FOREWORD

This monograph contains a collection of papers presented in the Roundtable Conference on Linguistic Corpus and Corpus Linguistics in the Chinese Context, organized by the Research Centre on Linguistics and Language Information Sciences of the Hong Kong Institute of Education in 2011. It reminds us of very pleasant memories when a similar meeting was held at the City University of Hong Kong some years back in 1998. That was also a very successful meeting which, among other benefits, resulted in a very useful volume. Over these years we are hopefully a little bit wiser, and the topics that we are pursuing get a bit more mature.

My contribution in the 1998 volume was on a specific topic[1], but now I would like to make some more general remarks. To begin with, let us reflect a little bit on the term 'corpus' as in 'corpus linguistics'. Even though this term has become quite accepted these days, I actually found it rather strange sounding, when I first heard it a little while back. For instance, I would not expect someone to be working in corpus physics, or corpus psychology, or corpus history, or corpus any discipline. Presumably corpus, which comes from a Latin word meaning 'body', is used to call attention to the fact that the large bodies of data are produced and used in research. But why should any discipline not involve large bodies of data, when they are appropriate? Back in the 16th century, at the beginning of modern science, Copernicus had obtained large bodies of data on the movements of the planet. These data led to the view that our

[1] W.S-Y. Wang, "Representing Relationships among Linguistic Elements," in *Quantitative and Computational Studies on the Chinese Language*, ed. Benjamin K. T'sou, Tom B.Y. Lai, Samuel W.K. Chan, & William S-Y. Wang, (Hong Kong: Language Information Sciences Research Centre, City University of Hong Kong, 1998), 1-14.

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

2409-2878/2015/25-Foreword

planet is not at the centre of the universe, and from then on, we have gained a whole lot of new insights, fundamental insights of what are the planets in our universe. To take another more recent example: molecular genetics, for which huge corpora of DNA of humans, chimpanzees and various other species have been compiled. These corpora are already online, for scientists everywhere, to share and make use of in the worldwide effort to understand the basis of life, and to discover cures for diseases which are of genetic origins. If no one would think of calling a field 'corpus astronomy' or 'corpus genetics', why is it the most natural thing in the world to study language with large bodies of data?

In this sense, corpus linguistics is simply what mainstream linguistics should be doing, and that is how linguistics should be done. The reason for adding corpus in one's description, I think, is just a historical one. It contrasts with a style of research, in which a linguist used to sit in his study and conjure up interesting sentences, and then propose various ingenious ways of representing these sentences. This style of research has been soundly criticized by William Labov as 'monastic', suggesting that the isolation is very unhealthy like in a monastery separated from the rich and vibrant complexities of reality, i.e. languages. Indeed, with computational devices available everywhere, it will be foolish not to make significant use of corpora in linguistics.

Over the past decades, major corpora have become available for general use, especially for the study of the English language. Here we can name the British National Corpus, with a hundred million words. On a smaller scale and with different orientation, we may recall corpora underlying important lexical databases like WordNet, pioneered by George Miller at Princeton University in 1985, and FrameNet, started by Charles Fillmore at Berkeley a year after that. In 1992, the Linguistic Data Consortium (LDC) was initiated by Mark Liberman of the University of Pennsylvania, to coordinate the accumulation of materials being developed at many government and university laboratories. The LDC includes materials from a variety of other languages as well as audio visual files. For the Chinese language, many researchers in many parts of the world had been developing corpora, some synchronic, some historical, some from written texts, and some from speech samples. A major example here is the synchronous corpus LIVAC that was pioneered and

maintained by Benjamin Tsou for almost 20 years now.

Alongside these general corpora, there are also certain others developed for specialized studies, and I will mention only three categories. The first category is on language ontogeny, i.e. children's acquisition of language; the second category on language phylogeny, i.e. the evolution of language across time; and the third category on language typology, i.e. the distribution of languages across space.

The one on language ontogeny launched by Brian MacWhinney and Catherine Snow in 1984, called the Child Language Data Exchange System or CHILDES, and maintained at the Carnegie Mellon University, contains data on child language acquisition dating back to the 1970s. The success of this database is evidenced by the many published papers which cited it (over 4,000 according to Wikipedia) and the various affiliated databases subsequently developed, such as the one on Hong Kong Cantonese developed by Thomas Lee. How languages are acquired either as a mother tongue or as a foreign language is obviously a central question for linguistics. The availability of these corpora has been a crucial stimulus in recent developments in this area of research.

The second category that I will mention on specialized corpora is the one compiled at the Santa Fe Institute in New Mexico,[2] directed by Murray Gell-Mann, who is a Nobel Laureate in Physics, and was a very good friend of the late Joseph Greenberg, the linguist. Gell-Mann has always been interested in language origins. He has initiated the 'Evolution of Human Languages' project and has ambitiously dedicated it to research on the deepest tree of language families, hopefully to arrive at a single source. Joseph Greenberg proposes that all of the thousands of languages in the world can be grouped under this source. Among the people who have contributed to this outlook is a Russian linguist by the name of Sergei Starostin. He is one of those who proposed that Chinese is, among other things, related to some of the languages in the Caucasus, some of the languages in Siberia, as well as many languages in the New World. The database compiled by this project contains phonetic transcriptions of hundreds of languages spoken in many parts of the world. These transcriptions provide the basis for reconstructing the earlier stages of

[2] "Language Evolution," *Santa Fe Institute Bulletin*, (2001):14-16.

previously spoken languages. Actually the idea of such a database, though not as ambitious as the one in Santa Fe, can be found in the work of linguists in Peking University starting in the 1950s. Under the direction of Yuan Jiahua, a series of handbooks were compiled for several thousand basic words and their pronunciations in numerous Chinese dialects. The *Hanyu Fangyin Zihui* (Phonetic Dictionary of Chinese Dialects), first published in 1962, has recently re-appeared in a newer form, edited by Wang Futang. When a group of us at Berkeley obtained a copy of the *Zihui* in the 1960s, we were immediately impressed by its importance as a tool for research on the history of Chinese. Instead of perhaps illustrating linguistic changes with a handful of examples, we wanted to make it possible to suggest and verify hypotheses by exhaustive and large-scale searching through databases such as *Zihui* on a computer. We put in the *Zihui* data on a computer, as well as extended it, for example, with Zhongyuan Yinyun, pronunciations in Korean, and several layers of Japanese, thus created the DOC, Dictionary on Computer.[3] The data were first punched on in old teletype machines, so old that they can now only be seen in museums of history of technology. This was a gigantic task in the 1960s, which took a dedicated amount of efforts of many good friends. Among them are the major contributors like Cheng Chin Chuan, who became so much an expert with the teletyped paper tape that he could pick up a piece off the floor and immediately recognize the encoded content, and Hsieh Hsin-I, who was so impressed by Chin Chuan's super-human ability that he once jokingly told us 'Chin Chuan is not a human'. It is a truism about corpora that how useful it is depends on how it is used. From the point of view of language change, DOC studies have contributed by providing empirical support for lexical diffusion, the theory that originated from studies on the Chinese language. Recently at the invitation of the Santa Fe Institute, Chin Chuan shipped them the latest version of DOC, and DOC now lives on as part of the 'Evolution of Human Languages' project.

[3] William S-Y. Wang, "Project DOC: Its Methodological Basis," *Journal of American Oriental Society* 90 (1970):57-66; Mary L. Streeter, "DOC 1971: A Chinese Dialect Dictionary on Computer," *Computers and Humanities* 6 (1972):259-70; Chin-Chuan Cheng, "DOC: lts birth and life," in *In Honor of William S- Y. Wang: Interdisciplinary Studies on Language and Language Change*, ed. by M.Y. Chen and O.J.L. Tzeng, (Taibei: Pyramid Press, 1994), 71-86.

The third category of specialized linguistic corpora is on language typology, an area that derives its major inspiration also from the late Joseph Greenberg and his research on language universals. The pioneering work on the *World Atlas of Language Structures* (WALS) was done at the Max Planck Institute for Evolutionary Anthropology in Leipzig, and the current online version became available in April 2011. Again the success of compiling a corpus depends crucially on how well it is used. Along this scale, the WALS is remarkably successful. For example, in 2007, there is a paper published in the *Proceedings of the National Academy of Sciences* with the title 'Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes: ASPM and Microcephalin'. The point of this article is that we carry certain genes that predispose us to speak tone languages. The evidence that it gives is a very strong correlation between tone speaking populations and the genetic makeup. Such an ambitious claim of course would not go unchallenged. A year later, there is a rather critical review on the result in the new journal *Biolinguistics*, challenging the causal inference reported in the paper.[4] The claim is the relationship between genetic and linguistic diversity, in this case, may be causal: genes cause language and influence the trajectory of language change through iterated cultural transmission.

There is another example in the very prestigious journal *Science*. Another use of WALS was made in the paper titled 'Phonemic diversity supports a serial founder effect model of language expansion from Africa'.[5] The conclusion is: this result points to parallel mechanisms

[4] Dediu, Dan & D. Robert Ladd,. "Linguistic Tone is Related to the Population Frequency of the Adaptive Haplogroups of Two Brain Size Genes, ASPM and Microcephalin," *Proceedings of the National Academy of Sciences of the United States of America* 104 (2007):10944-49; Ladd, D. Robert, Dan Dediu & Anna R. Kinsella. "Languages and Genes: Reflections on Biolinguistics and the Nature–Nurture Question," *Biolinguistics* 2(2008):114-26; Joshua Bowles, "Some Questions about Determining Causal Inference and Criteria for Evidence: Response to Ladd, Dediu & Kinsella (2008)," *Biolinguistics* 2(2008):247-55; Ladd, D. Robert, Dan Dediu & Anna R. Kinsella, "Reply to Bowles (2008)," *Biolinguistics* 2(2008):256-59.

[5] Quentin D Atkinson, "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa," *Science* 332 (2011):346-9; Jaeger, T. Florian, Daniel Pontillo & Peter Graff, "Comment on 'Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa,'" *Science* 335(2012):1042a; Michael, Cysouw, Dan Dediu & Steven Moran, "Comment on 'Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa,'" *Science* 335(2012):657b; Rory Van Tuyl & Asya Pereltsvaig, "Comment on 'Phonemic Diversity Supports a Serial

shaping genetic and linguistic diversity and supports an African origin of modern human languages. Earlier anthropologists have been telling us that genes, peoples, and modern peoples all came out of Africa perhaps about 100,000 years ago. This paper makes a similar claim but for languages, and it does this on the basis of counting the phonemes in the entries of languages. So this is based on 504 languages, and the use of a large number of samples from the WALS. This article has captured a lot of media attention and I was invited by the *South China Morning Post* to comment on it. To give it a popular appeal, I entitled the column 'Do you speak African?'.[6] I have again various reservations about such a conclusion, because Atkinson is a psychologist and quite often people outside of the field are not quite sensitive to the difficulties within the field. As linguists, of course, we know that it is not so easy to count phonemes. This message was signalled to us very clearly in a classical article by Chao Yuen Ren in 1934 called 'The non-uniqueness of phonemic solutions of phonetic systems'.[7] So before you count, you got to know what you are counting, and people are not agreed on how to actually get phonemes for phonetic systems. For instance, English words like 'bait' and 'boat' are sometimes written with diphthongs. Some are written with consonant plus glide, and some linguists prefer to spell them with single phonemes. Depending on the choice, English will have a different number of vowel phonemes. Daniel Jones once said Chinese, or Putonghua, has two vowel phonemes; Charles Hockett said that there are three. If you look at the literature, sometimes you have five, sometimes you have seven. So based on this uncertainty, how can you come up with a very far-reaching hypothesis like this? This was one of the reservations that I have. But corpora will continue to improve in the coming years in efficiency and in power. Computer science will invent better and better data structures as we connect our world with other bodies of our

Founder Effect Model of Language Expansion from Africa,'" *Science* 335(2012):657c; Chuan-Chao Wang, Qi-Liang Ding Hui Li & Huan Tao, "Comment on 'Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa,'" *Science* 335(2012) 657d; Quentin D. Atkinson, "Response to Comment on 'Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa,'" *Science* 335(2012):1042.

[6] William S-Y Wang, "Voices out of Africa?" *South China Morning Post* 13, May 29, 2011, Sunday.

[7] Yuen-ren Chao, "The Nonuniqueness of Phonemic Solutions of Phonetic Systems," *Bulletin of the Institute of History and Philology*, *Academia Sinica* 4(1934):363-97.

knowledge. For instance one very strong feature of WALS is that it connects to Google Maps, an extremely powerful service on the web. I think future corpora must look in that direction.

I would like to return to the legacy of the 1998 volume on the quantitative study of language mentioned at the beginning. In that volume, Cheng Chin Chuan contributed a very interesting paper called 'Quantification for Understanding Language Evolution'.[8] In that paper he raised a question which concerns the number of words that a person can actively use. Towards answering this question, he prepared a graph, which on the one hand shows the collections of characters in time, starting from Jiaguwen going through the various centuries over 2,000 years. For instance *Shuowen Jiezi* has over 9,000 characters and *Hanyu Da Zidian* over 56,000 characters. As the culture moved on, it accumulated new vocabulary, invented new characters for it, so the set of characters grew. On the other hand, the graph also shows the number of distinct characters used in various dynastic histories. For instance *Shiji* has around 5,000 characters, and then *Tangshi*, *Mingshi*, up to *Qingshigao* with over 8,000 characters. So this is a very interesting contrast. As a greater number of characters become available, we do not use them but instead stick with a kind of limit of about 8,000.

To generalize, we find that this result reminds us of something that Charles Darwin said: "We see variability in every tongue, and new words are continually cropping up; but as there is a limit to the powers of memory, single words, like whole languages, gradually become extinct."[9] I will make just two comments. Is our memory really so limited for words? What happens when you learn a new language – suddenly you learn another batch of thousands of new words? How does the memory bank work for learning words? Is it stratified? Certain words go to one bank and certain other words go to another bank? Chinese goes to one bank, English goes to another bank, and French goes to another bank?

[8] Cheng, Chin-chuan.鄭錦全, "Cong jiliang lijie yuyan renzhi 從計量理解語言認知" (Quantification for understanding language cognition), in *Hanyu jiliang yu jisuan yanjiu* 漢語計量與計算研究 or *Quantitative and Computational Studies on the Chinese Language*, ed. by B.K. T'sou, T.B.Y. Lai, S.W.K. Chan & W.S.-Y. Wang (Hong Kong: Language Information Sciences Research Centre, City University of Hong Kong, 1998), 15-30.
[9] Charles R. Darwin, *The Descent of Man, and Selection in Relation to Sex* (London: John Murray, 1871).

Nouns go here and verbs go there? How does it work? Is it some kind of a 'push-through' storage so that as new words come in, old words go out? Or is it some kind of a 'push-down' storage: the earlier the word is learned, the more strongly it is fastened? Then after you reach a limit, it is very difficult to learn new words? All these, I think, are just fundamental questions about the nature of language, things I hope that more and more emphasis on corpora study will shift to.

The other comment that I want to make with respect to Chin Chuan's graph is that he was not really working on words. He was working with characters. Characters, syllables, morphemes and words are all very different cognitive units. Can we somehow relate them together? I think we are still quite far from being able to come to a coherent picture on how these cognitive units actually relate to each other, but I will just very briefly mention the experiment that we have completed. Hopefully this will add just another chip to this big area of questions concerning syllables, characters, morphemes and words.

The experiment started because we saw a paper in *Physica A*,[10] in which a few Portuguese linguists took the syllables in Portuguese and analyzed them in terms of networks. There are many types of networks, and one especially interesting type is called 'Small World Networks', which became really a hot topic in science since it was described in a very important paper in *Nature* in 1998.[11] So in an example of a very small network based on the paper that Peng Gang, James Minett and I published in the *Journal of Quantitative Linguistics*,[12] huǒ (火) for instance has three partners: huǒchē (火車), huǒzāi (火災) and huǒjǐng (火警), but huò (貨) in this very limited network has only one partner. So there can be a lot of difference among the number of partners that a particular character has in terms of distinct words it can form association with. If we do an extensive analysis for Putonghua and for Cantonese, we find that such networks do follow power-law distributions. So at one end, we have very rare syllables or characters like hú (蝴) as in húdié (蝴蝶) or xī (犧) as in

[10] Medeiros M. Soares, G. Corso & L.S. Lucena, "The network of syllables in Portuguese," *Physica* A 355(2005):678-84.
[11] J. D. Watts & S. H. Strogatz, "Collective Dynamics of 'Small-World' Networks," *Nature* 393(1998):409-10.
[12] Gang Peng, James W. Minett & W.S-Y. Wang, "The Networks of Syllables and Characters in Chinese," *Journal of Quantitative Linguistics* 15(2008) 243-55.

xīshēng (犧牲). At the other end, we have syllables or characters which participate in hundreds of words as in fūzǐ (夫子) and dàrén (大人). This is true for both Putonghua and Cantonese. I have been trying to relate these findings with Chin Chuan's graph on characters diachronically and the LIVAC findings which also show that quite often a very small number of characters can cover a huge percentage of the actually used words in running corpora.

Systematic development and creative use of large corpora is becoming a major trend in research on language, and that is as it really should be. It should not be marginalized like 'Ah, that is corpus linguistics', which is a ridiculous statement because that is what linguistics should be. The study of the Chinese language within this perspective is already in the position of strength thanks to the contribution of many research teams. So, resources like LIVAC, for instance, must be further enhanced and used creatively. After all, a corpus is only as good as how it is used.

June 2015 William S.-Y. Wang