LANGUAGE AND LINGUISTICS 6.3:505-512, 2005
2005-0-006-003-000146-1

Linguistics from a Computational Perspective— Review of *Computational Linguistics and Beyond*

Tao Gong
Chinese University of Hong Kong

Since the advent of the computer in 1945, computational research has by now become pervasive in just about all newly created as well as traditional fields. Nor has linguistics escaped this tidal surge. In fact, a pre-computer computational perspective had already been attempted when linguists examined their data utilizing scientific methods; computational linguistics was formalized after the emergence of computer science, which then catalyzed the development of the whole field of linguistics. The book under consideration here, *Computational Linguistics and Beyond*, edited by Chu-Ren Huang and Winfried Lenders as Series B in Language and Linguistics Monographs, collects and discusses some recent contributions in computational linguistics, shedding light on the many aspects of this field.

Key words: Computational Linguistics, Natural Language Processing (NLP), Computational Simulation (CS)

1. Components and features of Computational Linguistics

Computational linguistics dates back to machine translation in the 'sixties (p.1), but now includes a great variety of topics, including traditional questions of intrinsically linguistic aspect, in such areas as semantics and syntax, and modern tasks in computer-aided natural language processing and the human-machine language interface. The book under review touches upon three representative components in Computational Linguistics, i.e., *Natural Language Processing* (*NLP*), *Computational Simulation* (*CS*) of Language Evolution, and *Construction & Analysis of Language Database*. All these components are interdependent and cover major fields in computational linguistics.

Computational linguistics shares many properties with traditional linguistics, yet it has some new characteristics:

First, a key feature of computational linguistics is to solve linguistic problems with computational techniques, such as the statistical analysis of linguistic data, the cross-language feature extraction, and the knowledge-based perception and production of linguistic materials. Many new techniques have already been adopted in NLP and CS. For example, Fillmore's FrameNet adopts a hierarchical network structure, which views

linguistic features as nodes and their relationships as connections. Wang et al.'s CS models import dynamic analysis to study the lexical diffusion process. Other computational techniques, such as web structure, rule-based system, and decision-tree, pave their way in other NLP methods and CS models (e.g., Katz et al. 2002, Ke et al. 2002, Steels et al. 2002).

Second, the three components of computational linguistics covered in this book can assist linguistic research from different angles. The starting point for exploring a language may be the construction and analysis of its database. Information exchange via this language between human users and computers is possible when NLP implements some rules extracted from its database. Meanwhile, the evolutionary panorama of this language, projected by CS, may guide the collection of linguistic data, the comprehension of linguistic features, and modification of NLP methods.

Third, computational linguistics is a multi-disciplinary subject, similar to traditional linguistics. In the series' preface, Wang (p. iv) clearly states that (besides mathematics and computer science) anthropology, psychology, physics, and biology are all influential in computational linguistics. These disciplines present new records in language database, provide theoretical backgrounds in CS, and introduce new techniques in NLP. Meanwhile, the great advancements in computational linguistics require some cross-language analysis, cross-discipline research, and cross-regional collaboration. The COLING conference series, as Huang and Lenders indicate (p.4), provides a good forum for discussion and cooperation from multiple perspectives.

This review summarizes the main trends in the three components computational linguistics, which are reflected in the representative contributions collected in this book.

2. More human-like Natural Language Processing (NLP)

According to Huang and Lenders (p.6), rule-based systems are found to lack efficiency, robustness, and coverage, so computational linguists began to shift to the structural systems and import different approaches to study structure-meaning links. As exemplified in this book, Fillmore's group develops the FrameNet based on semantics-syntax interactions. It combines semantic and syntactic features within lexicons in their kernel dependency graphs (KDGs). Uszkoreit's group introduces deep linguistic processing, which considers more semantic or syntactic information and offers new opportunities for further developments in NLP. Bryant, Lin, and Ide develop a multi-layer structure of semantic web and discuss how the hybrid annotation viewpoint helps computers in natural language perception.

All the above research represents new tendencies in NLP, i.e., to use structural and connectionist views in processing linguistic materials and to consider more intrinsic

linguistic features. Original NLP methods treat language processing as matching between independent input and unrelated data stored in computers. Therefore, the heavy load in data mining operations greatly restricts the performance and efficiency of these methods. New NLP methods introduce multiple tags (e.g., semantics and syntax) into lexicons and consider vertical, horizontal or cross-range connections among these tags. This approach reduces the calculation load, and provides opportunities for developing more "human-like" methods and designing a more user-friendly human-machine interface. In addition, structural viewpoints can also be adopted in speech recognition and synthesis, in which computational linguists and engineers can introduce phonetic and phonological relations in acoustic models to develop more productive audio processing approaches.

To a certain degree, NLP's focus is to handle natural languages; it does not touch upon language origin or influences of different linguistic and nonlinguistic factors. These questions are specialties of another aspect in computational linguistics, i.e., Computational Simulation.

3. New film to record language evolution—Computational Simulation (CS)

Language evolution, as one of the oldest topics in linguistics, dates back to the corresponding exchange between Darwin and Schleicher in the nineteenth century (Wang, p. i). Reconstructing language history has long relied upon anthropological findings, but these findings are "indirect evidence", inasmuch as ancient speech has not been preserved in this language history's "film". With the aid of computational techniques, the recently developed methods of computational simulation provide an alternative way to project the evolutionary trajectories of human languages on meso- or macro-historical scales (Wang 1991); see Kirby (2002) and Gong & Wang (2005). For the problem of language evolution, theoretical argumentations alone are not sufficiently reliable, and incomplete empirical findings may not allow the original history to be established. For example, many linguistic theories of language evolution only focus on the general process and the main cause, without clear explanations of details or exceptions, or of influences of other related forces throughout the process. Meanwhile, most empirical findings or empirical research only cover limited time periods in history or touch upon some short-term phenomena.

The prolific development of CS is indicated in many collections (e.g., Cangelosi & Parasi 2001, Briscoe 2002). In the book under review, Wang et al., using multi-agent computational models, exemplify a **phase transition** (a transient change of state), emergent process of meaning-utterance mappings (primitive lexicons), and a lexical

diffusion process with snowball effect.

To evaluate CS models, we need to examine attitudes towards CS shared by both traditional and modern linguists:

1) **What is the plausibility of this "game-like" CS?** First, most CS model assumptions are plausible and supported by empirical findings and theories in linguistics and other disciplines. For instance, "follow the majority" is a plausible psychological assumption adopted in many CS models studying collective behaviors (e.g., Ke et al. 2002). Second, CS uses objective, realistic mechanisms, and follows accurate, traceable procedures to obtain replicable results. For example, some simulation results match the experimental findings in the task of artificial language perception by human subjects (e.g., Christiansen et al. 2002). Finally, most results of CS can be verified by empirical data. For instance, the lexical diffusion process demonstrated by the Wang group's models has already been traced by much linguistic data (e.g., Shen 1997).

2) **What can CS do beyond the instantiation of linguistic theories and the demonstration of certain existent phenomena?** First, due to its explicit formulation, CS can help to check rigorously whether a certain *explanandum* follows certain explanation (Christiansen & Kirby 2003). And sometimes, computational model results can guide researchers to design actual experiments for testing the explanation in-depth (e.g., Christiansen et al. 2002, Perruchet et al. 2004). Second, the exploratory computational models (Holland 2005) can predict new problems in linguistics and lead to new theories.

3) **What are the shortcomings of CS?** First, sometimes, the validity of some assumptions in CS models is questionable. For example, "the impetus to iteratively communicate" is an arbitrary assumption in Wang et al.'s models. Second, results of CS models are sensitive to parameter setting and there are chances of building in results beforehand, which reduces the plausibility of these models. A global analysis and a reasonable parameter assignment can avoid this disadvantage. Finally, not all CS models can match or get support from real data, because one CS model usually covers one aspect and/or its influence, while, the real data usually results from multiple factors. A complete replication of human language evolution must incorporate many consistent CS models, each covering certain aspects.

With these attitudes clarified, the efficient development of CS still requires interactions between modelers and non-modelers (traditional linguists). These interactions include the acceptance of CS by traditional linguists and the sharing of theoretical guidance to models. It is encouraging to see the undergoing of more and more such interactions.

4. Challenge and promise in Chinese Language Processing (CLP)

CLP offers an appropriate case study for NLP (Huang, p.187). Chinese dialects, however, in contrast with European languages, confront computational linguists with three unique hurdles (T'sou, p.189):

1) non-alphabetic writing system;

2) word segmentation and absence of word morphology;

3) flexible syntax.

Advances in computer technology have resulted in major breakthroughs for the first two problems. However, due to the flexibility of the writing system and the many phonograms in Chinese, overcoming the third hurdle requires in-depth linguistic analysis from a massive language database. In addition, acoustic information for a tone language like Chinese is mandatory when constructing and analyzing its language corpus.

Confronted with these challenges over the past two decades, computational linguists within China and without, through cross-disciplinary and cross-regional cooperation, have been working hard to construct many Chinese language databases, including Mandarin, Cantonese, minority languages, and some other dialects. Several segmentation standards of Chinese based on massive language database have been announced. The construction of a formalized Chinese database makes possible an in-depth-analysis from various perspectives—semantics, syntax, phonetics, and phonology. All these efforts provide resource materials for the design of more efficient CLP methods and boost development of the field of computational linguistics as a whole.

Computational Linguistics and Beyond traces a productive development in computational linguistics by introducing creative NLP methods, instructive CS models, and profuse language database. This book is a good introduction for the novice and a useful documentation and update for the specialist. As a student just stepping into this field, I look forward to the next volume in the series of Language and Linguistics Monographs from Academia Sinica.

References

Briscoe, Ted. (ed.) 2002. *Linguistic Evolution through Language Acquisition*. Cambridge, New York: Cambridge University Press.

Cangelosi, Angelo, and Domenico Parisi. (eds.) 2001. *Simulating the Evolution of Language.* London: Springer Verlag.

Christiansen, Morten H., and M. H. Ellefson. 2002. Linguistic adaptation without linguistic constraints: the role of sequential learning in language evolution. *The Transition to Language*, ed. by Alison Wray, 335-358. Oxford: Oxford University Press.

Christiansen, Morten H., and Simon Kirby. 2003. Language evolution: consensus and controversies. *Trends in Cognitive Sciences* 7.7:300-308.

Gong, Tao, and William S-Y. Wang. 2005. Computational modeling on language emergence: a coevolution model of lexicon, syntax and social structure. *Language and Linguistics* 6.1:1-41.

Holland, John H. 2005. Language acquisition as a complex adaptive system. *Language Acquisition, Change and Emergence: Essays in Evolutionary Linguistics*, ed. by James W. Minett and William S-Y. Wang. Hong Kong: City University of Hong Kong Press.

Katz, Boris, Jimmy Lin, and Dennis Quan. 2002. Natural language annotations for the Semantic Web. *Proceedings of the International Conference on Ontologies, Databases, and Application of Semantics* (ODBASE 2002), 1317-1331.

Ke, Jinyun, James W. Minett, Ching-Pong Au, and William S-Y. Wang. 2002. Self-organization and selection in the emergence of vocabulary. *Complexity* 7.3:41-54.

Kirby, Simon. 2002. Learning, bottlenecks and the evolution of recursive syntax. *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, ed. by Ted Briscoe, 173-205. Cambridge: Cambridge University Press.

Perruchet, Pierre, D. Michael Tyler, Nadine Galland, and Ronald Peereman. 2004. Learning nonadjacent dependencies: no need for algebraic-like computations. *Journal of Experimental Psychology: General* 133.4:573-583.

Wang, William S-Y. 1991. The three scales of diachrony. *Explorations in Language*, ed. by William S-Y. Wang, 60-71. Taipei: Pyramid.

Shen, Zhongwei. 1997. *Exploring the Dynamic Aspect of Sound Change*. Journal of Chinese Linguistics Monograph No. 11. Berkeley: Project on Linguistic Analysis, University of California.

Steels, Luc, K. Kaplan, A. McIntyre, and J. V. Looveren. 2002. Crucial factors in the origins of word-meaning. *The Transition to Language*, ed. by Alison Wray, 252-271. Oxford: Oxford University Press.

[Received 14 February 2005; revised 13 June 2005; accepted 14 June 2005]

Department of Electronic Engineering
Chinese University of Hong Kong
Hong Kong
tgong@ee.cuhk.edu.hk

語言學中的計算觀點——評《計算語言學和其他》

龔　濤
香港中文大學

繼 1945 年計算機出現後，計算化研究方法席捲了各個新興和傳統的研究領域。曾被認爲是傳統學科之一的語言學也沒有“倖免”。實際上，早在語言學家蒐集數據並對其進行比較分析時，計算方法就已經被使用了。計算語言學在計算機科學出現後得以規範化並推動了整個語言學的發展。本文將評價由黃居仁和蘭德斯編寫的語言和語言學專論之一的《計算語言學和其他》一書。該書系統介紹了計算語言學的研究現狀，並指明了其未來發展的方向。

關鍵詞：計算語言學，自然語言處理，計算機仿真