpus. The analysis is aimed in several directions:

(1) We have counted all occurrences of each BabelNet sense (counting together all lemmas representing the sense) and analysed: (a) senses found

across different time periods; (b) discrepancy in frequency in different periods which possibly marks changes in the usage of these concepts in the language. Below we give a number of examples.

The examples show possible analysis based on the data which is only reliable for high-frequency words with more occurrences in the corpus. Proper conclusions and hypothesis testing can be carried out after thorough analysis on more data. In the examples below we label frequency across periods using 6 symbols (e.g., XXXXXX, 00XXXX, etc.) where X marks occurrence in the period and 0 marks no occurrences.

Example 1. Words with general meaning occurring in all periods but with lower frequency in certain periods and relatively equal distribution in the rest of the periods
bn:00036632n A group of people with a common ideology who try together to achieve certain general goals движение_NOUN (dvizhenie / movement) Total freq. 203 (XXXXXX)
1850-1880 43 21.2 (%)
1881-1910 35 17.2
1911-1930 57 28.1
1931-1950 59 29.1
1951-1990 7 3.4
1991-2021 2 1.0
In recent years the use of the word *dvizhenie / movement* in the sense of 'A group of people with a common ideology' decreases.

Example 2. Words occurring in all periods but with lower frequency in certain periods and high frequency in a particular period
bn:00018819n One of the groups of Christians who have their own beliefs and forms of worship църква_NOUN (tzarkva / church) Total freq. 186 (XXXXXX)
1850-1880 108 58.1 (%)
1881-1910 24 12.9
1911-1930 18 9.7
1931-1950 30 16.1
1951-1990 3 1.6

1991-2021 3 1.6
There is a peak in the usage of the word *tzarkva / church* in the period of the Bulgarian Revival and it is not particularly relevant in more recent periods.

Example 3. Words occurring in earlier periods but not in later ones
(a)
bn:00060072n A civil or military authority in Turkey or Egypt паша_NOUN (pasha / Turkish leader) Total freq. 92 (XXXX00)
1850-1880 19 20.7
1881-1910 24 26.1
1911-1930 22 23.9
1931-1950 27 29.3
1951-1990 0.0
1991-2021 0.0

(b)
bn:00069885n A particular orthography or writing system писменост_NOUN (pismenost / written word) Total freq. 76 (XXXX00)
1850-1880 10 13.2
1881-1910 62 81.6
1911-1930 3 3.9
1931-1950 1 1.3
1951-1990 0.0
1991-2021 0.0

(c)
bn:00074113n A boat propelled by a steam engine параход_NOUN (parahod / steam ship) Total freq. 40 (XXX000)
1850-1880 10 13.2
1881-1910 62 81.6
1911-1930 3 3.9
1931-1950 1 1.3
1951-1990 0.0
1991-2021 0.0

(d)
bn:00012135n A learned person (especially in the humanities); someone who by long study has gained mastery in one or more disciplines книжар_NOUN (knizhar / learned person, writer); книжовник_NOUN (knizhovnik / learned person, writer)
Total freq. 28 (XXXX00)
1850-1880 13 46.4
1881-1910 7 25.0
1911-1930 6 21.4
1931-1950 2 7.1
1951-1990 0.0
1991-2021 0.0

The low or zero frequency of some words in most recent periods might be due to the fact that the concept was relevant in a particular period and then lost its significance (e.g., *pasha / Turkish leader* and *knizhovnik / learned person* were relevant to the period of the Bulgarian Revival and are now considered historical concepts) or is not used in contemporary times (e.g., *parahod / steam boat*).

Example 4. Words occurring only in recent periods
(a)
bn:00023101n An upholstered seat for more than one person диван_NOUN (divan / sofa) 14 (000XXX)
1850-1880 0.0
1881-1910 0.0
1911-1930 0.0
1931-1950 1 7.1
1951-1990 9 64.3
1991-2021 4 28.6

(b)
bn:00013723n The act of constructing something строителство_NOUN (stroitelstvo / builidng) 8 (0000XX)
1850-1880 0.0
1881-1910 0.0

1911-1930 0.0
1931-1950 0.0
1951-1990 6 75.0
1991-2021 2 25.0

(c)
bn:00109067a Engaged in a profession or engaging in as a profession or means of livelihood професионален_ADJ (profesionalen / professional) 4 000XXX
1850-1880 0.0
1881-1910 0.0
1911-1930 0.0
1931-1950 1 25.0
1951-1990 2 50.0
1991-2021 1 25.0

(d)
bn:00055448n Representation of something (sometimes on a smaller scale) модел_NOUN (model / model) 7 00000X
1850-1880 0.0
1881-1910 0.0
1911-1930 0.0
1931-1950 0.0
1951-1990 0.0
1991-2021 7 100.0

Some new concepts and words emerge in line with the modernisation of the life and household (e.g., *divan / sofa* or *model / model*) or in some cases related to the changes in the political and social system (e.g. *stroitelstvo / building* became particularly relevant in Comunist times used both literally and metaphorically).

Example 5. Words occurring only in earliest and most recent periods and not in the middle periods

bn:00047693n The act of issuing printed materials издаване_NOUN; 7 XX000X
1850-1880 4 57.1

1881-1910 2 28.6
1911-1930 0.0
1931-1950 0.0
1951-1990 0.0
1991-2021 1 14.3
The word *izdavane / publishing* is relevant in the period of the Bulgarian Revival and gathers popularity in the most recent periods as well in particular related to digital media.

(2) For each BabelNet synset (sense) we considered the occurrences of all lemmas of that sense. We are interested in noticeable differences in the usage of different lemmas which can show speakers' preferences for certain lexical units in general or in different time periods.

Example 6. Synonyms that are both equally used across all time periods
bn:00021644n A state at a particular time състояние_NOUN (sastoyanie / state); положение_NOUN (polozhenie / state, position)

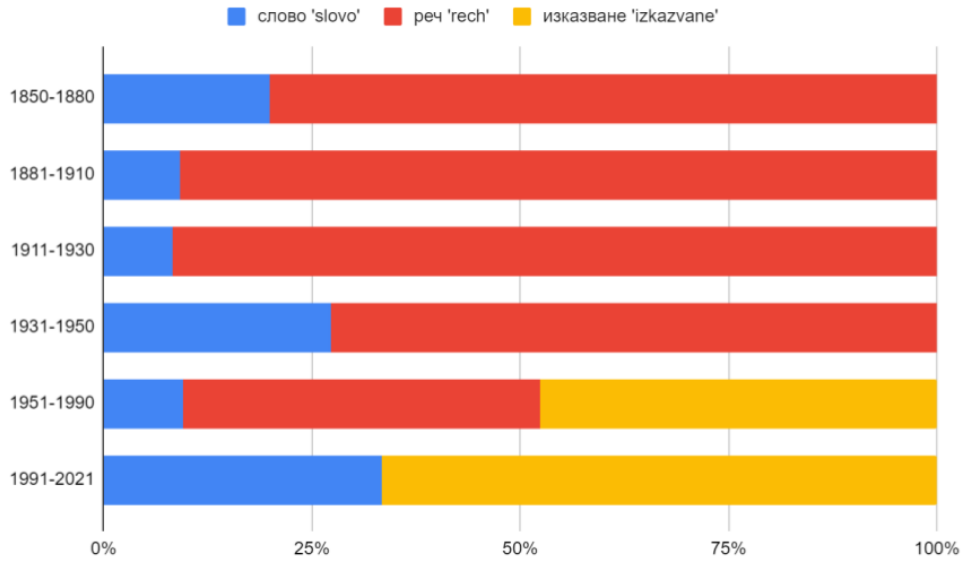
Example 7. Synsets for which a certain synonym dominates the usage across all time periods.
bn:00028934n The solid part of the earth's surface земя_NOUN (zemya / ground); суша_NOUN (susha / hard ground)
The word *zemya / ground* is preferred (375 occurrences) over *susha / hard ground* (3 occurrences), all appearing in different time periods.

bn:00005846n A distinctive odor that is pleasant мирис_NOUN (miris / smell); миризма_NOUN (mirizma / smell)
The word *mirizma / smell* is preferred (12 occurrences) over *miris / smell* (5 occurrences), all appearing in different time periods.

Example 8. Cases where a synonym is used only in earlier time periods.
bn:00003242n The feeling that accompanies something extremely surprising удивление_NOUN (udivlenie / amazement); изумление_NOUN (izumlenie / amazement)

The word *izumlenie / amazement* is much less frequent and appears only in earlier texts before 1910 while *udivlenie / amazement* is used in later periods as well (except in the most recent one).

Example 9. Cases where a synonym is used only in later time periods.
bn:00001304n The act of delivering a formal spoken communication to an audience реч_NOUN (rech / speech); изказване_NOUN (izkazvane / speech); слово_NOUN (slovo / speech)
As the diagram below shows, the word *izkazvane / speech* only appears in recent times and takes over the usage of other words in the synset which are prevailing in earlier periods.



(3) For each word (main form) we can further analyse the different senses it appears with. This can be useful to analyse the development of polysemy and occurrence of semantic neologisms.

Example 10. Metaphorical use of a word in certain periods.
bn:00074459n A violent commotion or disturbance буря_NOUN (burya / storm) (1850-1880 and 1881-1910) Total freq. 6
bn:00074458n bn:00074458n A violent weather condition with winds 64-72 knots (11 on the Beaufort scale) and precipitation and thunder and lightning буря_NOUN (burya / storm) (all time periods) Total freq. 58

The example shows the metaphorical use of the word *burya / storm* in texts of the earlier periods related to describing riots and national suffering before the liberation of Bulgaria from the Ottomans.

Example 11. Use of polysemous words and homonyms.
(a)
bn:00062658n A long-handled hand tool with sharp widely spaced prongs for lifting and pitching hay вила_NOUN (vila / pitchfork)
bn:00080000n Country house in ancient Rome consisting of residential quarters and farm buildings around a courtyard вила_NOUN (vila / villa)
The word *vila / pitchfork* is found only in texts from the period 1911-1930, while *vila / villa* occurs only from 1931 onwards.
(b)
bn:00036632n A group of people with a common ideology who try together to achieve certain general goals движение_NOUN (dvizhenie / movement)
bn:00056030n A natural event that involves a change in the position or location of something движение_NOUN (dvizhenie / movement)
Both appear in all time periods.

(c)
bn:00023310n An area wholly or partly surrounded by walls or buildings двор_NOUN (dvor / yard)
bn:00024541n The enclosed land around a house or other building двор_NOUN (dvor / yard)
bn:00023306n The family and retinue of a sovereign or prince двор_NOUN (dvor / royal court)
The first two senses appear in all time periods while the last one is used only in texts up to 1950; it probably loses popularity due to the fact that Bulgaria becomes a republic.

Example 12. Cases where a word develops new senses.
bn:00034949n An object firmly fixed in place (especially in a household) инсталация_NOUN (instalatziya / installation)

bn:00046934n The act of installing something (as equipment) инсталация_NOUN (instalatzita / installation)
The first sense of the word appears after 1930 and the second one only after 1990, probably related to the boom of new technologies.

Bulgarian National Corpus webpage
http://dcl.bas.bg/bulnc/en/

Bulgarian National Corpus search engine
http://search.dcl.bas.bg/

Diachronic Corpus of Bulgarian webpage
http://dcl.bas.bg/bulnc/en/diachronic-corpus/

Department of Computational Linguistics, Institute for Bulgarian Language
http://dcl.bas.bg/