Role of the Brown Corpus in the History of Corpus Linguistics

Olga KHOLKOVSKAIA[1]

[1] *Dept. of Economics, Management and Humanities, Czech Technical University, Zikova 4, 166 27 Praha, Czech Republic*

olga.kholkovskaya@gmail.com

Abstract. *The history of corpus linguistics as a field of science is relatively short. Commonly agreed that the first modern computerized corpus was created between 1961 and 64 at Brown University by two linguists Henry Kučera and Nelson Francis. It is no exaggeration to say that the Brown Corpus shaped the modern corpus studies. It served as the model not only for new English language, but also for all modern national corpora, and is still used as the dataset in different kinds of research.*

Keywords

the Brown corpus, corpus linguistics, history of corpus linguistics, language corpus.

1. Introduction

Corpus analyses as an approach to investigating language-related aspects in different spheres of science became wide-spread during last three decades. Corpus linguistics is the discipline which studies natural languages with the help of electronic language corpora and deals with methodological questions of design, development and implementation of such corpora [1].

Corpus, the core concept in this field, is generally a large enough, structured, and annotated database of language samples, stored and accessed electronically. Structure, volume, scope and overall design of a particular corpus depends on its purpose. They can be universal or special, single- or multilanguage, synchronic or diachronic etc. For example, Czech National Corpus (ČNK, Český národní korpus) [7] is a continuous project that creates and supports a number of corpora covering different aspects of the Czech language, both written and spoken: series of synchronic corpora SYN2000, SYN2005, SYN2010 and the most recent SYN2015, all containing around 100 million words of contemporary written Czech, diachronic written corpus covering texts from the 2nd half of 13th century to the middle of the 20th, and two corpora of transcribed oral speech [4,7].

Rapid development of computer science and IT industry during the last decades of the 20th century made it possible to create, store and access really large collections of linguistic data, which, in its turn, inevitably led to interdisciplinary cooperation between language and computer science. But first such datasets existed even earlier. In pre-computer era corpus was a collection of index cards (or dictionary slips) with text sample and annotation. Such manually prepared and searched collections had a number of intrinsic problems. Sampling and preparation was time-consuming and expensive, they were prone to errors, and data was sparse. It was difficult to add some new annotation later, difficult to search, or statistically assess the gathered data [3,5].

Transition to electronic form radicaly changed the way the corpus is prepared, stored and used. It gave researchers the opportnity not only to create much bigger corpora and use more complex method to work with them. It also added replicability and representativeness, significantly reduced the number of mistakes, and solved, at least partly, the sparseness issue. During the first stage the corpora was relatively small, if compared to current ones, and equipped with less sophisticated software, but these pioneers defined the future development of corpus linguistics and became the prototype for next generation of language corpora [4].

2. The Brown Corpus

Many sources states that the first electronic corpus, in the modern sense, was Brown University Standard Corpus of Present-Day American English, commonly known as the Brown corpus [2,17,1,4]. It is a synchronic corpus of contemporary written prose, printed in the United States in 1961. The Brown corpus was prepared in 1961-1964 by Nelson Francis and Henry Kučera, linguists at Brown University [8, 6].

2.1 Before the Brown Corpus

The Brown corpus was in many ways pioneering, but it still had predecessors in technical and conceptual terms. Transition of French Thesaurus (Trésor de la Langue Française, TLF) to electronic form started before the Brown corpus was planned [1,6]. The Rand corpus, developed by the Rand Corporation group of Machine Translation in Los Angeles, was available for researchers already in 1959 [18, 6].

Fig. 1. Henry Kučera. From http://www.brownalumnimagazine.com/content/view/2573/40/

The strongest source of influence for the Brown corpus was the Survey of English Usage (SEU) created at the University College London. Two groups were closely connected. Randolf Quirk, who decided to devise SEU (planned to be machine readable from the beginning), during his scholarship in the United States met Freeman Twaddell, American Germanist, who created departments of linguistics and Slavic studies at Brown University in Providence (Rhode Island) and invited first Kučera, then Nelson, to join these departments and his corpus project. Nelson during his scholarship worked at SEU in London, and Quirk was at the original conference, where major decisions about the Brown corpus were made [6,8].

But SEU was computerized only in the end of 1980s, TFL wasn't available for the public, and Rand was built on physics and math texts. So, the Brown corpus was first representative general purpose corpus, carefully designed, sampled, and prepared. Moreover, it immediately became available to any researcher, who asked for it. It opened a new era in corpus linguistics, and for years became the standard in many ways, and the most cited resource in the field [6, 2, 18].

Fig. 2. W. Nelson Francis.Picture by By Jfrancis51 (Own work)

The Corpus was recorded on the tape with density 556 characters per inch in BCD code, standard binary coded decimal seven-bit code, in which each character is represented by a six-bit code and one check bit, and on nine-channel tape with density of 800 characters per inch in the IBM System/360 Extended Binary Coded Decimal Interchange Code (EDCDIC).

The most part of data processing was done on the IBM 7070 Data Processing System at the Brown University Computing Laboratory. Programs for this computer were written mainly in the two available 7070 assembly programming languages, Autocoder 74 and Autocoder 76. The statistical data was obtained with the help of several FORTRAN programs. IBM 1401 was used to solve smaller problems. Some statistical calculations were also performed on the IBM System/360, which had its own assembly language [11].

The tagged version (Form C), which became available later, used the set of 82 tags, including part-of-speech tags, punctuation, and inflectional morphemes, which could be combined in compound tags. So, in grammatical tagging, as it was called then, the Brown team also was the first [9, 10, 12, 17].

2.2 The impact of the Brown corpus

In 1967 Kučera and Francis published their classical work Computational Analysis of Present-Day American English. The book contains very little text by Kučera and Francis themselves and a lot of data. The main part of the book constitutes two frequency lists, by rank and by alphabetical order, and frequency distribution lists. It also includes two articles by other researchers (analysis of word-frequency distribution by John B. Carroll, study of sentence-length distribution by Mary L. Marckworth and Laura M. Bell). By the time the book came out several

researches, based on the corpus, were already finished, and several more were in progress [9, 11].

During next decades a number of other corpora arose. The first corpus of spoken English (the London-Lund Corpus of Spoken English, LLC) was developed in 1975. Another collaboration between British and Scandinavian universities resulted in the Lancaster-Oslo-Bergen Corpus of British English (LOB) three years later [6, 14]. The Lancaster Corpus followed the design and sampling practice of the Brown Corpus. And all of them shared methodological grounds of the Brown corpus. These two corpora, in their turn, became the model for other comparable corpora of the so called 'Brown Family': seven corpora covering the period between 1931 and 2006 [17].

Mentioned earlier Form C, tagged version of the Brown corpus, became the start point for many works in the field. PARTS, Program for part of speech tagging, developed by Kenneth Ward Church at Bell Laboratories

used the Brown corpus for training to get probabilities estimates[13]. The elaborated versions of the Brown tagset were used in LOB and LLC, as well the corpus itself was used for training purposes [12, 15, 17]. The first deep parsed corpus, the Penn Treebank, used the Brown corpus for development of training set, and its tagset was based on that of the Brown corpus [12]. All of them preserved comparability with the Brown corpus. In five years, 1986-90, as Leech states, 89 institutions acquired copies of the tagged LOB Corpus [2].

Corpus-based approach became a new powerful method in different subfields of linguistics and computational linguistics: machine translation and translation studies

in general, information retrieval, studies of language acquisition and psycholinguistics. The latter also adopted the measure of word frequency, supposed by Francis and Kučera. For 50 years it became the norm for a wide variety of studies of human memory and word processing by human brain [14, 16, 21].

The Brown corpus has not lost its actuality even in 21st century. It still serves as training and testing set in different fields of science from natural language processing (parsing, disambiguation, spell-checking) [24, 19, 20, 22] to biomedicine [25]. It is included as one of the corpora in NLTK (Natural Language Toolkit), a leading platform for building Python programs for work with language data [26]. In 2006 Google released a trillion-word corpus including frequency counts for n-grams up to five words long. One of participants of that project, Peter Norvig, Brown graduate, remembers how excited he was about access to the Brown corpus during his undergraduate studies [23].

3. Conclusion

The Brown corpus, even though it has predecessors and sources of inspiration, in many senses was the first one. It was the first representative electronic language corpus, it was carefully designed and sampled, and it became available to other researchers. It defined the principles of corpus design and gave rise to corpus-driven approach, which is currently used in many different fields of science.

Fig. 3. IBM 7070

Acknowledgements

Study described in the paper was supervised by Ing. Jan Mikeš, Ph.D., FEE CTU in Prague.

About author

Olga KHOLKOVSKAIA born in Russia, graduated from Petrozavodsk University, philological department, worked in ABBYY software group as linguist, currently student of Czech Technical University, Prague.

References

[1] KENNEDY, GRAEME D. *An Introduction to Corpus Linguistics*. 1998.

[2] LEECH, G. The State of the Art in Corpus Linguistics. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, ed. by Aijmer Karin & Bengt Altenberg, London & New York, Longman, 1991, p. 8-29. Available at: http://ccl.pku.edu.cn/doubtfire/CorpusLinguistics/Introduction/The%20state%20of%20the%20art%20in%20corpus%20linguistics.htm

[3] ČERMÁK, F. Korpusová lingvistika: stručný historický přehled. Available at: https://ucnk.ff.cuni.cz/doc/korp_lingv_prehled.rtf.

[4] ČERMÁK, F., SCHMIEDTOVÁ, V. Český národní korpus – základní charakteristika a širší souvislosti. In *Národní knihovna*, 2004, roč. 15, č. 3, s. 152-168. Available at: http://full.nkp.cz/nkkr/nkkr0403/0403152.html

[5] ČERMÁK, F. Language corpora: The Czech Case. Available at: www.researchgate.net/.../226459248_Language_Corpora_The_Czech_Case

[6] LÉON, J. Claimed and unclaimed sources of corpus linguistics. In *Henry Sweet Society Bulletin*, May 2005, Issue 44, p. 36-50.

[7] Czech National Corpus. Available at: http://ucnk.ff.cuni.cz

[8]. NELSON, F. W., KUČERA, H. Brown corpus manual. 1979. Available at: http://www.hit.uib.no/icame/brown/bcm.html

[9] MAVERICK, G. V. Review. In *International Journal of American Linguistics*, Vol. 35, No. 1, 1969, p. 71-75.

[10] CHARNIAK, E., CARROLL, G., ADCOCK, J., CASSANDRA, A., GOTOH, Y., KATZ, J., LITTMAN, M., MCCANN, J. Taggers for parsers. In *Artificial Intelligence*, Volume 85, Issues 1–2, August 1996, p. 45–57. Available at: http://www.sciencedirect.com/science/article/pii/0004370295001085

[11] NELSON, F. W., KUČERA, H. *Computational Analysis of Present-Day American English*. Providence: Brown University Press, 1967.

[12] MARCUS, M. P., MARCINKIEWICZ, M. A., SANTORINI, B. Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics*, v.19, n.2, 1993, p. 313-330. Available at: http://ucrel.lancs.ac.uk/acl/J/J93/J93-2004.pdf

[13] CHURCH, K. W. A Stochastic Parts program and nounphrase parser for unrestricted text. In *ANLC '88 Proceedings of the second conference on Applied natural language processing*, p. 136-143. Available at: http://www.aclweb.org/anthology/A88-1019.pdf

[14] LEECH, G. 100 million words of English: the British National Corpus. In *Language Research*, 28(1): 1–13, 1992. Available at: http://s-space.snu.ac.kr/bitstream/10371/85959/1/1.%202235197.pdf

[15] LEECH, G., GARSIDE, R., ATWELL, E. The automatic grammatical tagging of the LOB Corpus. IN *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, 7.13-33. Available at: http://eprints.whiterose.ac.uk/81848/1/TheAutomaticGrammaticalTaggingLOBCorpus.pdf

[16] GILQUIN, G., GRIES, S. TH. Corpora and experimental methods: A state-of-the-art review. In *Corpus Linguistics and LinguisticTheory*, 5–1,2009, 1–26. Available at: http://www.linguistics.ucsb.edu/faculty/stgries/research/2009_GG-STG_CorpExpMeth_CLLT.pdf

[17] LEECH, G. The development of ICAME and the Brown family of corpora. Available at: https://bells.uib.no/bells/article/download/358/373.

[18] HUTCHINS, W. J. Machine translation: a concise history. In J*ournal of Translation Studies*, vol.13, nos.1-2 (2010). Special issue: The teaching of computer-aided translation. Ed. Chan Sin Wai, Chinese University of Hong Kong, 2010, p.29-70. Available at: www.hutchinsweb.me.uk/CUHK-2006.pdf

[19] FAN, J., BARKER, K., PORTER, B. Automatic interpretation of loosely encoded input. In *Artificial Intelligence*, 173, 2009, p. 197–220. Available at: http://www.sciencedirect.com/science/article/pii/S0004370208001434

[20] SHARMA, S., GUPTA, S. A correction model for real-word errors. In *Procedia Computer Science*, 70, 2015, p. 99 – 106. Available at: http://www.sciencedirect.com/science/article/pii/S1877050915032111

[21] BRYSBAERT, M., NEW, B. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. In *Behavior Research Methods*, 2009, 41 (4), p. 977-990. Available at: https://www.researchgate.net/publication/38075028_Moving_beyond_Kucera_and_Francis_A_Critical_Evaluation_of_Current_Word_Frequency_Norms_and_the_Introduction_of_a_New_and_Improved_Word_Frequency_Measure_for_American_English

[22] LEE, J., GOLDSMITH, J. Linguistica 5: Unsupervised learning of linguistic structure. In *HLT-NAACL Demos*, 2016. Available at: http://people.cs.uchicago.edu/~jagoldsm/Papers/lxa5.pdf.

[23] HALEVY, A., NORVIG, P. ,PEREIRA, F. The unreasonable effectiveness of data. In *IEEE Intelligent Systems*, v.24, n.2, p. 8-12, March 2009. Available at: http://research.google.com/pubs/archive/35179.pdf.

[24] RIEZLER, S., KING, T. H., KAPLAN, R. M., CROUCH, R., MAXWELL III, J. T., JOHNSON, M. Parsing the Wall Street Journal using a lexical-functional grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July 2002, p. 271-278. Available at: https://pdfs.semanticscholar.org/fc57/c5fa3a8a14fd100645db1dfe65aa62d6966f.pdf.

[25] AOUICHA, M. B., TAIEB, M. A. H. Computing semantic similarity between biomedical concepts using newinformation content approach. In *Journal of Biomedical Informatics*, 59, 2016, p. 258–275. Available at: https://www.ncbi.nlm.nih.gov/pubmed/26707454

[26] BIRD, S., KLEIN, E., LOPER, E. Analyzingt text with the Natural Language Toolkit. Available at: http://www.nltk.org/book/