Диахронен корпус на български език

Компилация на корпуса

Първоначално от Българския национален корпус бяха извлечени 54 текста, наброяващи 754 814 думи общо за времевите периоди до 1950 (за периодите 1951-1990 и 1990-2021 не бяха намерени подходящи текстове в БНК със свободни права).

Допълнително за обогатяване на данните за периодите 1951-1990 и 1990-2021 бяха събрани автоматично текстове от следните източници: Литернет, Словото, Project Gutenberg, Литературен свят, Култура и др.

Бяха включени 11 нови текста със свободен лиценз, наброяващи 341 926 думи.

Всички текстове са снабдени с подробни метаданни по модела на метаданните в Българския национален корпус.

Описание

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

Тематични области: художествена литература, публицистика, научна литература. Подборът е направен въз основа на наблюдението, че тези области имат покритие през повечето периоди, макар и с различно разпределение и покритие. Административни и други текстове не са включени, тъй като не са добре представени в по-ранните периоди.

Период	Брой текстове	Брой думи	Брой автори	Покритие на тематични области
1850-1880	5	154 886	4	Художествена, Публицистична, Научна
1881-1910	10	252 426	7	Художествена, Публицистична, Научна
1911-1930	24	180 241	10	Художествена, Публицистична, Научна
1931-1950	15	167 261	5	Художествена, Публицистична, Научна
1951-1990	5	195 500	5	Художествена
1991-2021	6	146 426	6	Художествена
ОБЩО	65	1 096 740	37	Художествена, Публицистична, Научна

Достъп

Достъпни за сваляне са подкорпуси за следните периоди:

- Корпус с текстове от периода 1850 – 1880: ➥ Корпус (с метаданни)
- Корпус с текстове от периода 1881 – 1910: ➥ Корпус (с метаданни)
- Корпус с текстове от периода 1911 – 1930: ➥ Корпус (с метаданни)
- Корпус с текстове от периода 1931 – 1950: ➥ Корпус (с метаданни)
- Корпус с текстове от периода 1951 – 1990: ➥ Корпус (с метаданни)

- Корпус с текстове от периода 1991 – 2021: ➥ Корпус (с метаданни)

Честотен анализ

Общи данни за Диахронния корпус на български език

Период	#токъни	#думи	#изр-я	#прости изр-я	#същ.	#глаголи	#прил.	#наречия
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

Статистическият анализ на данните показва сходно разпределение на думите, съотношение между брой графични думи и брой леми и т.н. през различните времеви интервали. Подобна е и мярката за сложност на изреченията, измерена въз основа на броя прости изречения в рамките на сложното (средно между 2.0 и 2.4 прости изречения). По-значителна е разликата в дължината на изречението, измерена като брой думи – 21.2 в най-ранния етап (1850-1880), която спада на 14.7 и 14.3 съответно в последните два етапа (1951-1990 и 1991-2021). По-задълбочен анализ е необходим, за да се потвърди дали причината за това е в подбора на текстовете.

Наблюденията върху 100-те най-често употребявани понятия (уникални значения, senses) за всеки времеви период показаха следното: (а) Общо 222 значения се появяват общо в списъците със 100-те най-често употребявани значения за всичките 6 периода (което показва значително припокриване

между тях). (b) 44 значения се срещат във всичките 6 периода, а още 13 се срещат в 5 периода – това са високочестотни думи от общата лексика като година, място, път, мисля, български, работа, имам, живот и др.

Броят думи (форми и леми) и изречения за всеки от шестте времеви периода в Диахронния корпус на български език

Период	#думи / изр.	#прости изр. / изр.	#уникални форми / 1000 думи	#уникални леми / 1000 думи
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

Примери за семантичен анализ

Разглеждаме филтриран списък със значения, извлечен след автоматично отстраняване на семантичната многозначност в текстовете от корпуса. Анализът е в следните насоки:

(1) Преброяваме срещанията на всяко значение от Бейбълнет (като преброяваме всички леми, които са използвани с това значение) и анализираме: (а) значения, които се срещат във всички времеви периоди – това са думи, запазили своята употреба във времето; (б) разминавания в употребата на думите в различните периоди, което може да показва промени в употребата на дадени понятия в езика в зависимост от времевия период. По-долу привеждаме примери и техния анализ.

Примерите представят възможен анализ въз основа на наличните данни, който е сравнително достоверен само за думите с по-висока честота и повече срещания в корпуса. Потвърждаването на изведените хипотези може да стане само след по-задълбочено разглеждане на повече данни. В примерите отбелязваме честотния модел с 6 символа (напр. XXXXXX, 00XXXX и т.н.), където с X означаваме, ако думата / значението се среща през периода, 0 – ако не се среща.

Пример 1. Думи с общо значение, срещащи се във всички периоди, но със значително по-ниска честота в даден период.
bn:00036632n Група хора с обща идеология и цели движение_NOUN Брой срещания 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
В последните периоди намалява употребата на думата ‘движение’ в значението ѝ на група хора с обща идеология.

Пример 2. Думи, срещащи се във всички периоди с подчертано по-малка честота в даден период и пик в друг период.

bn:00018819n Християнска група, подразделение на християнството, със собствени вярвания и религиозни обичаи църква_NOUN (church) Брой срещания 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
Има пик на употребите на думата ‘църква’ в текстове през Възраждането, а употребата ѝ в последните периоди намалява.

Пример 3. Думи, които се срещат в ранните периоди, но не и в по-късните.
(а)
bn:00060072n Представител на цивилната или военната власт в Турция и Египетпаша_NOUN Брой срещания 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

(б)
bn:00069885n Ортографична система писменост_NOUN Брой срещания 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

(в)
bn:00074113n Кораб, задвижван с пара параход_NOUN (steam ship) Брой срещания 40 (XXX000)
1850-1880 2 5.0
1881-1910 5 12.5
1911-1930 33 82.5
1911-1930 0.0
1951-1990 0.0
1991-2021 0.0

(г)
bn:00012135n Учен човек (в областта на хуманитарните науки) книжар_NOUN; книжовник_NOUN; Брой срещания 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
Несрещането или ниската честота на някои думи в последните периоди, може да се дължи на това, че понятието е било актуално за даден исторически период и е загубило актуалност (напр. ‘паша’ и ‘книжовник’ са актуални през Възраждането, а сега са вече исторически понятия) или не се използва в съвременния живот (напр. ‘параход’).

Пример 4. Думи, които се срещат само в по-късни периоди
(а)
bn:00023101n Тапицирана седалка за повече от един човек диван_NOUN; 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

(б)
bn:00013723n Дейността по построяване на нещо (сгради и под.) строителство_NOUN 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

(в)
bn:00109067a Който е свързан с професия или изкарване на прехраната професионален_ADJ; 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

(г)
bn:00055448n Представяне, образ на нещо (обикновено в умален размер или схематичен вид) модел_NOUN; 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

Пример 5. Думи, които се срещат само в най-ранните и най-късните периоди, но не и в междинните

bn:00047693n Дейността по публикуване, оповестяване 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
Думата ‘издаване’ е актуална през Възраждането и след Освобождението, а в последния период придобива и нови значения, свързани с издаване на електронни медии.

(2) За всяко значение в Бейбълнет разглеждаме срещанията на всички леми с това значение. Интересуваме се от значителни разлики в употребата на дадени думи пред други думи със същото значение, което показва тенденции и предпочитания на носителите на езика към дадени лексикални единици през определени периоди.

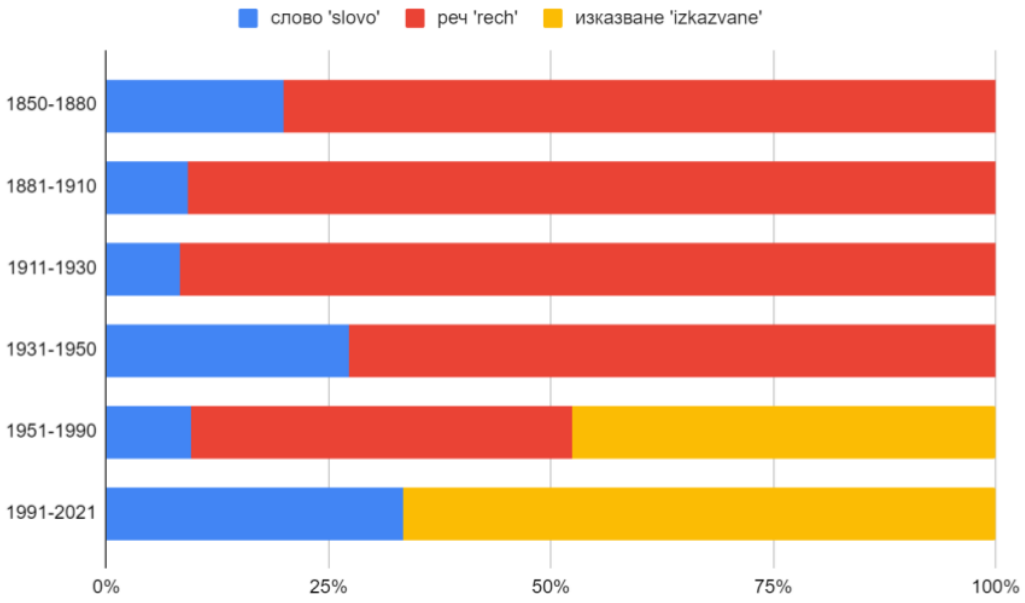
Пример 6. Синоними, които имат сравнително равномерна употреба през различните периоди.
bn:00021644n Съществуване (за даден период) в определена форма или по определен начин състояние_NOUN; положение_NOUN

Пример 7. Значения, при които даден синоним консистентно доминира в употребата през всички периоди.
(a)
bn:00028934n Твърдата външна повърхност на планетата Земя земя_NOUN; суша_NOUN;
Думата ‘земя’ чувствително се предпочита (375 срещания) пред ‘суша’ (3 срещания) и това е валидно за всички периоди.

(б)
bn:00005846n Излъчване от страна на обект, което се възприема чрез носа мирис_NOUN; миризма_NOUN
Думата ‘миризма’ се предпочита (12 срещания) пред ‘мирис’ (5 срещания) през различни периоди.

Пример 8. Случаи, при които даден синоним се използва само в по-ранните периоди.
bn:00003242n Чувство на силна изненада удивление_NOUN; изумление_NOUN;
Думата ‘изумление’ е със значително по-ниска честота и се среща само в по-ранните периоди преди 1910, докато ‘удивление’ се употребява и в по-късните периоди (с изключение на последния).

Пример 9. Случаи, при които даден синоним се използва само в последните периоди.
bn:00001304n Представяне в устна форма пред публика реч_NOUN; изказване_NOUN; слово_NOUN



Както се вижда и от диаграмата, ‘изказване’ се среща само в последните периоди, но започва да доминира над другите два синонима, които се срещат с голяма честота в ранните периоди.

(3) За всяка дума (основна форма) анализираме различните значения, с които думата се среща. Това може да е полезно за анализ на развитието на полисемия и появата на семантични неологизми.

Пример 10. Метафорична употреба на думи в даден период.

bn:00074459n Безредици, свързани с насилие буря_NOUN (1850-1880 и 1881-1910) Брой срещания 6
bn:00074458n Атмосферно време, свързано със силен вятър, дъжд, често и светкавици и гръмотевици буря_NOUN (всички периоди) Брой срещания 58
Примерът показва употреба на думата ‘буря’ в ранни текстове в метафорична употреба при описване на живота и борбите на българите по време на османската власт.

Пример 11. Употреба на многозначни думи и омоними.
(а)
bn:00062658n Селскостопански инструмент с няколко остри пръчки в края, който се използва най-често за слама вила_NOUN
bn:00080000n Провинциална къща за отдих вила_NOUN
Думата ‘вила’ (селскостопански инструмент) се среща само в текстове от периода 1911-1930, докато ‘вила’ (къща за отдих) се среща само след 1931.

(б)
bn:00036632n Група хора с обща идеология и цели движение_NOUN
bn:00056030n Спонтанно или причинено събитие, при което обект променя местоположението си движение_NOUN
И двете се срещат във всички времеви периоди.

(в)
bn:00023310n Пространство, оградено със стени, сгради или ограда двор_NOUN

bn:00024541n Прилежащо пространство към къща или друга сграда двор_NOUN
bn:00023306n Най-близкото обкръжение на монарх двор_NOUN
Докато първите две значения се срещат през всички периоди, последното се среща само до 1950 и вероятно губи актуалност, тъй като България става република.

Пример 12. Случаи, когато дума придобива нови значения.
bn:00034949n Обект, фиксиран на определено място (обикновено свързано с поддръжката на дома) инсталация_NOUN
bn:00046934n Привеждане на даден уред в готовност за употреба инсталация_NOUN
Първото значение се среща след 1930 година, а второто – само след 1990, вероятно свързано с бума на технологиите.

Страница на Българския национален корпус
http://dcl.bas.bg/bulnc/en/

Интерфейс за търсене в Българския национален корпус
http://search.dcl.bas.bg/

Страница на Диахронния корпус на български език
http://dcl.bas.bg/bulnc/en/diachronic-corpus/

Секция по компютърна лингвистика,
http://dcl.bas.bg/