BNC2014 Baby+

The BNC2014 Baby+ was compiled by Vaclav Brezina, Lancaster University.

The Written BNC 2014 Project Lead: Vaclav Brezina and Tony McEnery, Lancaster University.

Acknowledgement: A large number of people were involved in the data collection of the corpus. Special thanks are due to Robbie Love and Abigail Hawtin, whose PhD work focused on proposing the sampling frame for the spoken and written corpora respectively. Thanks are also due to Andrew Hardie for the creation and distribution of xml files of the Spoken BNC2014. A number of external partners were involved in providing the data for the corpus. The Spoken BNC2014 was created in collaboration with Cambridge University Press. The Written BNC2014 was created with the assistance of Cambridge University Press, Dunedin Academic Press, Elsevier and John Benjamins. The following researchers helped with the data collection: Zoé Broisson, Carmen Dayrell, Mathew Gillings, Andressa Rodrigues Gomide, Vasiliki Simaki, Matt Timperley and Isolde van Dorst.

Funding: The project has been supported by ESRC grants no. EP/P001559/1, ES/K002155/1 and ES/R008906/1.

Figure 1. BNC2014 data sources and family of BNC2014 corpora

Overview

The BNC2014 Baby+ has been extracted from the BNC2014 data pool to mark the second stage of release of The British National Corpus 2014 (BNC2014). The British National Corpus 2014 is a major project led by Lancaster University to create a 100-million-word corpus (a large collection of 'real life' language) of modern-day British English. This corpus will be used by researchers to understand more about how language works and how it is evolving. Educators, dictionary compilers and the interested public will also be able to access the corpus to find usage examples of modern British English in different genres.

In the first stage, the Spoken BNC2014 (Love et al. 2017) has been released. It can be accessed online via CQPweb or BNClab and is also available for download from http://corpora.lancs.ac.uk/bnc2014/.

The second stage involves a release of a balanced subset of the BNC2014, the BNC2014 Baby +. The BNC2014 Baby + is a mirror corpus to the BNC1994 Baby (Burnard 2003), a four-million-word sample of British English from the early 1990s. In addition to traditional genres (academic, fiction, newspaper and conversation), the BNC2014 Baby + includes data from the e-language subcorpus of the Written BNC2014. The corpus is made available via #LancsBox (Brezina et al. 2015), a corpus analysis package developed at Lancaster University. The corpus was built with different research purposes and research designs in mind (Brezina 2018).

Spoken component

The spoken component of the BNC2014 Baby + comes from the balanced subset (core) of the Spoken BNC2014 (Love et al. 2017). This subset was developed as part of the project 'The British National Corpus (BNC) as a sociolinguistic dataset: Exploring individual and social variation' funded by the ESRC (grant no. EP/P001559/1, Brezina et al. 2018). The core contains data that is broadly representative of the UK

population in terms of age, gender, region, and class and that does not have any major missing values in the meta-data categories. The speakers selected for the core also contribute substantially to the word count of the corpus (i.e. are prolific speakers). 29 speakers as the main interlocutors were randomly chosen from 250 core speakers. Their production is included as part of an interaction with other speakers. Table 1 presents an overview of the 29 core speakers included in the corpus.

Table 1. Overview of BNC Baby + core speakers

Speaker_ID	ID in Spoken BNC2014	File tokens	Core speaker tokens	Pseudonym	Gender	Age	Region	Social class	Occupation
BNCBSp1	S0197	81293	16622	Casey	male	51	London	middle class	buyer
BNCBSp2	S0518	49469	36102	Samantha	female	48	Midlands	working class	customer service
BNCBSp3	S0483	25592	15475	Kristin	female	37	North England	middle class	manager
BNCBSp4	S0026	33721	12902	Sam	male	23	North England	working class	engineer
BNCBSp5	S0307	9436	3364	Nora	female	30	Midlands	middle class	teacher
BNCBSp6	S0610	23444	7652	Tanya	female	28	Wales	middle class	editor
BNCBSp7	S0320	61328	35717	Stacy	female	34	Southwest	middle class	teacher
BNCBSp8	S0585	57450	35807	Marion	female	21	North England	student	student
BNCBSp9	S0450	36022	22323	Tara	female	24	Midlands	student	student
BNCBSp10	S0538	21756	12628	Brent	male	73	North England	retired	retired
BNCBSp11	S0393	95701	18289	Tim	male	47	North England	middle class	writer
BNCBSp12	S0680	34432	22367	Carla	female	64	Southeast	retired	retired
BNCBSp13	S0626	17206	4931	Harvey	male	29	North England	middle class	filmmaker
BNCBSp14	S0094	33477	10525	Georgia	female	33	Southwest	student	student
BNCBSp15	S0628	9562	5667	Viola	female	29	London	middle class	writer
BNCBSp16	S0601	17653	2750	Mathew	male	25	Southeast	working class	assistant
BNCBSp17	S0536	3210	1387	Jordan	male	74	Southwest	retired	retired
BNCBSp18	S0070	33770	17605	Dean	male	44	Southwest	student	student
BNCBSp19	S0639	30950	1906	Jenny	female	44	London	working class	teacher
BNCBSp20	S0251	107073	31579	Robyn	female	52	North England	middle class	teacher
BNCBSp21	S0420	14728	7494	Angel	male	32	Scotland	middle class	director
BNCBSp22	S0312	9065	1282	Ken	male	32	Scotland	middle class	engineer
BNCBSp23	S0586	20974	4250	Christy	female	20	North England	student	student
BNCBSp24	S0648	11069	1075	Jasmine	female	45	London	middle class	media producer
BNCBSp25	S0668	31405	6860	Edgar	male	31	Wales	middle class	housing officer
BNCBSp26	S0587	86900	28346	Calvin	male	20	North England	student	student
BNCBSp27	S0195	14626	5381	Ron	male	28	Midlands	middle class	analyst
BNCBSp28	S0034	18892	4722	Jared	male	29	Southeast	working class	engineer
BNCBSp29	S0458	11592	3759	Iris	female	21	North England	middle class	receptionist

Newspapers component

The newspaper component of the BNC2014 Baby + comes from the pool of newspaper articles collected for the Written BNC2014. This pool contains over 83 thousand articles published in the period of 2014-2016. These articles were downloaded from online sources based on a list of random dates within the given sampling period. The newspapers are divided into three major categories i) serious, ii) mass market and iii) regional newspapers. For BNC2014 Baby +, a random set of newspapers was selected within each category; each category includes approximately one third of the total word count of the newspaper component. Texts vary considerably in length (51-6,210 tokens). The minimum article length imposed in the selection was 50 tokens. Tables 2, 3 and 4 show an overview of the BNC2014 Baby + dataset according to the newspapers included in the selection and the type of content covered. Because of this rigorous sampling technique, the newspaper component of the BNC2014 Baby + reflects relatively reliably the newspaper production in the UK with a range of topics covered in different sections of the newspapers.

Table 2. Serious newspapers

Newspaper	Texts	Words (tokens)
Financial Times	89	60,694
The Guardian	139	94,440
The Observer	24	19,365
The Sunday Times	84	47,536
The Times	274	112,840
TOTAL	610	334,875

The serious newspaper sections include:

ADVERTISING WEEK EUROPE;NEWS, ANALYSIS, ART AND DESIGN, AUSTRALIA NEWS, BOOK REVIEW, BOOKS, BUDGET 2016 - BUSINESS, BUDGET 2016 - COMMENT, BUSINESS, BUSINESS LIFE, BUSINESS; FRONT PAGE, BUSINESS; OPINION COLUMNS, CHILDREN'S BOOKS, COMMENT IS FREE, COMPANIES, COMPANIES - ROUND-UP, CULTURE;FEATURES, DRIVING;FEATURES, EDITORIAL, EDITORIAL; OPINION COLUMNS, EDITORIAL; OPINION; LEADING ARTICLES, ENVIRONMENT, FASHION, FAST TRACK 100;FEATURES, FEATURES, FEATURES; OFFERS, FEATURES; OPINION COLUMN, FILM, FOOTBALL, FRONT PAGE - COMPANIES & MARKETS, FRONT PAGE - FIRST SECTION, FT REPORT - ASIA-PACIFIC INNOVATIVE LAWYERS, FT REPORT - CORPORATE AVIATION, FT REPORT - FUND MANAGEMENT, FT REPORT - HOUSE & HOME, FT REPORT - LATIN AMERICAN CAPITAL MARKETS, FT REPORT - WATCHES & JEWELLERY, FT WEEKEND SUPPLEMENT - HOUSE & HOME, FT WEEKEND SUPPLEMENT - LIFE & ARTS, FT WEEKEND SUPPLEMENT - MONEY, GOOD UNIVERSITY GUIDE 2016;BUSINESS, GOOD UNIVERSITY GUIDE 2016;FEATURES, GUARDIAN COMMENT AND DEBATE PAGES, GUARDIAN FEATURES PAGES, GUARDIAN FILM AND MUSIC PAGES, GUARDIAN FINANCIAL PAGES, GUARDIAN HOME PAGES, GUARDIAN INTERNATIONAL PAGES, GUARDIAN LEADER PAGES, GUARDIAN REVIEW PAGES, GUARDIAN SATURDAY COMMENT PAGES, GUARDIAN SPORT PAGES, GUARDIAN TRAVEL PAGES, HEALTHCARE PROFESSIONALS NETWORK, HOME;FEATURES, LAW, LETTERS, LETTERS TO THE EDITOR, LIFE AND STYLE, LOMBARD, MARKETS, MARKETS & INVESTING, MEDIA, MONEY, MONEY;BUSINESS, MONEY;NEWS, MUSIC, NATIONAL NEWS, NEWS, NEWS REVIEW;FEATURES, NEWS REVIEW;NEWS, NEWS; FRONT PAGE, OBSERVER HOME NEWS PAGES, OBSERVER NEW COMMENT PAGES, OBSERVER OBSERVER SPORT PAGES, OBSERVER REVIEW AGENDA, OBSERVER REVIEW BOOKS

PAGES, OBSERVER WORLD NEWS PAGES, ON WORK, PERSONAL TECHNOLOGY, POLITICS, SATURDAY REVIEW;FEATURES, SCIENCE, SMALL TALK, SOCIETY, SPORT, SPORT; OPINION COLUMNS, SPORT;FEATURES, SPORT;SPORT; OPINION COLUMNS, STAGE, STUDENT LAW;NEWS, STYLE;FEATURES, SUMMER BUDGET - EMPLOYMENT, SUMMER BUDGET - POLITICS, T2;EDITORIAL, T2;FEATURES, T2;FEATURES; OPINION COLUMN, T2;NEWS, THE DISH;FEATURES, THE GAME;SPORT, TOP 50 EMPLOYERS FOR WOMEN;FEATURES, TRAVEL;FEATURES, UK COMPANIES, UK NEWS, US NEWS, WEEKEND;FEATURES, WOMEN IN LEADERSHIP, WORLD NEWS

Table 3. Mass-market newspapers

Newspaper	Texts	Words (tokens)
Daily Star	286	66,527
Daily Star Sunday	66	16,977
Sunday Express	75	22,069
Sunday Mirror	59	20,144
The Express	283	79,932
The Mirror	181	44,781
The Sun	360	83,586
TOTAL	1310	334,016

The mass-market newspaper sections include:

BUSINESS, BUSINESS; OPINION, COLUMNS, EDITORIAL, EDITORIAL; OPINION COLUMNS, EDITORIAL; OPINION, COLUMNS, EDITORIAL; OPINION, LEADING ARTICLES, EDITORIAL; OPINION; LEADING ARTICLES, FABULOUS;FEATURES, FAVOURITE;SPORT, FAVOURITE;SPORTS; FRONT PAGE, FEATURES, FEATURES; EXCLUSIVE, PROFILES AND INTERVIEWS, FEATURES; OFFERS, FEATURES; OPINION COLUMN, FEATURES; OPINION, COLUMN, FEATURES; REVIEW, FINANCIAL;BUSINESS, FINANCIAL;BUSINESS; DIARY, FINANCIAL;BUSINESS; FRONT PAGE, FINANCIAL;NEWS, GOALS;SPORT, GOALS;SPORT; OPINION COLUMNS, LETTERS, ME;FEATURES, MIRROR FOOTBALL;SPORT, MIRROR RACING;FEATURES, MRS CRUNCH;FEATURES; OPINION COLUMN, NEWS, NEWS; DIARY, NEWS; EXCLUSIVE, NEWS; EXCLUSIVES, PROFILES AND INTERVIEWS, NEWS; FRONT PAGE, NEWS; OPINION COLUMNS, NEWS; OPINION, COLUMNS, NEWS; OPINION; COLUMNS, NOTEBOOK;FEATURES, RESULT, RESULT;SPORT, ROYAL ASCOT;SPORT, SERIOUSLY FOOTBALL;FEATURES, SERIOUSLY FOOTBALL;SPORT, SPORT, SPORT; OPINION COLUMNS, SPORT; OPINION, COLUMNS, STAR FORM;FEATURES, STAR FORM;SPORT, SUNDAY MIRROR FOOTBALL;SPORT, SUPER GOALS;SPORT, THE ROYAL TOUR;FEATURES, THE TICKET;FEATURES; OPINION COLUMN, TRAVEL;FEATURES

Table 4. Regional newspapers

Newspaper	Texts	Words (tokens)
Belfast Telegraph	108	42,384
Bham Evening Mail	76	23,958
Daily Record and Sunday Mail	119	34,799
East Anglian Daily Times	48	17,648

Evening Standard	118	39,883
Hull Daily Mail	75	24,480
Leicester Mercury	80	26,241
Liverpool Echo	93	35,048
Manchester Eve News	85	27,598
South Wales Argus	44	14,157
South Wales Echo	60	20,837
The Herald	26	11,043
Western Morning News	48	18,550
TOTAL	980	336,626

The regional newspaper sections include:

AGENCY:CITY, AGENCY:ENTERTAINMENT, AGENCY:MEDIA, AGENCY:NEWS, AGENCY:OTHER, AGENCY:SPORT, BKMF, BMDS, BUSINESS, BUSINESS MONTH;BUSINESS, BUSINESS WEEK;BUSINESS, BUSINESS WEEK;NEWS, BUSINESS:AGRICULTURE, BUSINESS:CONSUMER, BUSINESS:ECONOMY, BUSINESS:OTHER, BUSINESS; DIARY, BUSINESS; OPINION COLUMNS, CITY LIFE;FEATURES, COURT:CROWN, COURT:MAGISTRATES, CURVE;FEATURES, DEATHS, EACO, EDITORIAL; OPINION COLUMNS, EDITORIAL; OPINION, COLUMNS, EDITORIAL; OPINION; LEADING ARTICLES, ENTS:BOOKS, ENTS:FILM, ENTS:MUSIC, ENTS:OTHER, ENTS:THEATRE, ENVI, ES MAGAZINE;FEATURES, FEAT, FEATURES, FEATURES:ADVERTORIAL, FEATURES:HISTORY, FEATURES:HOBBIES, FEATURES:MOTORS, FEATURES:OTHER, FEATURES:PEOPLE, FEATURES:PROPERTY, FEATURES; DIARY, FEATURES; OPINION COLUMN, FEATURES; OPINION, COLUMN, FEATURES; REVIEW, FEMAIL;FEATURES, FOOTBALL ECHO;SPORT, FRIDAY LIVE;FEATURES, FRNT, GAME ON;NEWS, GO GREEN;NEWS, HEAT YOUR HOME;FEATURES, HOME AND AWAY;NEWS, HOMES AND PROPERTY;FEATURES, HS - BUSINESS, HS - FEATURES, HS - MOTORING, HS - NEWS, HS - SPORT, JUNIOR XSPORT;SPORT, LETT, LETTERS, LIFE STYLE;FEATURES, MAGAZINE;NEWS, MUSIC, NEWS, NEWS:ARMED FORCES, NEWS:CHARITY, NEWS:CRIME, NEWS:EDUCATION, NEWS:EMPLOYMENT, NEWS:ENVIRONMENT, NEWS:HEALTH, NEWS:LETTERS, NEWS:MOTORING, NEWS:OTHER, NEWS:PEOPLE, NEWS:POLITICS, NEWS; DIARY, NEWS; OPINION; COLUMNS, NEWS; TEASERS, NI JOB FINDER;NEWS, ROAD RECORD;FEATURES, ROP, SATURDAY EXTRA;FEATURES, SATURDAY EXTRA;NEWS, SATURDAY;FEATURES, SEVEN DAYS;FEATURES, SPORT, SPORT:BOXING, SPORT:CRICKET, SPORT:FOOTBALL, SPORT:LOCAL SPORT, SPORT:OTHER, SPORT:RUGBY LEAGUE, SPORT:RUGBY UNION, SPORT; OPINION COLUMNS, SPORT; OPINION, COLUMNS, SPRT, SUNDAY BRUNCH;FEATURES, THE GUIDE;FEATURES, THE GUIDE;NEWS, THE PUNTER;SPORT, TIME TO BE EURO STARS;SPORT, UNSOLVED;NEWS, WEEKEND;FEATURES, WEEKEND;NEWS, WINNER;SPORT, WOW BRUM;NEWS, WOW BRUM;NEWS; TEASERS, YOUR CITY YOUR RIVER;FEATURES, YOUR CITY YOUR RIVER;NEWS

Fiction component

The fiction component includes samples (beginnings, middles and ends) of 69 randomly selected books from the BNC2014 data pool. Each of the sample is approximately 15,000 tokens in length. The front matter (including, TOC, dedication and preface) and the back matter (including notes and author info) were not included in the sample. The 69 fiction book samples represent current British fiction published between 2010 and 2017. Each author appears only once in the subcorpus. Overall, the fiction subcorpus consists of 1,007,359 tokens and 36,627 types.

Table 5 provides details about the works sampled.

Table 5. Fiction sources

Author	Year	Title	File	Category
Guy Adams	2017	New York: Queen of Coney Island	BNCBFict_b1	beginning
Colin Bateman	2010	Dr Yes	BNCBFict_b2	beginning
MR Carey, Linda Carey and Louise Carey	2012	The City of Silk and Steel	BNCBFict_b3	beginning
Louise Welsh	2012	The Girl on the Stairs	BNCBFict_b4	beginning
Neil Cross	2011	Luther: The Calling	BNCBFict_b5	beginning
Nathan Filer	2013	The Shock of the Fall	BNCBFict_b6	beginning
Helen Simpson	2010	In-Flight Entertainment	BNCBFict_b7	beginning
Ian Whates, Storm Constantine and Frances Hardinge	2014	La Femme	BNCBFict_b8	beginning
Howard Jacobson	2010	The Finkler Question	BNCBFict_b9	beginning
Tom McCarthy	2015	Satin Island	BNCBFict_b10	beginning
Michael Smith	2015	The Debs of Bletchley Park and Other Stories	BNCBFict_b11	beginning
Lawrence Osborne	2017	Beautiful Animals	BNCBFict_b12	beginning
Sunjeev Sahota	2015	The Year of the Runaways	BNCBFict_b13	beginning
Charles Stross	2014	The Rhesus Chart	BNCBFict_b14	beginning
Jonathan Stroud	2017	The Empty Grave	BNCBFict_b15	beginning
Adrian Tchaikovsky	2015	Children of Time	BNCBFict_b16	beginning
SJ Watson	2011	Before I Go to Sleep	BNCBFict_b17	beginning
David Whitley	2010	The Children of the Lost	BNCBFict_b18	beginning
Naomi Alderman	2013	The Liars' Gospel	BNCBFict_e19	end
Kate Atkinson	2013	Life After Life	BNCBFict_e20	end
Mark Billingham	2017	Love Like Blood	BNCBFict_e21	end
Ann Cleeves	2010	Blue Lightning	BNCBFict_e22	end
Catherine Doyle	2017	Mafiosa	BNCBFict_e23	end
Jenni Fagan	2016	The Sunlight Pilgrims	BNCBFict_e24	end
Ken Follett	2017	A Column of Fire	BNCBFict_e25	end
Kate Furnivall	2012	The White Pearl	BNCBFict_e26	end
Matt Haig	2013	To Be a Cat	BNCBFict_e27	end
Robert Harris	2015	Dictator	BNCBFict_e28	end
Charlie Higson	2014	The Hunted	BNCBFict_e29	end
Katy Regan	2012	How We Met	BNCBFict_e30	end
Lee Child	2012	A Wanted Man	BNCBFict_e31	end
Deborah Levy	2011	Swimming Home. And other stories.	BNCBFict_e32	end
Ken MacLeod	2014	Descent	BNCBFict_e33	end
Mark Logue and Perter Conradi	2010	The King's Speech	BNCBFict_e34	end
Robert Galbraith	2013	The Cuckoo's Calling	BNCBFict_b35	beginning
Frederick Forsyth	2010	The Cobra	BNCBFict_b36	beginning

Ruth Rendell	2011	The Vault	BNCBFict_e37	end
Sarah Hall	2017	Madame Zero	BNCBFict_e38	end
Jill Paton Walsh	2014	The Late Scholar	BNCBFict_e39	end
Sophie Kinsella	2011	I've Got Your Number	BNCBFict_e40	end
Tom Pollock	2014	Our Lady of the Streets	BNCBFict_e41	end
Will Mabbitt	2016	Mabel Jones and the Doomsday Book	BNCBFict_e42	end
Alan Wright	2010	Act of Murder	BNCBFict_m43	middle
Adele Parks	2014	Spare Brides	BNCBFict_m44	middle
Alastair Reynolds	2010	House of Suns	BNCBFict_m45	middle
Jeffrey Archer	2013	Best Kept Secret	BNCBFict_m46	middle
Kevin Brooks	2013	The Bunker Diary	BNCBFict_m47	middle
Jessie Burton	2016	The Muse	BNCBFict_m48	middle
John le Carré	2013	A Delicate Truth	BNCBFict_m49	middle
Philip Caveney	2010	A Buffalope's Tale	BNCBFict_m50	middle
David Solomons	2015	My Brother is a Superhero	BNCBFict_m51	middle
Lindsey Davis	2015	The Spook Who Spoke Again: A Flavia Albia Short Story	BNCBFict_m52	middle
Louise Doughty	2010	Whatever You Love	BNCBFict_m53	middle
Paul Finch	2011	Hunter's Moon	BNCBFict_m54	middle
James Goss	2011	First Born	BNCBFict_m55	middle
L.S. Hilton	2017	Domina	BNCBFict_m56	middle
Tom Holt	2010	Blonde Bombshell	BNCBFict_m57	middle
Anthony Horowitz	2011	The House of Silk	BNCBFict_m58	middle
Lars Iyer	2013	Exodus	BNCBFict_m59	middle
Jacqueline Wilson	2015	Katy	BNCBFict_m60	middle
Joe Abercrombie	2016	Sharp Ends: Stories from the World of The First Law	BNCBFict_m61	middle
Jill Mansell	2013	Don't Want to Miss a Thing	BNCBFict_m62	middle
Mari Hannah	2010	The Murder Wall	BNCBFict_m63	middle
Jojo Moyes	2010	The Last Letter from Your Lover	BNCBFict_m64	middle
Adam Nevill	2015	Lost Girl	BNCBFict_m65	middle
James Robertson	2016	To Be Continued	BNCBFict_m66	middle
Robert Webb	2017	How Not To Be a Boy	BNCBFict_m67	middle
Simon Edge	2017	The Hopkins Conundrum	BNCBFict_m68	middle
Chris Wooding	2013	The Ace of Skulls	BNCBFict_m69	middle

Academic component

The academic component consists of i) academic books and ii) academic journals evenly distributed across six broad disciplinary areas:

1) Humanities and arts, 2) Medicine, 3) Natural science, 4) Politics, law and education, 5) Social science, and 6) Technology and engineering.

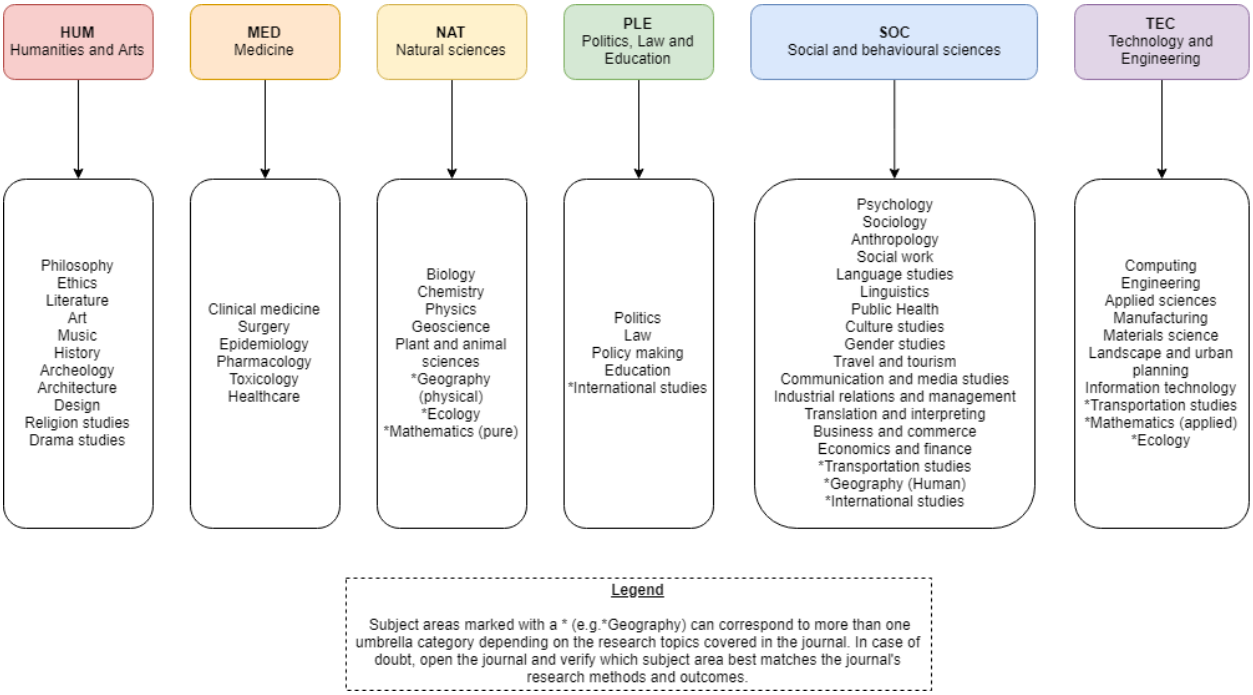

Figure 2. Disciplinary division of BNC2014 (thanks to Zoé Broisson)

Table 6. Academic sources

Category	Texts (books)	Tokens (books)	Texts (journals)	Tokens (journals)
Hum_arts (H)	11	85,117	22	86,328
Medicine (M)	7	85,600	19	83,584
Nat_science (N)	3	85,112	12	91,279
Polit_law_edu (P)	2	78,156	17	87,670
Soc_science (S)	2	80,331	26	70,873
Tech_engin (T)	2	85,237	20	83,960
Total	**27**	**499,553**	**116**	**503,694**

E-language component

Finally, the e-language component includes four types of new and emerging genres/registers reflecting the use of language in the online environment. This new component does not have a mirror sample in the original BNC1994 Baby or the original BNC for that matter. The e-language component includes 1) social media (facebook and twitter posts), 2) emails, 3) sms, 4) blogs, 5) discussion forums and 6) product reviews.

Table 7. E-language sources

Category	Texts	Tokens
Social media	2	196,282
Emails	84	24,333
Sms	22	196,911
Blogs	295	194,981
Discussion forums	39	194,675
Product reviews	39	195,699
Total	481	1,002,881

References

Burnard, Lou. 2003. Reference guide for BNC-baby. *Distributed by Oxford University Computing Services on behalf of the BNC Consortium*.

Brezina, V. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Brezina, Vaclav, Dana Gablasova & Susan Reichelt. 2018. BNClab. http://corpora.lancs.ac.uk/bnclab [electronic resource], Lancaster University.

Brezina, Vaclav, Tony McEnery & Stephen Wattam. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2). 139-173.

Hardie, Andrew 2012. CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics* 17(3). 380-409.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014. *International Journal of Corpus Linguistics* 22(3). 319-344.

How to build and analyze a general corpus?

Building a large general corpus is a major undertaking. The following exercises are intended to help discuss main ideas in the corpus design and analysis.

Key terms British English | sample | sampling frame | current

Task 1

What makes a good general corpus? Think of features that a large national corpus should have. Make a list of these in the space provided.

Task 2

In many ways, the four words in the name of the corpus i.e. 'British National Corpus 2014' summarize the essence of the dataset. Use the terms provided below and match them with the four words in the name of the corpus. What do these terms mean? Are they important for the quality of the corpus? Explain why.

BNC
BRITISH NATIONAL CORPUS 2014

1) .....
2) .....
3) .....
4) .....

balanced, British variety, current, range of genres/registers, representative, research design, robust, sampling frame, speech and writing, synchronic

Task 3

Search for words connected with technological development and recent social and political processes. Can you find these words in the BNC2014 Baby +?

e.g. Brexit, to milkshake, cloud, iPad, facebook, climate change

Task 4

Investigate the distribution of different words across the genres/registers of present-day British English. The graph below shows the distribution of the swearword 'shit' in the BNC2014 Baby+.