ВЕСТНИК ПЕРМСКОГО УНИВЕРСИТЕТА

2013 РОССИЙСКАЯ И ЗАРУБЕЖНАЯ ФИЛОЛОГИЯ Вып. 1(21)

УДК 811.111: 81’37

BACK TO THE FUTURE: THE BRITISH NATIONAL CORPUS

Valentina A. Kononova
Reader of the Department of Linguistics and Cross-cultural Communication
Siberian Federal University
660041, Krasnoyarsk, Svobodny Prospect, 82A. v.kononova@mail.ru

The article is devoted to student research opportunities within one of the largest language resources – the British National Corpus. The user-friendly annotations of this rich collection make it possible to effectively apply to the BNC for both research and English language learning. The British National Corpus is well-known as a completed project, it was created in 1990s, and since then it has been regarded as one of the largest and most varied corpora of spoken and written data yet compiled. In this article, we reintroduce some issues of the British National Corpus in its modernized BYU architecture of the year of 2012, in sense of new software applications for computer-assisted language research and language learning. The article provides some accessible research samples for students coming to empirical linguistics. The article is written within post-doc research of the ERANET MUNDUS project, at the University of Barcelona, Spain.

Key words: the British National Corpus; BNC-BYU; corpus linguistics; John Sinclair; Mark Davies; corpus annotation.

Introduction: new habitus

The British National Corpus is a huge structure with a lot of residents inside. The house is not full; the doors are open to all the newcomers, whether they are worldly-wise scholars or university students. It is also not brand-new, and some people [Moskovkin 2009] call it a bit outdated. In this article, we will reintroduce some issues of the British National Corpus in its modernized BYU architecture of the year of 2012. In general, here we focused again on the BNC content, and the opportunities it gives to researchers as exemplified in a number of real linguistic case-studies.

Huge electronic corpora collections have been available to researchers for almost half a century. In October 2012, Mark Davies announced seven new resources and improvements associated with the BYU corpora, including, but not limited to the newer CLAWS 7 tagset which allows better comparisons with the four other BYU corpora. [Announcements from corpus.byu.edu.]. Mark Davies is a world-known professor of Linguistics at Brigham Young University (that is where *BYU* in *BYU-BNC* comes from) in Provo, Utah, USA. His primary areas of research are corpus linguistics, the design and optimization of linguistic databases, language change and genre-based variation, and frequency and collocational analyses. Mark Davies became a Big-Name personality in the first instance thanks to the intelligent user-friendly universal design and annotation of the high-demand corpora he created, including BYU-BNC [British National Corpus, 100 million words, 1980s-1993], COCA [Corpus of Contemporary American English, 450 million words, 1990-2012], COHA [Corpus of Historical American English, 400 million words, 1810-2009], SOAP [Corpus of American Soap Operas,100 million words, 2001-2012], TIME [Time Magazine Corpus,100 million words, 1923-2006]. In most cases, the corpora creation involved collecting the texts, editing and annotating them, creating the corpus architecture, and designing and programming the web interfaces. Some other people and organisations also contributed to the latest improvements, including Paul Rayson who provided the CLAWS tagger for the COCA, COHA, and the TIME corpora. Many BYU students helped to scan novels, magazines, and non-fiction books, and process and correct the files and lexicon for the COCA and COLA corpora. As for the BYU-BNC, the original texts were licensed for re-use from Oxford University Press. The British National Corpus has other interfaces as well, but Mark Davies believes that his corpus architecture and interface allows for speed, size, annotation, and a range of queries that is unmatched with other architectures. It is also free, in addition to that.

Here we are talking about exploring the British National Corpus both in hindsight and in the light of recent improvements. It might be particularly timely for young researchers who are not entirely in the know of all the advantages, privileges and benefits of BYU-BNC. The corpus linguistics approach (and

the BNC in particular) seems to be drawing attention of the eastern European researchers and university faculty over the last years: new projects are being launched, new researches are being performed, new national corpora appear (Национальный корпус русского языка, Корпус української мови, Narodowy Korpus Języka Polskiego, Český národní korpus, Slovenský národný korpus, Hrvatski nacionalni korpus, Български национален корпус, etc.)

There are thousands of more influential corpora and corpora of less studied languages in the world now. The British National Corpus, 'the first and best-known national corpus' [Xiao 2008: 384], was initiated in 1980s under the management of the BNC consortium, and the project was finished by 1994. The consortium was led by Oxford University Press, together with two other major dictionary publishers Longman and Chambers, as well as the research centres at the Universities of Lancaster and Oxford, and the British Library. Today the BNC is being widely used by lexicographers to create dictionaries, by linguists to describe the English language, by language and translation teachers and students to teach and learn the language –to name but a few of its applications.

Good old words

The British National Corpus is a collection of 4124 samples of modern British English from a wide range of sources. They are not selected for enjoyable reading, they are documents that should properly be regarded as contexts for research, samples of original written texts, of size between 40 and 50 thousand words, cut out of entire texts for two main reasons – of length and of copyright. Samples of spoken texts have been also partially transcribed in most cases. The BNC comprises approximately 100 million words of written texts (90%) and transcripts of speech (10%) of modern British English. It has been estimated that the BNC would take 4 years to read aloud, at 8 hours a day. The overall size of the BNC corresponds to roughly 10 years of linguistic experience of the average speaker in terms of quantity [Aston & Burnard 1998: 28].

The BNC is not a random collection of texts. Written texts were selected according to three criteria: '*domain*', '*time*' and '*medium*'. Domain refers to the content type of the text (i.e. subject field); time refers to the period of text production, and medium refers to the type of text publication such as books, periodical or unpublished manuscripts (Xiao 2008: 384). The table below summarizes the distribution of these criteria [Aston & Burnard 1998: 28–30]:

Tab.1. **Composition of the written BNC**

Domain %	***Date %***	***Medium %***
Imaginative 21.91	1960-1974 2.26	Book 58.58
Arts 8.08	1975-1993 89.23	Periodical 31.08
Belief and thought 3.40	Unclassified 8.49	Misc. published 4.38
Commerce/Finance 7.93		Misc. unpublished 4.00
Leisure 11.13		To-be-spoken 1.52
Natural/pure science 4.18		Unclassified 0.40
Applied science 8.21		
Social science 14.80		
World affairs 18.39		
Unclassified 1.93		

The transcribed spoken material was collected on the basis of two criteria: *'demographic component'* and *'context-governed component'*. The demographic component is composed of informal encounters recorded by 124 volunteers selected by age group, sex, social class and geographic region, while the context-governed component consists of more formal encounters such as meetings, lectures and radio broadcasts in four broad context categories. Table 2 summarizes the distribution of these criteria [Xiao 2008: 385]. At large, the sole source of the BNC spoken data is the orthographic transcription of the corpus. The BNC original sound recording is not available for general use. The entire set of recordings is lodged in the National Sound Archive in the UK [McEnery 2003: 451]

Tab.2. **Composition of the spoken BNC**

Region %	***Interaction type %***	***Context-governed %***
South 45.61	Monologue 18.64	Educational/Informative 20.56
Midlands 23.33	Dialogue 74.87	Business 21.47
North 25.43	Unclassified 6.48	Institutional 21.86
Unclassified 5.61		Leisure 23.71
		Unclassified 12.38

Here anything goes

The first questions the corpus first-timers may ask about this huge data-base look trivial: *What can I get out of it? What can I do with the BNC?* A number of scholars would note in favour of corpora use.

1) "… in studying corpora we observe a stream of creative energy that is awesome in its wide applicability, its subtlety and its flexibility" [Sinclair 2004: 1].

2) "… by using a corpus, the linguist can investigate more material and get more exact calculations of frequencies. The results from corpora are usually presented in one of two ways: as a concordance or as frequency figures" [Lindquist 2009: 5].

3) Corpora can give you three different kinds of data: (1) empirical support for hypothesis as the corpus linguistics approach enables a good recall of relevant examples and acts as a qualitative method of corpus exploitation; (2) frequency information for words, phrases or constructions that can be used for quantitative studies to show similarities and differences between different groups of speakers or different kinds of texts, to provide frequency data for psycholinguistic research, etc.; (3) extra-linguistic information (or meta-data) on such factors as the age or gender of the speaker/writer, text genre, temporal and spatial information about the origin of the text, etc. [Lüdeling & Kytö 2008: ix].

4) … Until recently, it has been unfeasible to analyze the full range of texts, registers, and linguistic characteristics required for comprehensive analysis of register variation. With the availability of large on-line text corpora and computational analytical tools, such multi-dimensional analyses have become possible [Biber 2008: 823].

5) "Without corpus tools in the form of concordances, word frequency counters, and collocate profilers, many of the actions that we usually complete in seconds would take years of work" [Anthony 2009: 87].

6) …a corpus aims "for *balance and representativeness* within a specific sampling frame, in order to allow a particular variety of language to be studied and modeled" [McEnery 2003: 449].

7) "… indirectly, corpora can help with decisions about what to teach and when to teach it", directly, they can assist in the teaching process [Römer 2008: 113].

8) Corpus allows you "to adopt a principle of total accountability, retrieving all the occurrences of a particular word or structure in the corpus for inspection" [Aston & Burnard 1998: 6].

9) "… From parallel corpora we can extract a larger variety of translation equivalents embedded in their context, which makes them unambiguous" [Teubert & Čermáková 2004: 123].

10) "… *corpus* has become less of a buzzword and more of a necessary, acknowledged reference source for students, linguists, language professionals (teachers, translators, technical writers, lexicographers etc.). As a consequence, discovery learning is now a workable option for many teachers, that can easily be adapted and made to appeal to most students, not necessarily very advanced ones or language specialists" [Bernardini 2004: 32].

In general, almost any kind of computer-based research on the nature of the language is possible in a format BYU-BNC. The BYU-BNC annotation is intelligible and user-friendly, it answers McEreny's set of advantages of corpus annotation which includes ease of exploitation, reusability, multifunctionality, and explicit analyses [McEnery 2003: 454]. The first-timers may also ask a question about annotation. In essence, Tony McEnery argues, corpus annotation is the enrichment of a corpus in order to aid the process of corpus exploitation. It does not necessarily occur from the viewpoint of the expert human analyst – corpus annotation only makes explicit what is implicit, it does not introduce new information (ibid: 453). In detail, different interfaces to the BNC enables a researcher (1) to look for different linguistic constructions, with *'part of speech'* tags; (2) to query syntax: the BNC provides variable length syntactic searches such as noun phrases, relative clauses, etc.; (3) to find the most frequent collocates for a given word, which often provides useful insight into word meaning and usage; (4) to study frequency in five genres – spoken, fiction, magazines, newspapers, academic – and four time periods since 1990; (5) to compare the collocates of two words (or lemmas); (6) to find the frequency and distribution of the synonyms of a word, and see

which synonyms are used more in different registers or historical periods; (7) to do word comparisons: it shows which collocates occur with *Word 1* but not *Word 2* and vice versa. Moreover, by modifying the values of the options, researchers can create their own queries. A range of application areas for scholars, teachers and learners include lexicography, natural language understanding systems, and all the branches of applied and theoretical linguistics. Corpus evidence can be also found for verifying hypothesis on each linguistic level from speech sounds to entire conversations or texts [Lüdeling & Kytö 2008: ix].To provide inspiration for new-comers, we could have a thorough look at some of these areas. A wide variety of linguistic information can be introduced by concordances and frequency counts. G. Aston and L. Burnard [Aston & Burnard 1998: 7-10] serve plenty of food for thought on the three levels -- lexis, morphosyntax, and semantics/ pragmatics - in "The BNC Handbook", which was written for previously used SARA software though.

Lexis

- How often does a particular word-form appear in the corpus? The mean frequency is approximately 150, but the standard deviation of the mean is very high (over 11,000), indicating that there are very many words with frequencies far removed from the mean.
- With what meaning is a particular word-form, or group of forms, used? Is *'back'* more frequently used with reference to a part of the body or a direction? Do we *'start'* and *'begin'* the same sorts of things?
- How often does a particular word-form appear to other, which *collocate* with it within a given distance? Does *'immemor*ial' always have *'time'* as a collocate? Is it more common for prices to *'rise'* or to *'increase"*?
- How often does a particular word-form appear in particular grammar structures, which *colligate* with it? Is it more common to *'start to do something'* or to *'start doing it"*?
- How often does a particular word-form appear in certain semantic environment, showing a tendency to have positive or negative connotations? Does the intensifier *'totally'* always modify verbs and adjectives with a negative meaning, such as *'fail'* or *'ridiculous"*?
- How often does a particular word-form appear in a particular type of text, or in a particular type of speaker and author´s language? Is *'little'* or *'small'* more common in conversations? Do low-class speakers use more (and different) expletives?
- Whereabouts in texts does a particular word tend to occur? And is it in fact true that *'and'* never begins a sentence?

Morphosyntax

- How frequent is a particular morphological form or grammatical structure? How much more common are clauses with active than with passive main verb?
- With what meaning is a particular structure used? Is there a difference between *'I hope that'* and *'I hope to"*?
- How often does a particular structure occur with particular collocates or colligates? Is *'if I was you'* or *'if I were you'* more common?
- How often does a particular structure appear in a particular type of text, or in a particular type of speaker and author's language? Are passives more common in scientific texts? Is the subjunctive used less by younger speakers?
- Whereabouts in texts does a particular structure tend to occur? Do writers and speakers tend to switch from the passive tense to the 'historic present' at particular points in narratives?

Semantics or pragmatics

- What tools are most frequently referred to in the texts talking about gardening?
- What fields of metaphor are employed in economic discourse?
- Do the upper-middle classes talk differently about universities from the working classes?
- How do people close conversations, or open lectures? How do chair-persons switch from one point to another in meetings?
- Are pauses in conversations more common between utterances than within them?
- What happens when conversationalists stop laughing?

Not all of these types of information are equally easy to obtain. To disambiguate homographs or to identify particular uses of words or structures, it is necessary to inspect the lines in the output, classifying them individually. It is relatively easy to calculate the frequency of a word-form or of its collocates. It may be more difficult to calculate its frequency of use as a particular part of speech, with a particular sense, or in a particular position or particular kind of text. To help in such tasks, BYU-BNC is increasingly marked up with a detailed encoding which encompasses both external characteristics of each text and its production, and internal characteristics such as its formal structure. Newcomers may try the opportunities of *SEARCH SYNTAX* options, to start with.

Within BYU corpus family, all these linguistic areas can also be examined contrastively, comparing data from different corpora, different historical periods, dialects or geographical varieties, modes (spoken or written), genres, and registers. By comparing corpora collected 20 years ago with an analogous

corpus of today, it is possible to investigate recent changes in English. By comparing corpora collected in different parts of the world, it is possible to investigate differences between, for instance, contemporary British and American English.

Three linguistic case studies

The choice of the language material to be used in a linguistic study depends upon the research questions a researcher asks. Below, three Russian language users-driven case studies in the format *trigger situation + problem + study + discussion* are presented. They are in no way claim to provide an exhaustive treatment of the linguistic problems mentioned but can be considered as aimed at familiarizing corpus newcomers with a small portion of the BNC linguistic content and BYU-BNC software.

1) *'To my mind'* case study

Trigger situation: About fifteen years ago, two scholars from Great Britain came to Krasnoyarsk, a Russian city in the centre of Siberia, to provide a five-day methodological seminar on Business English for the regional faculty of so called Presidential Programme on Training Managers for Enterprises of National Economy of the Russian Federation. The seminar was a success, and by the end of the fifth day the methodologists left a couple of hours to discuss the topical issues. And then one of the participants asked a question about the language changes and the English language the seminar participants spoke. "You speak perfectly well, - the answer was. – We really feel at home. We mean it. The only thing we can if we may, say is that you a bit overuse of *'to my mind'* phrase. *'To my mind'* is a little out of date. There are other simpler discourse markers like *'I think'* or *'in my opinion'*. Feel free to better use them." Many years passed, but still *'to my mind'* in wide use among Russian English-speakers, it is in the textbooks, and in the young generation's written and spoken discourses.

Problem: frequency as evidence for language shift or change. All of us have been observing linguistic change in progress around us. Sometimes we suffer from its effects, which may "range from petty inconveniencies to crushing disabilities that can consume years of our lives with unrewarding struggle against hopeless odds" [Labov 2006: 4]. The problem of linguistic change of discourse markers use does not carry heavy consequences but still is worth considering. In general, a discourse marker is a word or phrase that helps to link certain ideas. As a rule, these words are more formal lexical items that find little use in speech – which is perhaps why they do not always come naturally. They help to create a clear structure by acting as a kind of 'linguistic signposts' that contribute to well-constructed arguments. They provide a sense of clarity, coherence, fluency and logic to, in most cases again, a piece of writing. For finding evidence about *'to my mind'* use, four discourse markers from the category "expressing attitudes" have been chosen. They are basically applied to express somebody's opinions: *'to my mind', 'I think', 'in my opinion', 'it seems to me'*.

Study: For the analysis, we applied to the BNC and four other BYU Mark Davies's corpora . We analysed all the registers in all the corpora, and additionally, spoken corpora in the BNC – BNC (S) and the COCA – COCA (S). The results (Tab. 3) show that by now *to my mind* is the least discourse marker among the four mentioned. In spoken discourse, *to my mind* is 516 times less occurred in the BNC compared with *I think* (50 vs. 25,825) and 1,415(!!!) less occurred in the COCA (181 vs. 256,156).

Tab.3. **Frequency of four "expressing attitudes" discourse markers in BYU corpora**

CORPORA	*TO MY MIND*	*I THINK*	*IN MY OPINION*	*IT SEEMS TO ME*
BNC	264	40,971	554	826
BNC (S)	50	25,825	78	329
COCA	640	339,086	2,544	4,982
COCA (S)	181	256,156	1,242	3,341
COHA	1,562	92,928	1,962	5,201
TIME	99	8,465	342	451
SOAP	38	131959	229	664

Discussion: The discussion on meaning shift might be built from various perspectives:

a) historical pragmatics, particularly in sense of 'external' and 'internal' language change [Traugott 2004: 539];

b) competing forces of speaker's economy vs. hearer's economy (*I think* is shorter than *to my mind*) [Horn 1996: 313];

c) morphosyntax: reanalysis and analogy [Harris & Campbell 1995]; etc.

2) *'Language communication'* case study

Trigger situation: Naming is a challenging and time-consuming process. Whoever or whatever people name – a new-born baby, a new street, a company or a programme – they carefully study the essence, choicely discuss the options, and finally, delightedly present the end product, which is the name. Six years ago, on the basis of four major Krasnoyarsk institutions of higher education, Siberian Federal University (Krasnoyarsk, Russia) was founded. One of its nineteen new-made institutes gathered philology, linguistics, languages, and journalism beneath the same root. It was given the name of the Institute of Philology and Language Communication (Institut Filologii i Yazikovoi Kommunikasii, in Russian). The faculty were very much gone with the excitement and hurry, and finally this name was democratically voted for, and juridically approved as well. Within the time from then on, the hosts themselves sometimes feel the awkwardness of the name, and many university visitors asked and checked back again and again about the *Language Communication*.

Problem: words and their company. Corpora give us opportunity not to get lost in opinions, but see the typical. If anything exists in the corpus, it exists in the language. Both words – *'language'* and *'communication*' at first sight belong to the same sphere – society, the way people deal with each other, make decisions, etc. To some extent, a language is a tool to build communication. How closely are they associated? Do they collocate?

Study: The word *'communication'* belongs to very frequent words, and among the BYU corpora we encounter:

- 6,042 tokens in the BNC (100 million words);
- 21,851 tokens in the COCA (450 million words);
- 11,522 tokens in the COHA (400 million words); interestingly, that for the period of 1810-1820 it occurred 38 times, whereas in 1990-2000 the number of occurrences reached its maximum of 802;
- 1,771 tokens in the TIME (100 million words);
- 417 tokens in the SOAP (100 million words).

If we focus only on academic discourse, we will get 1,601 *'communication'* tokens in the BNC.

Within the BYU corpora, the word *'language'* shows:

- 18,515 tokens in the BNC;
- 60, 173 tokens in the COCA;
- 43,468 tokens in the COHA;
- 8,105 tokens in the TIME;
- 810 tokens in the SOAP.

The frequency of *'language'* exceeds the former twofold (for the SOAP), threefold (for the BNC and COCA), fourfold (for the COHA) and almost fivefold for the TIME. Now let us see how these two words (*'language'* and *'communication'*) go together. The entire set of BYU corpora counted only 6 tokens: 2 times in the COCA and 4 occurrences in the BNC, including just three in the latter for the academic discourse. If we study the contexts, we will see that two of them belong to *sign* language communication, which is far from our context:

1) A comparison of the UK and USA also indicates a difference in function and use of fingerspelling. In both countries fingerspelling has become incorporated into sign language communication and deaf people will use fingerspelling with one another [BNC. Source information: Sign language. Woll, B. and Kyle, J. G. Cambridge: CUP, 1993].

2) If on close inspection, educators can begin to understand the great importance of these supposed' gestures' in portraying the syntax and semantics of sign language communication, then a more effective view of the children's needs in language will emerge [BNC, idem.].

The third context gives us hope that the words collocate anyway, and that *language communication* is not a 'non-existent object' [Carlson 2004: 95]:

There is a whole range of language communication, particularly that which involves the interrelation between speaker and hearer, which cannot be fitted into this conceptual view of semantics [BNC. Source information: Style in fiction. Short, Michael H. and Leech, Geoffrey N. Harlow: Longman Group UK Ltd, 1987].

Considered all, we could confirm the only occurrence of *Language Communication* as a phrase within BYU BNC corpus and three occurrences in the whole family of BYU corpora.

Discussion:

The problem of phrases could be well-discussed in terms of descriptive and prescriptive grammar: there are many opinions that linguists should retain an objectively descriptive stance and base their theories on observed behaviour in the speech community and not on structures and not on structures that some people prefer on ground of aesthetics or faulty logic.

3) *'An arm is a hand'* case study

Trigger situation: An English student learning the Russian language in the Russian university was confused with the use of the Russian equivalents for *an arm*/ *a hand* words and their collocations. The point is that, contrastingly to the English Language which has two major words to describe a human upper extremity – *an arm* (either of the upper limbs from the shoulder to the wrist) and *a hand* (the prehensile part of the body at the end of the arm), the Russian lan-

guage makes do with one word – *рука* [ruká] (both meanings are represented by the same word).

… your brother refused to raise his **arm** [BNC].

[…tvoi brat otkazalsya podnyat' **ruku.**] (Russian accusative case for **ruká**, here **an arm**)

I wanted to shake his **hand** [BNC].

[Ya hotel pozhat' emu **ruku**.] (Russian accusative case for **[ruká]**, here **a hand**)

Problem: corpora, lexical meanings and lexis discovery. These complicated cross-language lexical issues could be researched through exploring the lexical items and their behavior in advanced queries of the 'word comparisons' option in the SEARCH SYTAX section in BYU-BNC.

Study: Word comparisons are a variation on the CONTEXT searches in BYU-BNC. This query allowed us to enter two words – *hand* and *arm* in [1] and [2], and then compare the adjective collocates – j* for [3] that occurred with each of these words. A comparison of the collocates provided insight into differences in meaning between the words. These sets of collocations are presented in comparison in Tab. 4 and Tab. 5.

Tab.4. **Collocations of 'adjective + *hand (W1)*' in comparison with '*adjective + arm (W2)*'**

WORD 1 (W1): **HAND** (3.88)

	WORD	W1	W2	W1/W2	SCORE
1	OTHER	5525	56	98.7	25.4
2	LITTLE	31	4	7.8	2.0
3	FREE	249	36	6.9	1.8
4	LEADING	20	3	6.7	1.7
5	RIGHT	1093	221	4.9	1.3
6	HUMAN	28	6	4.7	1.2
7	BONY	14	3	4.7	1.2
8	BROWN	13	3	4.3	1.1
9	LEFT	883	215	4.1	1.1
10	OUTSTRETCHED	76	25	3.0	0.8
11	EXTENDED	15	5	3.0	0.8
12	STRONG	56	30	1.9	0.5
13	GOOD	34	19	1.8	0.5
14	WHOLE	12	8	1.5	0.4
15	THIN	10	7	1.4	0.4
16	UPPER	124	88	1.4	0.4
17	FRONT	12	16	0.8	0.2

The BNC, like most corpora, contains large quantities of figurative language, *hand* and *arm* collocations notably contribute to this collection. The first line of Table 4 shows a hundred-fold preponderance of *'other hand'* (W1) over *'other arm'* (W2). It might be explained by the high frequency of *'on the other hand'*, but requires a thorough study. In the Russian language, *рука* [ruká] does not carry this meaning at all. A closer look at lines 3 and 5 allows to suppose high *free hand* frequency due to metaphorical meanings of *free hand* (total freedom, carte blanche):

He's given me a **free hand** to buy horses [BNC].

Mr McCloy gave me a **free hand** to go where I chose, bar the two big outhouses [BNC].

The Russian *рука* [ruká] is not appropriate in both cases either.

Line 5 represents a five-time advantage occurrence of *'right hand'* over *'right arm'*. A fleet glance shows *'right hand'* metaphor- excessiveness ('indispensable assistant') as well:

as Phil's **right hand** man I felt I could cope with most situations [BNC].

… this time it's Eddie Lawson's old **right hand** man Kel Carruthers [BNC].

In the Russian language, the same connotational metaphoric meaning is carried by similar phrase – *правая рука* [pravay'a ruka]:

Now Resin, Luzhkov's *right hand* ([**prava'ya ruká**]; [**ruká**] for **the hand**) became a high powered person in Moscow [Russian National Corpus].

Tab.5. **Collocations of 'adjective + *arm (W2)*' in comparison with '*adjective* + *hand (W1)*'**
WORD 2 (W2): **ARM** (0.26)

	WORD	W2	W1	W2/W1	SCORE
1	LONG	50	5	10.0	38.8
2	BROKEN	61	7	8.7	33.8
3	BARE	14	3	4.7	18.1
4	SHORT	12	5	2.4	9.3
5	HEMIPLEGIC	12	6	2.0	7.8
6	INJURED	12	8	1.5	5.8
7	FRONT	16	12	1.3	5.2
8	UPPER	88	124	0.7	2.8
9	GOOD	19	34	0.6	2.2
10	STRONG	30	56	0.5	2.1
11	OUTSTRETCHED	25	76	0.3	1.3
12	LEFT	215	883	0.2	0.9
13	RIGHT	221	1093	0.2	0.8
14	FREE	36	249	0.1	0.6
15	OTHER	56	5525	0.0	0.0

Figurative use can serve as reasoning for higher frequency of 'long arm' (power, strength) as well:

The **long arm** of the law is reaching a bit too close for comfort if you're taken short in sunny Singapore [BNC].

To compare: In Mexico, Stalin's *long arm* ([**dlinna'ya ruká**]; [**ruká**] for **the arm**) reached him [Russian National Corpus].

Discussion: It is up to the researcher to decide how the data should be interpreted. A number of conclusions might be drawn from these tables, concerning particularly lexical meanings, metaphor exploration, teaching and learning collocations. BNC also can provide translators with way of identifying the differences and of formulating and testing hypothesis to appropriate translation strategies. And, systematic studies of metaphor and metonymy, R. Moon notes [Moon 2012:204], may start by searching for a specific item, such as a metaphor-rich word like *heart*. Words *hand* and *arm* also have a lot of metaphorical traces in the BCN.

Conclusion

Corpora in general and the British National Corpus in particular, are not a universal panacea for language researchers of any rank, including student level. Nevertheless, a closer look at the BYU-BNC architecture and design reveals plenty of opportunities for multiple applications. The tools of the BNC corpus analysis can be applied to lexis on the whole, multi-word units, grammar, registers and genres, texts and discourses, semantic and pragmatic issues, language change, translation studies and translation as such, and not to know what all. Microsoft SQL Server as the backbone of the relational database approach makes the BYU-BNC available in any remote location. Even complex queries take one or two seconds. Our main aim was to highlight the advances that corpora carry for smaller or bigger linguistic research, which can only serve to strengthen the relationship between corpus linguistics and other branches of the same tree.

Literature

Anthony, L. Issues in the Design and Development of Software Tools for Corpus Studies: The Case for Collaboration. In Baker, P. (ed.), *Contemporary Corpus Linguistics*. London; New York: Continuum, 2009. P.87–104.

Aston, G. & Burnard, L. The BNC Handbook: Exploring the British National Corpus with SARA. UK: Edinburg University Press, 1998. 256 p.

Bernardini, S. Corpora in the classroom. In: Sinclair, J. McH. (ed.), How to Use Corpora in Language Teaching. John Benjamins Publishing Company, 2004. P.15–36.

Biber, D. Multi-dimensional approaches. In: Anke Lüdeling, A./ Kytö, M. (eds.), *Corpus Linguistics: An International Handbook*. Berlin; New York: Walter de Gruyter GmbH & Co, 2008. P.822–855.

Carlson, G. Reference. In: Horn, L.R. & Ward, G. (eds.), The Handbook of Pragmatics. Blackwell Publishing, 2004. P.74–96.

Davies, Mark. The Corpus of Contemporary American English: 450 million words, 1990-present. 2008–. URL: http://corpus.byu.edu/coca/ (дата обращения: 24.09.2012).

Davies, Mark. The Corpus of Historical American English: 400 million words, 1810-2009. 2010–. URL: http://corpus.byu.edu/coha/. (дата обращения: 08.10.2012).

Davies, Mark. TIME Magazine Corpus: 100 million words, 1920s–2000s. 2007–. URL: http://corpus.byu.edu/time/. (дата обращения: 16.10.2012).

Davies, Mark. BYU-BNC. (Based on the British National Corpus from Oxford University Press). 2004–. URL: http://corpus.byu.edu/bnc/. (дата обращения: 09.10.2012).

Harris, A.C. & Campbell, L. Historical Syntax in Cross-Linguistic Perspective. Cambridge: Cambridge University Press, 488 p.

Horn, L. R. Presupposition and Implicature. In Lappin, S. (ed.), The Handbook of Contemporary Semantic Theory. Oxford: Blackwell, 1996. P.299–320.

Labov, W. Principles of Linguistic Change: Social Factors. UK: Blackwell Publishers, 2006. 572 p.

Lindquist, H. Corpus Linguistics and the Description of English. UK: Edinburgh University Press, 2009. 219 p.

Lüdeling, A. & Kytö, M. (eds.), Corpus Linguistics: An International Handbook. Berlin, New York: Walter de Gruyter GmbH & Co, 1353 p.

McEnery, T. Corpus Linguistics. In: Mitkov, R. (ed.), The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press, 2003. pp. 448–463.

Moon, R. What can a corpus tell us about lexis? In: O'Keeffe, A. & McCarthy, M. (eds.), The Routledge Handbook of Corpus Linguistics. UK: Routledge, 2012. P.197–211.

Moskovkin, L. Illiterate Oak Tree. 2009. URL: http://leo-mosk.livejournal.com/138230.html. (дата обращения: 30.10.2012).

Römer, U. Corpora and language teaching. In: Anke Lüdeling, A. & Kytö, M. (eds.), Corpus Linguistics: An International Handbook. Berlin, New York: Walter de Gruyter GmbH & Co, 2008. P.112–131.

Sinclair, J. Introduction. In: Sinclair, J. (ed.), How to Use Corpora in Language Teaching. John Benjamins Publishing Company, 2004. P.1–10.

Teubert, W. & Čermáková, A. (). Directions in corpus linguistics. In: Lexicology and Corpus Linguistics. London, New York: Continuum, 2004. P.113–165.

Traugott, E.C. Historical Pragmatics. In: The Handbook of Pragmatics. Blackwell Publishers, 2004. P.538–561.

Xiao, R. Well-known and influential corpora. In: Anke Lüdeling, A. & Kytö, M. (eds.), Corpus Linguistics: An International Handbook. Berlin; New York: Walter de Gruyter GmbH & Co, 2008. P.383–457.

НАЗАД В БУДУЩЕЕ: БРИТАНСКИЙ НАЦИОНАЛЬНЫЙ КОРПУС

Валентина Анатольевна Кононова
к. пед. н., доцент кафедры лингвистики и межкультурной коммуникации
Сибирский федеральный университет

Статья «Назад в будущее: Британский национальный корпус» посвящена возможностям использования BNC для студенческого лингвистического исследования при написании студенческих научных сочинений, а также для изучения английского языка на реальных примерах использования его в жанровом и стилистическом разнообразии. Британский национальный корпус – проект завершенный, он был создан в Великобритании в 1990-е гг. специалистами-лексикографами; это один из лучших, крупнейших и наиболее известных корпусов в мире, своего рода эталон. Статья апеллирует к BNC нового формата, представленного в 2012 г. американским лингвистом Марком Дейвисом. Возможности использования языковых корпусов не должны недооцениваться: корпус является очень эффективным ресурсом и инструментом для исследования не только узкими специалистами, но и студентами. Статья написана в рамках постдокторского исследования проекта ERANET MUNDUS, университет Барселоны, Испания.

Ключевые слова: Британский национальный корпус; BNC-BYU; корпусная лингвистика; Джон Синклер; Марк Дейвис; корпусная разметка.