A Data-driven Approach to the Mental Lexicon: Two Studies on Chinese Corpus Linguistics

Chu-Ren Huang, Kathleen Ahrens, and Keh-jiann Chen

In this paper, we attempt to show i) that corpora offer real instances of language use (production) in a non-controlled environment, ii) that corpora constitute of a large sampling of the real input to linguistic perception, and iii) that corpora extracted from mass media represent the shared linguistic information of the language-speaking community.

Corpus-based studies are studies of linguistic theories based on linguistic objects (instead of on non-linguistic acts like naming, picture pointing, story-telling, or making decisions on yes-no questions.) We use two corpus-based studies to show that they can complement the traditional psychology-oriented studies based on controlled experiments. The two studies shed important light on the psychological reality of the notion of a word in the mental lexicon.

Our first study examines the definition of compounds based on M.I. (mutual information) values extracted from a corpus. We show that this empirically based definition of compounds easily resolves the previous controversies involving intuitive judgements (e.g. Bates et al. 1992 and 1993, and Zhou et al. 1993).

The second study involves the complex cognitive process of *suo1xie3* (abbreviation) and a simple statistical model. We show that while a rule-based model can only capture incomplete aspects of Chinese abbreviation, corpus-based statistical values nicely reflect their status in the mental lexicon.

In conclusion, we argue that corpora reflect shared uses of language and are efficient tools for establishing baseline facts in (psycho-/neuro-)linguistic research.

Keywords: Mental lexicon, Corpus, Word, Mutual information, Abbreviation

由語料出發驗證心理詞庫——漢語語料庫語言學研究二例

黃居仁 ? 安可思 ? 陳克健

中央研究院語言所 ? 國立台灣大學外文系 ? 中央研究院資訊所

本文試圖由語料著手來探索語言之心理真實性。傳統研究是以實驗為依據。這類心理或腦神經語言學研究雖然得到了不少突破。但仍有其限制。首先實驗室迫使受試者在受控制的非自然環境中使用語言；其次實驗的設計往往只限於少數幾個句子；最後限於受試者注意力的限制，實驗語句限制長度而缺乏自然的上下文語境。本文認為大量語料除可補足上述實驗方法之不足，且可表現出語言的心理真實性。

以語料庫探索心理真實性的前提有三：一、語料庫提供了在自然環境下語言使用（生成）的實例。二、語料庫正好也代表了日常語言辨識對象的大量取樣。三、適當抽取的語料正可以呈現使用該語言的人所共有的語法知識。

文中討論了兩個研究，這兩個研究均是根據中央研究院現代漢語語料庫為基礎。第一個研究探討中文的複合詞，第二個研究探討中文特殊的構詞現象——「縮寫」。這兩個研究都支持了一個基本假設——即「詞」這個觀念在漢語的心理詞彙庫中的確存在而且可以利用語料庫資料判讀。也就是說語料庫反映了語言的心理現象，可提供了我們由資料入手研究語言真實性的另一蹊徑。

關鍵詞：心理詞庫 詞 語料庫 互見訊息