INTRODUCTION

About half a century ago two major developments took place relevant to research on language. The field of linguistics saw the ascendency of the School of Generative Linguistics which focused linguistic investigations on the "ideal speaker" through the language generated by such a speaker. It was a paradigmatic change from the School of Taxonomic Linguistics which focused on generalization of linguistic data within traditional and strict divisions of language, such as phonology, morphology and grammar. At the same time the field of computer based investigation of language emerged from cryptography and went on to process massive textual data. Cross-fertilization between the two fields was not fruitful until recently in part because of the limitations of computational power and methodology at that time, and of the difficulties beyond serious armchairs deliberations in extending generalization from an ideal-speaker's language to cover that of an entire community of speakers.

In the subsequent decades, increasing realizations have surfaced that for generalizations on language to be ultimately meaningful they must go beyond individual introspection to embrace the communal norms on the full use of language in the relevant community of speakers. Also the all-important Darwinian mutant variations should not be overlooked because language exists in a dynamic continuum in time and always embodies the seeds for germination of new traits within the relevant social and cultural ecology. In parallel to these developments there have been rapid advancements in rigorous modeling methodology in the natural sciences, which are also found to be applicable to cover large datasets of natural language. Thus the stage is set for the new era of corpus cultivation and corpus linguistics.

Linguistic Corpus and Corpus Linguistics in the Chinese Context
Journal of Chinese Linguistics Monograph Series 25 (2015): 1-8

2409-2878/2015/25-Introduction

A fundamental function of linguistic corpora is the provision of quantitative information on the distribution of words or other linguistic artefacts to facilitate qualitative judgments. In recent years 'corpus' or 'corpora' has probably become a very frequent term itself among articles found in different areas of language studies. The volume *Quantitative and Computational Studies on the Chinese Language* published back in 1998 has marked a critical period of time when corpora have significantly gained attention and recognition in Chinese linguistics and language processing. The development of new Chinese corpora has never ceased, and the use of old and new corpora in linguistic investigations and applications has been on the increase since then. This trend has been evident in international gatherings such as the International Conference on Language Resources and Evaluation (LREC) and the Asian Language Resources (ALR) workshops, and more recently the workshops on Building and Using Comparable Corpora (BUCC), in addition to numerous others on various sub-areas of Chinese linguistics and computational linguistics. The trend will certainly continue in the future. More importantly, we are entering an era of 'big data', where voluminous text (or speech or even multi-media) materials are readily available especially from the web. It has brought convenience and has simultaneously brought necessary but often unrealized changes in corpus cultivation and data curation. It has not only given rise to new challenges in corpus-based methods but also considerably expanded their applicability for different purposes, as well as broadened intellectual horizons. The stage is also set for the desirable convergence in the study of language between the perspective of the abstract stereotypical speaker at a point in time and the broader perspective involving the continuum of communal linguistic and related developments in space and time.

This monograph consists of selected revised papers from the Roundtable Conference on Linguistic Corpus and Corpus Linguistics in the Chinese Context, held in May 2011 at the Hong Kong Institute of Education. The papers, presented in four sections, give a timely account of the latest status of Chinese corpus development and of their use in a wide range of linguistic studies and applications in this fast-moving information age. Some new developments since the 1998 volume are foregrounded.

Part I of the monograph contains discussions on the use of Chinese

corpora for various linguistic investigations, as well as their computational processing and applications in different settings.

'Web as Corpus' is probably a trend which is difficult to resist for different languages alike. Despite the scalability and cost-effectiveness in corpus cultivation from web data which is virtually unlimited in terms of variety and amount, the resulting data must nevertheless be evaluated with care and treated with caution in terms of the efficacy of the curation efforts and the suitability of their applications. **Hsieh** explicates the considerations necessitated by this trendy practice, illustrated with a comparison between a newly constructed web corpus based on microblog texts and some traditional corpora with respect to variations in frequency spectrum and lexical coverage.

Studies with more specialized corpora are diversely exemplified in this volume. **Cheung and Yang** are particularly concerned with the representativeness of a child language corpus especially with individual differences and contextual variations taken into account. They compare the lexical diversity, noun type and verb type indices among samples from children of different age groups. Normal children and those with specific language impairment were compared, based on spontaneous conversational samples and elicited narratives.

Children's stories constitute a specific text genre. **Kwong** compiled a bilingual corpus based on different published versions of the Aesop's Fables in English and Chinese, with annotations done at the structural, semantic and emotional level. Various surface linguistic properties including the use of sentiment-bearing words, dialogues and ending strategies, were explored for their contribution to the realization of the morals and subsequently to story understanding.

With a corpus of classical Chinese poems, **Fang et al.** demonstrated the usefulness of ontological annotation in addition to traditional bag-of-word features for automatic authorship attribution. They observed a difference in the distribution of imagery descriptors in the poems by Liu Yong and Su Shi, and tested various feature sets for their automatic classification.

On more traditional corpus-based linguistic investigation, the comparative nature of corpus linguistics remains central. **Xiao** proposed wide-angle corpus-based contrastive studies for comparing and

contrasting very distinctly different languages. Such an approach can serve as a common platform for corpus linguistics, contrastive studies, and even translation studies, as long as issues like cross-language grammatical terminology and corpus comparability can be satisfactorily handled. He presented major outcomes for comparing English and Chinese with respect to passive constructions and classifiers.

Examining evidence from concurrent and early corpus data of Cantonese and Hakka, **Kataoka** studied the grammaticalization process of a Chinese progressive marker and argued that its grammaticalization follows different development paths in the two dialects.

Part II of the monograph contains discussions on all-time issues of Chinese corpus processing, including word segmentation and part-of-speech (POS) tagging, as well as the extraction of compound words and grammatical relations.

Chan et al. discussed a system employing recursive means of predicting the POS tags of Chinese words based on morpheme properties and word neighbors in raw text. Out-of-vocabulary words were first subject to a similarity-based module for predicting their POS with respect to known words in the database, relying on morphemic, radical, phonetic and collocational features to reveal the syntactic contexts of the words.

Sun and Sun put forward the idea of using lexicalized statistical pattern matching as a strategy for judging the termhood of compound words. By devising a set of structural templates and capitalizing on the web as a huge corpus, statistics obtained from the hits returned by existing search engines may supplement conventional mutual information for analysis at various linguistic levels.

Similarly concerned with compounds, **Chung and Chen** classified morphemes into various semantic types. Based on their morpho-syntactic behavior and logical compatibility, semantic composition rules were devised to form an automatic scheme for predicting the POS and senses of compounds.

Huang et al. tackled the automatic extraction of grammatical relations from corpora, capitalizing on the word sketch function available from the Sketch Engine. Considering the properties of Chinese, they enriched the originally simple sketch grammar with more fine-grained lexico-grammatical knowledge, which demonstrated itself to be an

important and effective move.

Chen compiled a lexico-syntactic knowledge base for analyzing and deriving kinship relations, which starts with a set of templates observed from corpus data covering various forms and correspondence of kinship relations in different structural patterns. Given a set of possibly incomplete relations as premises, they undergo a series of logical operations and simplification in a reasoning model to resolve the unknown relation.

As the number of corpora grows and the variety expands, resource sharing and interoperability become a critical issue, partially determining the influence and popularity of any given corpus. A general solution is to adopt a standardized protocol for corpus markup. International practice has demonstrated the effectiveness of XML as a data markup language. Based on this, **Fu and Zhang** presented a schema designed for marking up written Chinese corpora. The schema accommodates meta-data, such as the original document structures and typographic features, as well as linguistic annotations at various levels of analysis. Important data in the form of components and relations are analogously captured by means of the elements and attributes in XML. In addition to annotation, the schema is expected to offer a common platform to facilitate document management and data processing.

Part III of the monograph contains five papers showcasing the continuous work in Chinese corpus development, spanning traditional general corpora and a variety of others in specialized domains.

The Peking University (PKU) Corpus is inarguably one of the earliest and most influential endeavors in the history of Chinese corpus development. After almost three decades since its inception, the corpus has continued to move toward multi-level annotation beyond being anchored in the People's Daily. Together with the many lexical resources, tools and applications derived from the corpus, they collectively constitute the Comprehensive Language Knowledge Base. **Duan et al.** have the full story to share.

In view of the cost in corpus annotation, optimizing the use of existing resources is often a more attractive solution than starting from scratch. **Huang and Song** demonstrated the feasibility of constructing a structurally annotated corpus with respect to Combinatory Categorial Grammar, by means of a 'translation' of an existing treebank based on

Phrase Structure Grammar, relying on a verb-subcategorization algorithm and a set of pre-defined Chinese sentence patterns.

Language variety has always been a core consideration in corpus design and compilation. Despite many observable regional variations, most Chinese written corpora capture the main characteristics of Modern Standard Chinese, while the issue of language variety is much more salient when spoken corpora are concerned. **Luke and Wong** described the design and compilation of a spoken corpus consisting of everyday conversations between Cantonese speakers in Hong Kong. Though small in size, the naturally occurring conversations captured in the corpus are believed to faithfully reflect the structure of a language, and this work adds to existing corpora of the kind which mostly comprise transcriptions of radio programmes.

In addition to monolingual corpora, bilingual and multi-lingual corpora are particularly important resources typically for machine translation and cross-lingual information retrieval. Parallel corpora are nevertheless both uncommon and expensive to obtain. **Lu et al.** demonstrated the combined use of conventional text alignment approaches, including those based on length, lexical items and translation models, for effective screening and extraction of good quality parallel sentence pairs from bilingual and multi-lingual comparable patent documents, to alleviate the bottleneck in the acquisition of adequate parallel data involving Chinese, English and Japanese.

On a more specific domain and genre, **Zhu et al.** compiled a parallel corpus of Chinese-Sanskrit Buddhist canons. It is one of the first serious endeavors of its kind, especially at its scale and details. The project has admirable aspirations to attempt an exhaustive comparative analysis between the translated Chinese and the original Sanskrit texts, to provide detailed annotations to the corpus data and to assess the impact of language contact on the development of Chinese.

The papers in Part IV of the monograph consider Chinese corpus linguistics from a macro perspective, relating language and society through corpus-based evidence.

With reference to the LIVAC corpus, **You** demonstrated the convergence of several lexical items among various Chinese speech communities. In addition, while neologism consisting of indigenous

Chinese words tends to be found more in northern regions, the adoption of loanwords with foreign origins is more typically found in Hong Kong.

He and Tu studied the characteristics of the language used in blogs. They reported some interesting findings from a corpus of blog texts which offer clues on the difference between famous and ordinary bloggers, as well as between male and female bloggers.

Sheng provided a descriptive account of frequent Chinese morphemes, their productiveness and the structural patterns of compounds produced from such root words.

On language development, **Chew** argued that while a synchronous corpus like LIVAC provides a good snapshot of the similarity and difference between various Chinese speech communities, expanding and extending it diachronically is essential for revealing the divergence and convergence of the Chinese language among them especially over crucial historical periods.

With a sociolinguistic approach, **Su** demonstrated that a number of phonetically adapted loanwords in Chinese commonly used in the early days are gradually replaced by their Chinese-originated counterparts, thus substantiating a claim made in earlier research with corpus-based frequency data.

In the final paper, **Tsou and Kwong** showed how LIVAC has matured from simply a synchronous corpus to become a diachronic corpus after some twenty years of cultivation. Moreover it can be seen as a collection of sequential time capsules and functions well as a monitoring corpus not only for the latitudinal analysis of linguistic phenomena among the pan-Chinese communities but also for exposing relevant and concurrent cultural and societal undercurrents over time. They also revisited the longstanding issue of 3,000 characters as the threshold for literacy with reference to more extensive and updated corpus data.

The publication of this monograph is made possible with the huge support from Prof. William Wang, Editor of the Journal of Chinese Linguistics, and the exemplary professional assistance from Ms. Yifeng Wu, which we gratefully received. We also acknowledge the publication subsidy for this monograph approved while the first editor was the founding Chiang Chen Chair Professor of Linguistics and Language Sciences at the Hong Kong Institute of Education. We thank all

contributing authors for exemplifying the many faces of Chinese corpora and Chinese corpus linguistics and for probing deeper into the many-faceted Chinese language. Their work has shown how corpus-based approaches have borne fruits in different areas of research, linguistic or otherwise, in the Chinese context, and will surely inspire many more interesting, innovative and multifarious investigations and applications in the future.

June 2015

Benjamin K. Tsou
Oi Yee Kwong

引言

約半個世紀前，生成語言學派在語言學界迅速冒起，鋒頭直蓋傳統分類語言學派。前者以人類先天語言機制及其具代表性的理想語用者為研究對象，後者則傾向把語言預設為個別獨立大小不同的單位（如音韻、形態、句法等）進行歸類研究。差不多同一時期，計算機科學界對自然語言處理開始從資訊加密或解密擴展至大規模文字處理。礙於當年計算機運算功能未臻成熟，並且所謂語言的理想標準亦未能簡單地在一整個語言群體中得到足夠的驗證，因而這兩方面的發展當時並未有交流和互相推動。

此後數十年，語言學家意識到按任何理論框架來描述語言，都必須涵蓋整個語言群體的常模，更重要的是，這些常模是動態的，因為語言行為可以隨環境而改變和進化，一些既有的特性可能會消失，亦會有新的特性相繼出現和擴展。要建立常模就要以有系統的方法大量採樣和歸納，正值統計學在自然科學研究中發揮了可觀的成效，可供語言學界借鑑，並應用於大量自然語言數據，從此為語料庫語言學的新時代揭開了序幕。

語料庫主要的功能是提供字詞成語和其他語言現象的頻率分佈等量化數據，作為研究分析的基礎。近年，「語料庫」本身就成了高頻詞，在各範疇的語言研究文獻中不停出現。1998 年出版的《漢語計量與計算研究》標誌着當時語料庫在漢語語言學和自然語言處理研究中獲得關注和認受的重要時刻。直至今天，語料庫的構建和應用有增無減，此趨勢已見於大大小小的國際學術會議（如 International Conference on Language Resources and Evaluation, Asian Language Resources Workshop, Workshop on Building and Using Comparable Corpora 等專題會議系列），亦必將持續下去。尤其是當今已是海量數據的時代，網絡流傳的大量語言文字材料，既為構建語料庫帶來極大的契機與方便，亦為傳統構建及使用方法帶來不少改變與挑戰，更

Linguistic Corpus and Corpus Linguistics in the Chinese Context
Journal of Chinese Linguistics Monograph Series 25 (2015): 9-12

2409-2878/2015/25-引言

大大開拓了學術研究和技術應用的空間。

「漢語語料庫及語料庫語言學圓桌會議」於 2011 年 5 月在香港教育學院舉行，本書收錄了會後徵集的論文，與 1998 年的會議和論文集呼應，適時紀錄了漢語語料庫一日千里的發展，並展示了最新的研究方向與成果。

全書分四部分。第一部分集中討論漢語語料庫在各類語言研究（包括語言學和計算語言學）的分析及應用。

海量數據湧現，興起了「網絡就是語料庫」的潮流，而這種語料庫更需嚴謹的評測方法以保證品質。**謝舒凱**比較網絡語料庫與傳統平衡語料庫的詞彙豐富度和涵蓋率，說明大規模語料一般面對的問題。

本書亦不乏各類專用語料庫應用的例子。**張顯達**和**楊靜琛**利用台灣兒童語料庫中的兒童說話樣本，分析幾項詞彙的量化指標，比較不同語境的說話樣本，以及正常兒童與語言障礙兒童的發展差異。

兒童故事有別於一般敍事文本，是一種獨特的體裁。**鄺蔼兒**介紹了一個新構建、附有不同層次標註的中英雙語寓言故事語料庫，並探討幾項表層語言特徵與深層意義的關係。

古典詩詞的遣詞用字與意象和風格密不可分。**方稱宇等**利用包含本體知識標註的詩詞語料庫，以機器學習技術，對柳永和蘇軾的作品進行了一系列文學風格識別的實驗，特別用以鑑定著作歸屬權。

語料庫作為語言研究工具，能為對比語言學和翻譯研究等提供共同的研究平台。**肖忠華**就是運用大型語料庫，比較了英語和漢語中的被動結構和量詞，並指出此類大跨度語言對比的難點。

同樣是對比研究，**片岡新**用了早期和現代的語料，比較粵語和客家話中進行體標記「緊」的特性，以歷時語料為佐證，指出此體貌標記在兩種方言中的語法化過程不盡相同。

第二部分關注語料處理（如分詞及詞性標註）和信息抽取等歷久不衰的課題。

陳偉光等根據對複合詞內部結構的分析，利用語素等特徵，及上文下理關係，建立基於相仿性和遞歸推理機制的詞性標註器，解決未登錄複合詞的詞性標註問題。

孫茂松和**孫如穎**提出「詞匯化模板定量匹配」的方法，借助搜索引擎返回的查詢結果，估算詞與詞的結合緊密度，或有助判斷一個詞串為自由組合還是固定組合，甚至可輔助其他層次的漢語分析。

鍾友珊和**陳克健**分析了四千多個衍生性強的詞素，發現它們的構詞與語法行為，受語意和邏輯性的影響較大，並訂定語意合成規則，用作預測各種組合的詞性和語意。

黃居仁等使用速描引擎平台的詞彙特性速描功能，結合千萬字語料庫與廣泛詞語語法，整合現有的語法信息，以期向自動抽取語法知識的目標邁進一步。

陳振宇通過從語料庫觀察調查各種表示親屬關係的名詞、動詞、句法結構等，附以語義運算式，構建了「辭彙—句法知識庫」，作為漢語親屬專家系統的一個重要部件。通過此系統和知識庫，可為不完整信息進行推理，為未知的關係提供線索。

現今語料庫的數量與種類日益增加，帶來了資源共享、相容性、互操作性等問題，這些因素也局部影響着各語料庫的用途和流通度。**傅愛平**和**張弘**以可擴充置標語言 XML 為漢語書面語語料訂立了通用描述規則，除了有助保持語料原貌以及記錄各種語言信息和說明性信息，亦可方便文本與數據管理。

第三部分報告漢語語料庫構建的最新情況，如舊有語料庫加工、各類新語料庫開發等。

在漢語語料庫的發展歷史中，北大語料庫佔了很重要的一席位，經過近三十年仍極具影響力。**段慧明等**闡述了北大團隊把以《人民日報》為基礎的語料庫發展成多級標註語料庫、從中提煉出各種語言資源和語言處理工具、直至今天建成「綜合型語言知識庫」的歷程。

為建立以組合範疇語法為本的漢語樹庫，同時避免由零開始，**黃昌寧**和**宋彥**提出採用動詞次範疇和各類句型轉換算法，將清華短語樹庫自動轉換，並展示其可行性。

陸鏡光和**王麗賢**介紹了香港粵語語料庫的設計和構建過程，有別於其他以廣播節目為主要口語材料的語料庫，此庫以日常會話為基礎，語料收集和轉寫需花更多人力，因此規模雖小，卻是難得的資源。

平行語料庫對機器翻譯和跨語言信息檢索有多重要是眾所周知的，而這方面資源短缺也是不爭的事實。**路斌等**用中、英、日文的專利文本作試驗，討論如何從可比語料中挖掘大量的高品質平行句對，可望有助建立較大規模的雙語及多語平行語料庫，用於機器翻譯的研發工作。

專用平行語料庫另一例，是**朱慶之等**構建的罕有漢梵佛典雙語標註語料庫。該項目旨在建設一個帶有詳細標註的對比分析語料庫，提供實例供對比研究，以揭示語言接觸對漢語產生的影響。

第四部分主要透過語料庫提供的客觀數據，從宏觀角度探討語言與社會的關係。

游汝杰分析了 LIVAC 共時語料庫的實例，發現各華語地區的新詞有趨同傾向，同時原創的新詞大多源起較北方地區，而香港則偏向

使用字母形式的外來詞。

何婷婷和**涂新輝**對漢語網絡媒體語言做的調查，從博客語料中觀察到一般用戶與著名博客的特點，同時發現用字用語跟博客作者性別的關聯。

盛玉麒通過語料庫定量分析，描述了當代漢語中高頻根詞相關性的知識，包括其衍生性及短語結構模式，有助辨識和判斷新詞語。

周清海認為現代漢語的發展實在已經歷了不同的階段，像 LIVAC 共時語料庫，反映現代漢語在不同華語地區的現況，對語文教學和推動華文等固然十分有用，但要更全面地了解近代漢語過渡到現代漢語的情況，還需建立歷時語料庫，增強對早期現代漢語的研究。

蘇金智在較早前一項研究中，討論過漢語外來詞變化的情況，如今通過語料庫檢驗，印證了當時的結論，顯示漢語音譯詞正逐步為漢語固有詞取代。

鄒嘉彥和**鄺藹兒**闡述了 LIVAC 在近廿年間從一個共時語料庫發展成歷時語料庫，更是可供橫向及縱向分析，探索泛華語地區語言、文化和社會演變的「追蹤語料庫」。他們亦根據當中的數據，重新審視過去以三千常用字為掃盲標準的說法。

本書得以順利出版，有賴中國語言學報主編王士元教授大力支持、吳一丰女士專業且難得的協助、以及首席編輯在香港教育學院擔任蔣震「語言科學」講座教授期間批出的出版資助，謹此致謝。最後，感謝全體作者的辛勤付出，他們在各項研究的豐碩成果與精闢觀點，擴闊了我們對漢語語料庫及語料庫語言學的視野，對這個領域日後的發展，定必起到承先啟後的關鍵作用。

鄒嘉彥 鄺藹兒
2015 年 6 月