Chapter 4
Corpora and Concordancers

Charles Hall
The University of Memphis, USA

ABSTRACT

At the heart of almost all ANLP is the corpus. This chapter provides an overview of the history and development of the corpus and crucial criteria that define the modern corpus. It ends with a discussion of the most basic analytical tool for corpus linguistics, the concordancer.

INTRODUCTION

In the simplest terms, a traditional corpus is just a collection of authentic texts containing language a researcher wishes to examine. Corpora (the traditional plural of corpus) can contain a hundred million words, as in the Corpus del Español. On the other hand, some specialized corpora are composed of just a few words. For example, Stubbs (1996) compiled and researched a corpus of the 880 words in two letters from Lord Baden-Powell, the founder of the Boy Scouts.

In the field of ANLP, most corpora are usually quite large but well defined and finite; they also need to meet specific criteria that will be discussed later in this chapter. However, the amorphous and "infinite" Internet can also serve as a type of corpus for quick and dirty work, such as in Hall & Lee (2006) who show non-native teachers of English easy techniques to use commercial search engines, such as Google, to explore language structure, lexical distribution, and syntactic differences in World Englishes.

Most contemporary corpora are different from the simplistic "collection of texts" in that corpora are usually selected to represent a genre (e.g. the Corpus of Late Eighteenth-Century Prose) or a language variety (e.g., the Corpus of Spoken American English). There are no limits to the subject matter of a corpus, and it could reflect written, spoken, or even signed language. Although there are corpora of spoken work, most corpora are based on written work because transcribing recordings is both time consuming and expensive.

DOI: 10.4018/978-1-61350-447-5.ch004

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Corpora can consist of texts in one or more languages. Because corpora are now being used to help understand and perform translations, there are many new configurations that involve two or more source and target languages. For example, Xiao and Yue (2009) have built a corpus that is all in Chinese but contains two sub-corpora: one consisting of Chinese novels and the second of English novels translated into Chinese; likewise, bilingual corpora can contain the same materials in two languages or different content in two languages but obtained through the same sampling method (Xiao & Yue, 2009).

Because ANLP researchers routinely use corpora as the data for their analysis, it is important to begin with a basic look at what corpora are. From there, we can turn to the criteria for corpus construction that are almost standard in ANLP. Finally, we can examine the basic tool used to analyze corpora, concordancing programs (usually called *concordancers*).

BACKGROUND

In 2011 the Centre for English Corpus Linguistics (CECL) at the University of Louvain (Belgium), will celebrate its 20th anniversary and is among the oldest modern institutions dealing with corpus work. However, the history of corpus research and concordancers actually predates that center by almost 800 years.

One of the first research tools developed for corpus work was the written concordance. In contrast to an index that normally only lists important topics or names, a written concordance lists the location of every occurrence of every word in the corpus. The first concordance recorded was of the Vulgate Bible completed in the 13th century by Dominicans (www.catholic.org/encyclopedia/view.php?id=3229). Although their original purpose was not linguistic, Bible concordances were later essential in the 19th century efforts at authorship identification issues in Genesis, for example. Researchers were able to use lexical means to support the documentary hypothesis (Wellhausen, 1905) that there were several different writers for first books of Bible.

In the 19th century, individuals would spend many years of their lives preparing written concordances of the works of individual authors, such as the complete concordance of Milton's work by Cleveland (1867). These works could be used to investigate language patterns by other scholars; however, these printed concordances were both unwieldy and subject to human error in compilation. Indeed, these last two factors were crucial in limiting the use of corpus research. Before corpus linguistics could become widespread and accepted, two events were essential to the growth of contemporary corpus linguistics: the development of the computer and an awareness of the need for empirical data in language analysis.

From the end of the 19th century, many researchers conducted types of corpus analysis, all done by hand. For example, in 1897 Käding published an analysis of frequency accounts of German spelling using an eleven-million word corpus he had collected, coded and analyzed by hand (Baker, 2006). Even Thorndike's famous lists of words for teachers (1932) were collected in part from manually prepared corpora.

According to Leech (1992) the American structuralists of the 1940s and 1950s thought that, "A corpus of authentically occurring discourse was the thing that the linguist was meant to be studying." McEnery and Wilson (2009) point out that these structuralist works were usually based on two presuppositions that most linguists would now agree are incorrect: 1) There is a finite number of sentences in a language, and 2) these sentences can be collected. As a result, corpora were the only source of evidence for linguistic inquiry. Consequently, the pendulum was set to swing from this pure empiricism to the almost pure rationalism introduced by Chomsky and with it a wholesale discrediting of corpus work.

8 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/corpora-concordancers/64579

Related Content

Comparison between Internal and External DSLs via RubyTL and Gra2MoL
Jesús Sánchez Cuadrado, Javier Luis Cánovas Izquierdoand Jesús García Molina (2014). *Computational Linguistics: Concepts, Methodologies, Tools, and Applications (pp. 816-838).*
www.irma-international.org/chapter/comparison-between-internal-and-external-dsls-via-rubytl-and-gra2mol/108753

Machine Translation within Commercial Companies
Tomáš Hudík (2015). *Modern Computational Models of Semantic Discovery in Natural Language (pp. 256-272).*
www.irma-international.org/chapter/machine-translation-within-commercial-companies/133882

Digital Watermarking Techniques for Audio and Speech Signals
Aparna Gurijalaand John R. Deller Jr. (2007). *Advances in Audio and Speech Signal Processing: Technologies and Applications (pp. 132-160).*
www.irma-international.org/chapter/digital-watermarking-techniques-audio-speech/4685

Synthetic Speech Perception in Individuals with Intellectual and Communicative Disabilities
Rajinder Kouland James Dembowski (2010). *Computer Synthesized Speech Technologies: Tools for Aiding Impairment (pp. 177-187).*
www.irma-international.org/chapter/synthetic-speech-perception-individuals-intellectual/40865

Dynamic Effects of Repeating a Timed Writing Task in Two EFL University Courses: Multi-Element Text Analysis with Coh-Metrix
Kyoko Babaand Ryo Nitta (2012). *Applied Natural Language Processing: Identification, Investigation and Resolution (pp. 398-413).*
www.irma-international.org/chapter/dynamic-effects-repeating-timed-writing/61061