Building and analysing a large general corpus: Exploring the Written BNC2014 with #LancsBox

A half-day practical workshop in a computer room

Vaclav Brezina, Carmen Dayrell, Mathew Gillings, Abi Hawtin, Tony McEnery, Andressa Rodrigues Gomide, Matt Timperley and Isolde van Dorst

Lancaster University

Introduction

Building a large general corpus is a major undertaking. It involves careful planning and requires the availability of the infrastructure and resources that allow successful completion of the project. Many useful lessons can be learnt from previous projects (the original BNC 1994, the Czech National Corpus, COCA, the Australian National Corpus etc.), including the American National Corpus (ANC), which originally aspired to comprise the same amount of data (100 million tokens) as the British National Corpus and cover a wide range of registers (Ide 2008: 108); currently the ANC v. 2, released in 2005, includes 22 million running words. Apart from theoretical challenges (see below) the major practical challenge for a general corpus, as acutely experienced by ANC compilers, is the availability of the required data that covers a range of genres/registers, a challenge that corpus compilers have been confronted with from the early days of the field (Leech 2014). From the theoretical perspective a general synchronic corpus is designed to represent a broad variety of language uses of a given speech community at a particular moment in time (McEnery & Hardie 2011). The British National Corpus 2014 (BNC2014) thus reflects the many varied uses of British English around the mid sampling point of 2014. In addition, it seeks to be a comparable corpus to the original BNC 1994. This leads to a potential tension between representativeness and comparability (see Leech & Smith: 2005) that needed to be resolved. When reviewing the multiple challenges connected with building a large general corpus such as the Written BNC 2014, we may consider the building of a corpus a daunting task to sample a dynamic and complex phenomenon of language. Leech (1992: 12) remarks that “no corpus-building enterprise is ever complete or completely satisfactory”. Yet, this enterprise is also essential for the field because it provides a unique opportunity to reflect on the state of a language, in our case British English.

Description of the dataset

BNC2014 is a major project led by Lancaster University to create a 100-million-word corpus 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.

Currently, the first stage of the project has been completed with the Spoken BNC2014 (Love et al. 2017) released. The Spoken BNC2014 is available Lancaster’s CQPweb (Hardie 2012), BNClab (Brezina et al. 2018), a brand-new sociolinguistic tool, and is also possible to download
The second stage involves creating a written counterpart to the Spoken BNC2014: The Written BNC2014.

The Written BNC2014 is being compiled by a team of researchers at the ESRC Centre for Corpus Approaches to Social Science (CASS), Lancaster University led by Vaclav Brezina and Tony McEnery. The sampling frame was proposed by Abi Hawtin, based on her doctoral research and the sampling frame of the original British National Corpus (BNC1994). The project has been supported by ESRC grants no. EP/P001559/1, ES/K002155/1 and ES/R008906/1.

Aims of the workshop

The main aim of the practical workshop is to introduce the participants to the written British National Corpus 2014 (BNC2014) and provide a unique early insight into the corpus prior to the public access which is planned for the autumn of 2019. We will provide early access to a balanced sample of BNC2014 data, BNC2014 Baby +. BNC2014 Baby + is a mirror corpus to BNC1994 Baby (Burnard 2003), a four-million-word sample of British English from the early 1990s. In addition to traditional genres (academic writing, fiction, newspapers and conversation), BNC2014 Baby + includes data from the e-language subcorpus of the Written BNC2014. The corpus will be analysed using #LancsBox (Brezina et al. 2015), a new corpus analysis package developed at Lancaster University. The practical introduction to BNC2014 and #LancsBox will highlight the innovative features of the new tool and demonstrate its use with current British English data.

The participants will be given exercise sheets that will guide them thorough the major features of the Written BNC2014. They will be able to compare British English from the early 1990s (BNC1994 Baby) with the current use of English (BNC2014 Baby +). In addition, theoretical issues related to corpus sampling and representativeness will be discussed.

Technical requirements: a computer lab with a screen and a data projector. No specialized software is required. The computers need to be able to run external programs. Pre-testing of the #LancsBox tool (http://corpora.lancs.ac.uk/lancsbox/download.php) in the lab will be necessary to ensure smooth running of the workshop.

Call for participation:

Building and analysing a large general corpus: Exploring the Written BNC2014 with #LancsBox

In this workshop, we introduce the Written British National Corpus 2014 (Written BNC2014), a new corpus compiled at Lancaster University and #LancsBox v. 4.1, a new software package for the analysis of language data and corpora, also developed at Lancaster University. The participants of the workshop will be given access to BNC2014 Baby +, a five-million-word balanced sample of the BNC2014 which includes both traditional (academic writing, fiction, newspapers and conversation) and new (e-language) genres. Participants will be guided through the process of corpus compilation and analysis using interactive exercises that demonstrate different stages of the process. We will also outline the theoretical
considerations involved in building a large general corpus and invite participants to engage in a critical discussion about corpus representativeness and sampling.

In particular, the participants will learn:

- About the Written BNC 2014 project and decisions involved in the corpus compilation.
- Search and filter BNC2014 Baby + according to different categories.
- Compare BNC2014 Baby + with different corpora including BNC1994 Baby from the 1990s.
- Apply a variety of statistical measures.
- Visualize data.

The workshop is open to anyone interested in innovative exploration of language using computational tools. It does not presuppose any technical or statistical knowledge.

References


