EVALUATING CHINESE WEB-AS-CORPUS: SOME METHODOLOGICAL CONSIDERATIONS

Shu-Kai Hsieh[*]
National Taiwan University

ABSTRACT

Corpus development in the context of Web has become one of the most important issues due to its tremendous size, geographic and social range, up-to-datedness, multimodality and wide availability at minimal cost, etc. Many Web-as-Corpus (WaC) construction tools are made freely available as well. However, due to its intricate orthography, in this paper, I will argue that a sound methodology for evaluating newly emerging Chinese WaC resources is needed urgently. There has been a wide range of possible usages of the Web for corpus construction, as well as the measures for the comparison of traditional corpus and web corpus. Basically, main approaches include acquiring web content and processing it into a static corpus (WaC, Web-as-Corpus), and accessing it directly as a dynamic corpus (WfC, Web-for-Corpus). I will introduce our works in constructing twWaC (Taiwan Web as Corpus[1]) at National Taiwan University, with the explanation of problems encountered. Two statistic measures from the distributional point of view will be proposed to illustrate the difference of scaled twWaC and ASBC (Academia Sinica Balanced Corpus).[2]

KEYWORDS

Web corpus **C**hinese corpus **S**egmentation **C**orpora comparison

[*] Corresponding author: shukaihsieh@ntu.edu.tw

Linguistic Corpus and Corpus Linguistics in the Chinese Context
Journal of Chinese Linguistics Monograph Series 25 (2015): 15-29

2409-2878/2015/25-0001

漢語網路語料庫評測方法初探

謝舒凱
國立台灣大學

提要
方興未艾的網路發展中，不斷湧現的巨量語言使用資料，伴隨著地理、社會、多模態與多層脈絡等後設資料，建構與處理工具的可得性的提高等種種背景因素，使得語料庫語言學進入了一個前所未有的局面。「網路即語料庫」的想法因而應運而生。近年來，利用網路資料建構語料庫有許多種作法。除了利用或自創搜尋引擎來動態地擷取網路資料供語言研究使用之外，大部分的作法，是在一定的設計之下，利用工具蒐集與下載網路資料，處理並標記語言訊息之後，提供研究者重複使用該語料庫。本文要討論的是，網路語料庫在漢語的脈絡下，迫切地需要嚴謹的評測方法。中文沒有詞的邊界訊息，在處理上向來需要先做分詞的程序。傳統小規模語料庫在機器自動分詞之後，經由人工校正之後可以得到一定品質的保證。但是在大規模的巨量網路語料庫，機器分詞錯誤比率因而倍生，但人工校正的可能性卻因而降低，造成基本計量上的不確定，也連帶影響後續的語言處理與分析工作。本文以我們所建構的台灣噗浪網路社群語料庫，和中研院平衡語料庫與其他語料庫的對比出發，利用詞彙豐富度與涵蓋率等分布統計，計量上說明了大規模語料造成的問題，希望能引出日後更多的研究課題。

關鍵詞
網路語料庫　**現**代漢語語料庫　**中**文分詞　**語**料庫比較