DOI: 10.17469/O2106SLI000002

GU YUEGUO

Reflections on the Foundation of Corpus Construction: An Argument for Experience-based Conceptualization

This paper presents reflections on the foundation of corpus linguistics, specifically around two basic issues: (1) Does corpus linguistics, bearing the label *linguistics*, contribute to our understanding of language in general or only specific languages? (2) Does corpus linguistics contribute to our understanding of man, or the mind of man, or individuals as speakers? Four positions are critically reviewed, covering three orientations, viz. product-oriented, process-oriented and experience-oriented. It is argued that experience-oriented conceptualization is a viable direction for future development. The tripartite division of labour in corpus construction – conceptual corpus linguistics, ideal corpus linguistics and practical corpus linguistics – is proposed and demonstrated. It points out that Bunge's systemism and ontology should be adopted as the foundation of corpus construction.

Keywords: experience-oriented conceptualization, corpus ontology, systems thinking.

1. *Preliminary Remarks*

Reflections on my 35-year compilation of the Spoken Chinese Corpus of Situated Discourse (the SCCSD, www.multimodalgu.com) have made me become suspicious of some basic issues about corpus linguistics in general:

1. Does corpus linguistics, bearing the label *linguistics*, contribute to our understanding of language in general or only specific languages?
2. Does corpus linguistics contribute to our understanding of man, or the mind of man, or individuals as speakers?

All practitioners of corpus compilation know through years of sleepless nights that these issues are central to corpus building of any kind, big or small. Practitioners also are aware that they have also been pon-

dered upon by some corpus pioneers. This paper first presents a brief critical review of the pertinent literature. The bulk of the paper will be dedicated to an argument for a holistic and experiential approach to the theorization of corpus linguistics, which will throw, hopefully, some light on the two fundamental issues.

2. *Corpus construction and its bearing on language/languages*

Historically, corpus construction witnesses an event that has exerted ever-lasting consequences, namely the use of computer. So significant is it that Leech (1992), based on it, draws a distinction between computer corpus linguistics and non-computer corpus linguistics. His distinction goes beyond sheer recognition of technology. He writes (1992:106): "I wish to argue that computer corpus linguistics (henceforth CCL) defines not just a newly emerging methodology for studying language, but *a new research enterprise*, and in fact *a new philosophical approach to the subject*." (italics added). In what sense does CCL constitute a new research enterprise? Leech points out:

> "On the face of it, a computer corpus is an unexciting phenomenon: a helluva lot of text, stored on a computer. But the computer's ability to search, retrieve, sort, and calculate the contents of vast corpora of text, and to do all these things at an immense speed, gives us the ability to *comprehend*, and to *account for*, the contents of such corpora in a way which was not dreamed of in the pre-computational era of corpus linguistics." (Leech, 1992: 106; italics original)

Speedy comprehension and massive data-based accounting-for still make a vivid picture of the present-day corpus linguistics research. These technical advantages aside, what is Leech's "a new philosophical approach to the subject"? His view is best seen through a comparison he makes with "other approaches in linguistics", primarily "the Chomskyan paradigm". It is summarized in four focuses as follows:

1. Focus on linguistic performance, rather than competence
2. Focus on linguistic description, rather than linguistic universals
3. Focus on quantitative, as well as qualitative models of language
4. Focus on a more empiricist, rather than rationalist view of scientific inquiry.

Leech's four-focus characterization of "a new philosophical approach" does indeed indicate something new only if we view it comparatively.

Each focus by itself, however, can hardly be seen as something new, for it has been argued for and pursued elsewhere in linguistics long before CCL. There are two features standing prominently in Leech's argumentation. One is that it is bi-polemic in the sense that it is dichotomous between two opposing opposites: performance vs. competence; specific vs. universalistic; empiricism vs. rationalism. The other is reductionistic in the sense that the speaker, the ultimate owner, the agent, of any human language, is reduced to non-existence, since it plays no role in the theorization of what language is. In other words, the three pairs of dichotomies about language just mentioned is language without speaker. Language is thus treated as if it were an object independent from the owner/agent.

It must not be construed here that Leech has failed to see the fact that behind every human language, every piece of text there is an ultimate producer/speaker/owner. This fact is taken for granted and plays little role as a key concept in theory-building.

The weakness of Leech's conceptualization of CCL is somehow amended by Chafe's view about corpus linguistics. Chafe reports the outcome of his pondering "more deeply the place of corpora within linguistics. Since I believe they are an absolutely crucial part of the linguistic enterprise..." he still sounds quite apologetic: "I would like to take advantage of this symposium to try to articulate some ideas about how corpora further the *ultimate goal of understanding the nature of language*. I hope I will be pardoned for starting in a philosophical vein." (Chafe, 1992: 79; italics added). As indicated above, Leech's "philosophic" venture is through bi-polemic dichotomy and reductionism, Chafe's in comparison is certainly deeper, for he wants to explore, via corpus linguistics, the *nature of language*. He argues thus:

> "we will never get much farther than we have gotten up to now unless we accept and integrate into our work the realization that *language cannot be separated from the mind as a whole*, that understanding language and understanding the mind are at bottom the same endeavor". (Chafe, 1992: 80; italics added)

It is crucial to note that Chafe does not adhere to the modularity theory of mind, as most notably advocated by Chomsky (1986) and Fodor (1983). He holds an opposite view "that language is an inseparable part of total mental activity." (Chafe, 1992: 81). He feels "joy whenever they discover a way in which some linguistic phenomenon can

be characterized as motivated and functional – explainable within a larger, coherent picture of the mind." (Chafe, 1992: 81). So language, as seen by Chafe, is free from the tyranny of competence-vs-performance dichotomy. This view of language runs consistent and has been continuously pursued in Chafe's long academic career (e.g., Chafe, 1994, 2018).

Chafe's investigation commences from the mind's eye, and he wants to know how the mind processes "speaking", which "is natural to the human organism... humans are 'wired up' to speak and listen". In this endeavor, "corpora have led to the discovery of two important constraints on the way the mind processes information during the production ... of language." (Chafe, 1992: 88). The two constraints are: "the light subject constraint", and "the one new idea constraint". Simplistically paraphrased, the first constraint refers to the phenomenon that the sentence subject tends to be light in information weightiness, while the second to the hypothesis that each intonation unit is limited to no more than one idea that is new.

Chafe hence asks: "What, then, is a 'corpus linguist'?"

> "I would like to think that it is a linguist who tries to understand language, and behind language *the mind*, by carefully observing extensive natural samples of it and then, with insight and imagination, constructing plausible understandings that encompass and explain those observations". (Chafe, 1992: 96; italics added)

This paper, meanwhile applauding Chafe's depiction of the tasks of corpus linguists, feels somewhat disappointed at his step short of embracing the whole-person speaker as the basic naturally given unit on which the conceptualization of corpus linguistics should be founded. We shall turn to this theme in section 4 below.

3. *Sinclair's view of corpus linguistics*

Our brief review, no matter how brief, cannot be complete without a look at Sinclair's position. To the best of my knowledge, Sinclair (2004) is the most outspoken corpus linguist who explicitly calls for giving priority to dialogical model of language. Integrating the written and spoken language has remained a recurrent theme in many of Sinclair's works. His effort is best appreciated when it is seen in a larger context. Linell (2005) makes a list of 101 points proving the

written language bias in linguistics, the bias actually felt by quite a few corpus linguists as well. Biber et al. (2000: 1038) observe that "there is a compelling interest in using the resources of the LSWE[1] Corpus of spoken language transcriptions to study what is characteristic of the grammar of conversation." But immediately after they admit that the "fact that this Corpus material consists of transcriptions — speech rendered in written form — means that even here, the reliance on the written form of the language cannot be escape..." In spite of this written bias, they still feel that

> "the existence of such a large body of transcribed speech makes it feasible to seek an answer to the following question, which has recently excited considerable interest: *is there a distinctive grammar of spoken language, operating by laws different from those of the written language? If so, what are these laws, and what are the functional or other principles underlying them*?" (italics added)

Their answer is: "the same 'grammar of English' can be applied to both the spoken and the written language." What is extremely interesting to note, however, is that the authors of the chapter have to admit constantly the difficulty in putting the jacket of written language grammar onto rebellious conversation. Take the notion of *sentence* for example. They observe:

> "Whereas the **sentence** has been treated, traditionally and in modern theory, as the fundamental structural unit of grammar, *such a unit does not realistically exist in conversational language*". (Biber et al., 2000: 1039; bold original; italics mine)

Carter and McCarthy (2006: 9) write:

> "Most books on the grammar of English have had a bias towards the written language. For many centuries dictionaries and grammars of the English language have taken the written language as a benchmark for what is proper and standard in the language, incorporating written, often literary, examples to illustrate the best usage. Accordingly, the spoken language has been downgraded and has come to be regarded as relatively inferior to written manifestations".

Carter and McCarthy seem to spare no efforts in counterbalancing the written language bias by inventing, quite readily without feeling

[1] Longman Spoken and Written English.

awkward, fresh metalanguage to capture unique features found in spoken discourse.

So much for the larger context. Let us return to Sinclair. If Biber et al try to integrate the spoken into the written, Sinclair attempts to do the opposite, i.e., integrating the written into the spoken. In his paper "The internalization of dialogue" (2004 [1999]), he puts forward the hypothesis that "much of the complexity of sentence grammar can be explained as the internalization of features of spoken interaction. This hypothesis is quite plausible, considering the fact that the written form in any literate culture is a much later development than the spoken one. Sinclair observes that "it must be reiterated that by the time it became possible to develop a written form of a language, the structures were already capable of great complexity." (ibid., p. 104). This echoes Chafe's view of speaking as human species' wired-up property touched upon in section 2 above.

What is significant about Sinclair's position is that he goes beyond embracing spoken samples in corpus building, and proposes to conceptualize corpus linguistics on the basis of speaking instead of writing. Two terms, viz. *dialogic language* vs. *monologic language*,[2] are proposed to help conceptualize the transition from the spoken to the written. Dialogic language is language in an interactive mode, whereas monologic language does not require elaborate contributions from other participants. The internalization involves reduction and compression of information, with gains or losses of independence or interdependence. Take "a move" in dialogic language for example. It is independent in its dialogic environment, but when it is internalized in monologic language, it becomes a sentence, and "loses its property of constituting an independent utterance". The sentence, however, "gains the facility of posture as a property of main clauses". (Sinclair, 2004: 114).

The notion of internalization in fact originates from his earlier paper "Planes of discourse" (2004 [1982]), where Sinclair writes:

[2] Sinclair's use of the term *language* in the distinction needs to be construed with care in this paper. In oral-aural cultures, there exists no monologic language. In such cases the term *language* overlaps with the term *natural language*, i.e. the sense being used in this paper. In literate cultures, on the other hand, there exist both dialogic and monologic languages. Here the term *language* does not overlap with *natural language*.

> "Language in use has two aspects: at one and the same time it is both a continuous negotiation between participants, and a developing record of experience. The negotiation aspect highlights interaction and will be called the *interactive plane* of discourse. ...
> ...
> The other aspect of language in use is the developing record of experience. On a small scale, in a conversation, say, or reading a letter, it can be seen as gradual sharing of relevant experience by recalling previous words and phrases and reworking them in the new contexts provided by a movement on the interactive plane.
> ...
> The stage-by-stage tally of the record of experience will be called the *autonomous plane* of discourse, because it is concerned with language only and not with the means by which language is related to the world outside." (Sinclair, 2004: 52-53; italics original)

The autonomous plane of discourse, looked at on a larger scale, as pointed out by Sinclair, is "a continuous internalization of experience, from the world outside to the inner space of language. The process is both individual and collective, and, where written down, forms the most explicit record we have of human evolution." (Sinclair, 2004: 53).

In a word, Sinclair adopts the dialogical model of language in use as a reference framework against which the written sentence grammar is critically examined. He exceeds Chafe in two aspects. One is that, unlike Chafe, he does not take the current status of corpus linguistics as given, and attempts to make it anew through reconceptualization. The other is that Sinclair's notion of speaker is much fuller than Chafe's. As pointed above, Chafe incorporates human language and mind, whereas Sinclair incorporates human language use into (1) "continuous negotiation between participants", and (2) "a developing record of experience". Surely the concept of experiencing participant presupposes a thinking mind, but the reverse inference is invalid. The experiencing participant will be referred to as a whole person below. To which we turn.

4. *The participant's total saturated experience: A step further from Sinclair*

The experiencing participant possesses not only a thinking mind, but also multiple sensory modalities interacting with the outside world.

Language use in the real-life everyday world is naturally multimodal and experientially real. By "naturally multimodal" is meant that real-life situated discourse involves what Goffman calls "bodily activity" on the speaker/performer's side, and "naked senses" on the addressee/receiver's side. The speaker's *current bodily activity* makes ***embodied messages*** (Goffman's terminology).

> "When one speaks of experiencing someone else with one's *naked senses*, one usually implies the reception of embodied messages. This linkage of naked senses on one side and embodied transmission on the other provides one of the crucial communication conditions of face-to-face interaction." (Goffman, 1963: 15; italics mine).

As emphasized by Goffman, ordinarily in using the naked senses to receive embodied messages from others, one also makes oneself available as a source of embodied messages for others. This is why we emphasized above that Chafe's notion of mind is not the same with Sinclair's notion of participant.

The embodied messages, produced and received/interpreted by both sides, consciously and sub-consciously, make moment-by-moment experienced reality. Gu (2009) proposes the adoption of the term *total saturated experience* (TSE for short) to refer to Goffman's face-to-face interaction with naked senses and embodied messages, and of the term *total saturated signification* (TSS for short) to talk about the total meanings constructed out of the total saturated experience by the acting co-present individuals.

Gu in series studies (2006a, 2009, 2013, 2016) attempts to make Sinclair's suggestion about the experiencing participant a *central concept* in corpus linguistics. Sinclair's "a developing record of experience" is reconceptualized primarily in terms of three inter-related concepts, *experiencer-experiencing-experience* (i.e., EEE model for short), and secondarily the "developing recording" is dynamically modeled through the formula "the (dimensional) self {...{...}...}". Briefly, the participant fills this formula with multiple layers of data corresponding to the multiple selves through multiple sensory modalities of interacting with the outside world (see further discussion in 5.3 below).

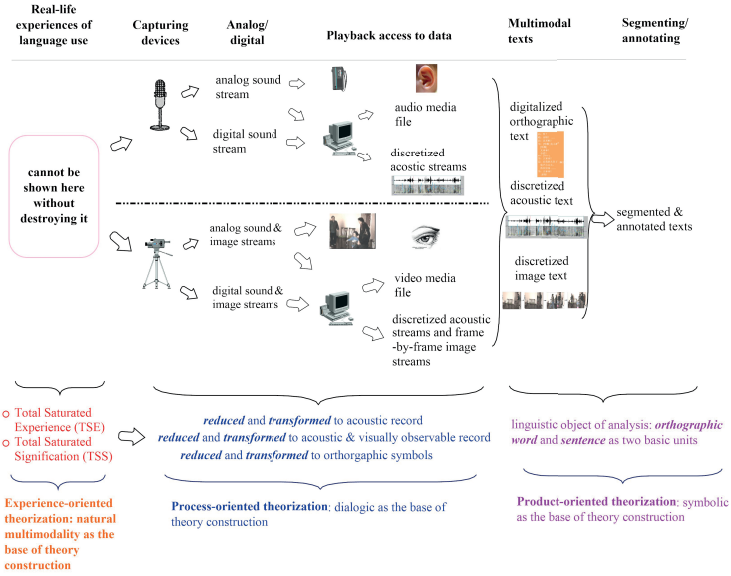
Gu (2009: 435) contrasts three perspectives in theorizing the study of language use: (1) product-oriented, (2) process-oriented, and (3) real-life experience-oriented. The first has been the most predominant one, whereas the second has been voiced as a potential alternative as

suggested by Sinclair. The third, in contrast, is little heard of. The three contrastive perspectives can be illustrated by the ensuing scenario.

> "Think about what happens to language use when the audio/video-recording device (or any other data-capturing devices) is switched on? It transforms real-life language-using experiences onto cassette tapes or hard disks. What is recorded on the tape or hard disk becomes data — real-life language use data. A corpus linguist then has an option of accessing the data: (1) by playing the medium back to listen to via the naked ear or watch it via the naked eye; and/or (2) transcribing the data into orthographic words or symbols before analyzing them; and/or (3) segmenting and annotating the data using tools such as PRAAT, ELAN, etc."

The whole process is graphically represented in Figure 1.

Figure 1 - *From real-life experience to text annotation*

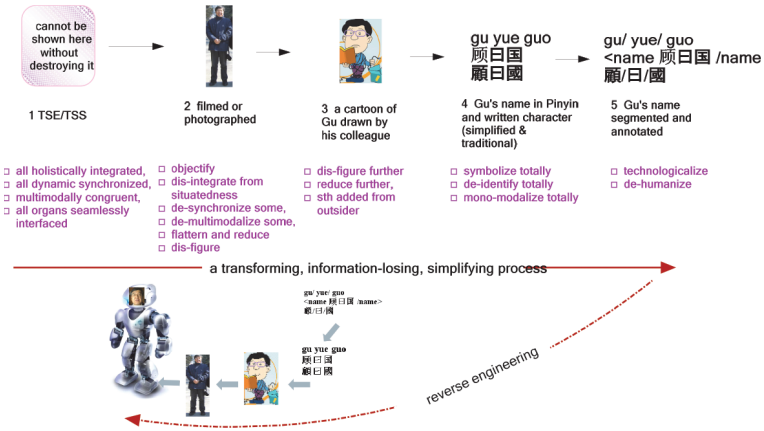


The product-oriented theorization shown on the bottom of the right side draws data from orthographic texts; the process-oriented theorization in the middle relies on dialogic data; the experience-oriented theorization attempts to develop a theory by modeling the natural multimodal interaction in the real-life world.

As shown in the figure, on the leftist side, it is the occurrence of real-life language use. It is a moment-by-moment creation of life experience, of which language use is a crucial part. It constitutes a total saturated experience and hence total saturated signification. This real-life experience cannot be captured by any device without destroying it. This echoes the fundamental tenet of Taoism: "Conceived of as having no name, it is the Originator of heaven and earth; conceived of as having a name, it is the mother of all things."[3]

The move from the product-oriented to the process-oriented, then to the experience-oriented and finally to the TSE-TSS can be seen as a reverse engineering process. This view assumes that language use experience lived by participants constitutes the ultimate goal of understanding (in the sense as defined by Chafe) to be achieved by corpus linguistics. Those corpus linguists adopting this view will endorse Gu's argument that the product-oriented theorization is based on *very much impoverished, if not distorted data*. This may be illustrated by the analysis of the author himself from the real-life person, alive and kicking, to his photograph, to his cartoon, and finally to his orthographic name (see Figure 2).

Figure 2 - *From real life to orthographic text and vice versa*



The bullet items under each stage indicate its data properties, and the transformations and distortions are clearly seen by comparing the

[3] Legge's translation of Laozi's Daodejing (Legge, 2008 [1891]).

data properties between the stages. At Stage 1 represented by TSE-TSS Gu is an all-round living system; at Stage 2 he is objectified, and flattened into a lifeless 2-dimensional figure; at Stage 3 he is further distorted and disfigured; at Stage 4 he is totally symbolized and de-identified – which can be testified if someone, having never seen him, is asked to pick Gu up from a crowd only by a name slip; at Stage 5 Gu's orthographic personal name is segmented and annotated for computer to do personal name recognition.

Now a corpus linguist is given a task of making a hominoid robot of the author capable of speaking Chinese, if given a corpus of orthographic texts. This is reverse engineering from Stage 5 back to Stage 1. Note that such reverse engineering tasks are not imaginative or mind-experiment: Paleoanthropologists, paleographers, and historiographers face challenges like this in their daily routines.

The illustration shows a self-defeating element intrinsically built in as a result of its linguistic stance and research methodology corpus linguists assume. By the time they finish crunching words, sentences, patterns, etc., and declare their findings about language use, the picture drawn is likely to be skewed in the way Indian blind men report their findings about the elephant!

Those corpus linguists who do not endorse Chafe's ultimate understanding view may counterargue that, since it is impossible to attain full data sets mapping the TSE-TSS, it becomes pointless to pursue experience-oriented theorization. This counterargument, if adhered to, will hinder advancement of corpus linguistics enterprise. We shall show some benefits of such theorization and discuss pertinent ontological and methodological issues.

5. *Tripartite division of labour in theorizing corpus linguistics*

5.1 Thinking in general systems theory

This paper proposes a solution to fending off the self-defeating element mentioned above. It proposes a tripartite division of labour of theorizing corpus linguistics, namely *conceptual corpus linguistics*, *ideal corpus linguistics*, and *practical corpus linguistics*. Conceptual corpus linguistics adopts the stance of looking at language use as it being experienced and lived by participants, and extracts know-how knowledge from it by way of constructing experiential constructs. Ideal corpus

linguistics, on the other hand, adopts a stance of looking at language use as an object to be modeled, and makes data models on the basis of the experiential constructs. Finally practical corpus linguistics looks at language use as texts, and its central task is to handle practicalities of text processing, meanwhile adopting experiential constructs and data models as its blueprints.

The hallmark of such conceptualization of corpus linguistics is *its general systems theory approach from real-life experience to text processing, and to data reverse engineering*. It does not stop at looking at language use as phenomenological reality. It goes further than that. For technically speaking, language use as it being experienced is an open-ended, continuous, and emergent analogue data type. It has to be modeled, sampled, quantized, and discretized to produce machine-friendly data type for efficient and effective processing. Experiential reconstruction and data modeling as research methodologies are no less important than orthographic transcription and word-crunching.

5.2 Bunge's ontological theory: A synopsis

Before we proceed, a synopsis of Bunge's theories of ontology (or metaphysics) and systemism will be helpful in case that they have not caught interest of some corpus linguists. Bunge differentiates "exact philosophy" (Bunge, 2016: Chapter 9) from inexact philosophies. His exact philosophy refers to philosophizing in such a way that all the basic concepts of philosophy, e.g., substance, thing, process, property, state, possibility, are defined in the metalanguage of basic set theory and pertinent mathematics. In Bunge's theorization, the terms ontology and metaphysics are interchangeable. He adopts "a naturalistic (or materialistic) world view", which is taken to "be the ontology of factual science and technology" – keep it in mind that corpus linguistics is technology-driven.

> "On this view the world is composed exclusively of concrete changing things: everything else is the invention of particular concrete things such as ourselves. Clearly, on this view language cannot exist in the same way as stars and people exist, i.e. in and by themselves. On this view *what are real are not languages but people, or other rational beings, engaged in producing, conveying or understanding linguistic expressions.* Asking whether language exists is like asking

> whether life or mind exist. The answer is an unqualified 'No'. There are no autonomous languages any more than there is life or mind by itself. There are instead minding animals and, in particular, animals capable of speaking and understanding speech. This, *the production and understanding of speech, we take to be the primary linguistic fact*. Everything else about language is construct – starting with language itself. Shorter: speech is real, language is not." (Bunge, 1984: 112-3; italics added)

It is essential to bear in mind that it is the minding animals capable of speaking and understanding that are real and concrete. Language is a construct out of speech. Bunge draws a distinction between "the philosophy of language", a part of ontology and epistemology, and "the philosophy of linguistics", a part of the philosophy of science (Bunge, 1984: 111). The former is concerned with such basic questions as "What is language?"

> "The basic question, 'What is language?' is an ontological one in the same category as 'What is life?'
>
> ...
>
> The nature of the question is likely to be better understood in attempting to answer the related question 'How does language exist?'. According to idealism language exists by itself, either as a sort of Platonic idea pre-existing people and hovering above them, or as a human creation though one that is immaterial. Needless to say, there is no empirical evidence for either variety of idealism." (Bunge, 1984: 112)

Bunge advocates materialism that holds that language does not exist by itself, contrary to idealism's position. The postulate that language does not exist by itself becomes self-evident only with reflection on the fact that there is no language without speakers. If language exists, its existence is secondary and derivative. Ironically, linguists accept this basic fact on the one hand, they ignore it while theorizing language on the other. It is a wide spread practice that language is regarded as an autonomous, independent system existing by itself. In other words, while linguists adopt a materialistic philosophy of language, they apply an idealistic philosophy of linguistics in practice. There are some serious consequences as a result of the discrepancy between the materialistic ontology and the idealistic practice. As Dillinger points out:

> "... linguists' theory is not developed as explicitly as their practice; the study of language and languages is fragmented, each subspecialty proceeds quite autonomously from the others; theoretical writing and textbooks present the field as a potpourri of activities without any explicit relations between them; and mutually exclusive "approaches" proliferate, each championing the study of one or a few fragments of the whole." (Dillinger, 1990: 10)

Dillinger's critique can readily be testified by the reading experiences students of linguistics have had while reading literature covering branches of linguistics, e.g. phonetics, phonology, formal syntax, formal semantics, pragmatics, discourse analysis, you name it. The reading of each, put in a Chinese proverb, constitutes an experience as difficult as crossing a mountain between them.

Bunge offers a revealing comment about "philosophical problems in linguistics" as follows:

> "... whoever regards language as an ideal object cuts the ties of pure linguistics with the other five branches of linguistics. Worse: *he isolates linguistics from the system of factual sciences*, all of which study concrete (material) things. And insulation is the mark of pseudoscience." (Bunge, 1984: 112; italics added)

Two pieces of background information are helpful in order to appreciate fully Bunge's critique: (1) Bunge's application of general systems theory to linguistics, and (2) his application of factual sciences to linguistics.

Bunge's favourite term for general systems theory is systemism, which is in contrast with individualism and holism. It refers to "a whole systemic worldview". It is centered in the following postulates:

1. Everything, whether concrete or abstract, is a system or an actual or potential component of a system;
2. systems have systemic (emergent) features that their components lack, whence
3. all problems should be approached in a systemic rather than in a sectoral fashion;
4. all ideas should be put together into systems (theories); and
5. the testing of anything, whether idea or artifact, assumes the validity of other items, which are taken as benchmarks, at least for the time being.

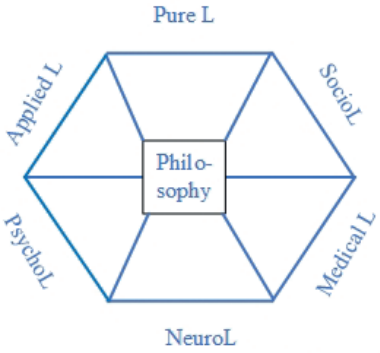
(Bunge, 2000: 149)

A *system,* concisely put, is "an organized whole in which parts are related together, which generates emergent properties and has some purpose." (Skyttner, 2005: 58). Systems are usually classified into four types: (1) concrete, (2) conceptual, (3) abstract, and (4) unperceivable. Alternative classifications are often made (1) between natural and man-made systems; (2) between living and non-living systems. The concrete system is the most common and also called physical system.

Everything can be modeled as a system, so language is a system, and linguistics is also a system. System-modeling like this, surly advantageous, has a defect of making people overlook the fact that *language system and linguistics system have very different referents, the former of which is the set of real people, whereas the latter is the set of man-made theories*. The factual sciences, on the other hand, study concrete/material things, and linguistics, if it wants to protect itself from falling into pseudoscience, must study concrete/material objects, e.g., speeches as recorded and collected in corpus. Logically, linguistics aiming at idealized and abstract phenomenon violates the postulates of factual sciences, hence it is pseudoscientific at best.

Bunge has drawn a "linguistic hexagon" visualizing the philosophy of linguistics – not to be confused with the philosophy of language – as shown in Figure 3.

Figure 3 - *Bunge's linguistic hexagon (Bunge, 1984: 111)*



Note: The linguistic hexagon or system of disciplines that study language. Here 'L' stands for 'linguistics'. Pure L is construed as the study grammars, which – since Chomsky (1965) – include syntax, semantics, and phonology.

Bunge's linguistic hexagon must not be construed strictly literally, that is, it can easily be expanded into a polygon like heptagon, octagon, and so on. The essential message is that "the different conceptions of language are related not only to the diversity of linguistic schools – each of them attached to its own philosophy – but also to the current fragmentation of the study of language into half a dozen different disciplines. These disciplines, that are only tenuously connected to one another..." (Bunge, 1984: 109).

5.3 Conceptual corpus linguistics: A whole person model

With Bunge's theories of ontology and systemism as our theoretical backbones, we are ready to give some meat to the tripartite division of labour introduced in 5.1 above. First, the tripartite division presupposes oneness of the overall task, viz. conceptualizing corpus linguistics as a systemic system in Bunge's sense of the term. Only in this way do we hope to avoid the fragmentation and mutual incomprehensibility found in the current branches of linguistics.

Conceptual corpus linguistics abides by the postulates of factual sciences. Accordingly, it is concerned with concrete/material objects with real-life existence, viz. individual speakers alive and kicking. The individual speaker is seen as a system which is part of a bigger system, viz. a society where s/he makes a living. Like any other systems, the individual speaker system is understood basically through four stages of human knowledge development: (1) intuition, (2) fact-finding, (3) analysis and (4) synthesis (Skyttner, 2001: 30). Corpus compilers are by nature individual speakers, hence they have naturally grown intuitions about themselves, about what to speak, and about many other things besides. Fact-finding is delineated by one's ontological view about what constitutes a fact. "Wittgenstein started his famous Tractatus of 1921 asserting that 'the world is the totality of facts, not of things'"(requoted from Bunge, 2016: 249). Bunge counterargues that

> "in the factual sciences 'fact' denotes either a state of a thing or a sequence of states of a thing: there are *no facts in themselves, without material things*. For example, there would be no car collisions without cars, or changes of government without rulers and their subjects. Remember Aristotle's criticism of Plato's idea of movement in itself rather than moving things." (Bunge, 2016: 249; italics added)

Following Bunge, there is no such linguistic fact in itself: *There are only linguistic facts as states of a speaker*. This statement is taken as our fundamental commitment to the ontological foundation of conceptual corpus linguistics. What is recorded when a corpus compiler switches on a tape recorder? Is he recording a linguistic fact? *It is not unless it is taken as a state or a sequence of states of the speaker being recorded*. We shall come to this later in section 8 below.

Let us return to the third and fourth stages above: analysis and synthesis. Analysis examines the components of a system, while synthesis expounds the synergetic properties of a system as a whole. In other words, analysis decomposes the system and synthesis elucidates how the system functions as a whole in a larger system.

Both analysis and synthesis are aided by a variety of tools. The tools we employ for conceptual corpus linguistics is systems theory, and in particular Bunge's systematerialist ontology expounded in Bunge's *Treatise on Basic Philosophy* (Volumes 3 and 4, 1977, 1979). Bunge's systematerialist ontology is a formidable general theory dealing with the "furniture of the world" and "a world of systems". Accurately elucidated key concepts include, among others, substance, property, state, process, thing, system, emergence, submergence, space, time, causality, randomness, space, time, chemism, life, evolution, mind, society, social structure, participation, marginality, social cohesion, and history. The metalanguage employed for this giant task is

> "such elementary formal tools as set and function. In addition, the exposition follows the axiomatic format: primitive (undefined) concepts and defined ones, axioms (or postulates), theorems, and comments. However, the motivation and justification of my principles (axioms) originated in the sciences." (Bunge, 2016, pp. 260-261)

What I have benefited most is his way of dealing with a very complex system in the simplest possible but accurate way. His systemism is concisely captured by himself in his *Memoir* as follows:

> "I have proposed ... that any system σ may be schematically modeled by the ordered quadruple composition-environment-structure mechanism, or
>
> $$\mu(\sigma) = \langle C(\sigma), E(\sigma), S(\sigma), M(\sigma) \rangle,$$
>
> where

C (σ) = Set of constituents of σ at the given level of analysis;

E (σ) = Immediate environment of σ;

S (σ) = Set of bonding and non-bonding relations among the system's constituents;

M (σ) = Set of processes that keep σ going."

The quadruple model is illustrated thus:

> "For example, the composition of an organism at the cellular level is the set of its cells, whereas at the organ level it is the set of its organs; the structure of an organism is the totality of its bonding relations, such as connecting tissues and hormonal fluxes, and nonbonding relations, such as those of position and succession; and the mechanism of a system is composed of the processes that keep it alive, in the first place metabolism and interaction with its environment." (Bunge, 2016: 252-3)

Now the basic living-speaker scenario conceptual corpus linguistics faces is this:

> "A baby, once born, engages itself in postnatal experience of all kinds non-stop until death. During this "from womb to tomb" lifespan, it undergoes stages from pre-speech, to babbling, to talking freely, to writing (if educated) and finally to speechless death."

To conceptualize such lifespan development of speaker in such a way that it can be dealt with in terms of corpus linguistics, Gu and Xu (2013) and Gu (2016) have proposed a lifespan scheme of data development. In a nutshell, living is coextensive with experiencing which is coextensive with meaning-making that is equivalent to building a lifespan big data store. The big data formula reads like this:

The multi-dimensional self {...{...}...}

1. "Self", a technical term here, is a special data folder as it were that maintains the dynamic fluid data from the living-experiencing-meaning-making scheme;
2. "Dimension", also a technical term, stands for a specific aspect or property of the speaker under investigation. "Multi-dimensional" indicates our ontological commitment to the postulate that the human speaker is a system with multi-complexity.
3. {...{...}...} stands for a lifespan data set with many sub-sets.

The big data formula gives rise to a whole person model of speaker, as shown in Figure 4.

Figure 4 - *A whole person model of speaker (Gu, 2016: 488)*

The living being consists of the experiencing self, the meta-self, and the institutionalized general self. The experiencing self refers to the ever-going, moment-by-moment, multimodal interactions with the world, while the meta-self refers to the online or offline reflections on the experiencing self. The developing infant does not have a meta-self until about year 3. The institutionalized general self, on the other hand, is a set of identities the self has co-constructed with the community while making his living in it. In China, not until about the late 1980s, did the newborn baby was simultaneously given an institutionalized general self, i.e., national ID number.

Each self is modeled from a range of perspectives. Take the experiencing self for example. It has dimensions such as physiological, psychological, linguistic, learning, familial, working, socializing, spiritual/religious, etc. These dimensions of the experiencing-self undergo various phases of development. The meta-self, the ability of formulating second-order beliefs, desires, evaluations, etc., has the dimension of online reflection, and the I-me reflection.

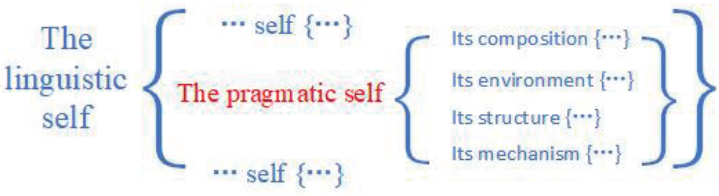
5.4 Critique of whole person model

The whole person model is in accordance with Bunge's systemic world view mentioned in 5.2 above. The five postulates – everything being a system, having emergent property, systemic not sectoral, all ideas into systems, and coherent validity for all – are upheld by the model. We cannot demonstrate them here for lack of space.

The model, against the yardstick of Bunge's composition-environment-structure-mechanism quadruple, suffers from many loopholes. Each dimensional self {...{...}...} should be further specified in terms of composition, environment, structure, and mechanism separately. Take the experiencing self for example. In view of the environment, it should be further specified into the experience-expectant growth self and the experience-dependent growth self (for this distinction see Greenough et al., 1987). The former refers to the fact that the infant's certain functions require basic experiences in order to develop, e.g., visual development requires visual stimulations of light from the environment, rooting behavior requires the stimulation from the breast nipple or the mock-up one, and most importantly, the development of speech requires the stimulation of speech sounds from caregivers. The latter, on the other hand, refers to the fact that some brain functions depend upon particular experiences, e.g., acquisition of a specific speech/language, e.g., a baby born and grown in Rome will acquire Italian due to specific input of speech, whereas in Beijing Chinese (i.e., Putonghua, a common language).

Take the pragmatic self for another example. It is a sub-set of the linguistic self, meanwhile it has its own sub-sets, which specify the quadruple, as shown in Figure 5.

Figure 5 - *The pragmatic self and its quadruple*

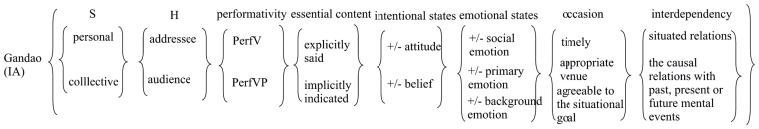


The pragmatic self shown in the figure can be further fine-tuned in terms of sub-sub-sets e.g., the illocutionary act self {...}. Gu (2013: 317) proposes an octet scheme as follows:

> "The Illocutionary act/force {S{ }, H{}, performativity {}, the essential content {}, the intentional states{}, the emotional states{}, the occasion {}, the interdependency {}}"

Adopting this scheme, the Chinese illocutionary act of *gandao* (feel) is conceptually analyzed as follows (see Figure 6):

Figure 6 - *Illocutionary act of gandao: a conceptual model (Gu 2013: 318)*



As the sample indicates it, Bunge's quadruple is substantiated in the octet. The bonding/non-bonding relations are handled in the sub-set of interdependency.

Up to this point one may wonder if such conceptual corpus linguistics leads to a jumble and a dead end. Our answer is NO. Because the whole enterprise is built on the basic set theory and Bunge's factual sciences. All sorts of selves {...{...}...} make a consistent and coherent whole through such logical operations as relations by union, and/or disjoint. As shown in Figure 6, members of the octet set hold logical union relations between them, while the values of each member, viz. the sub-sub-set members, are primarily disjoint between them, only under rare conditions are union. We cannot go to details here. Admittedly all we have achieved so far about conceptual corpus linguistics is only a tip of iceberg. It only points to a direction that may be fruitfully explored.

6. *Ideal corpus linguistics*

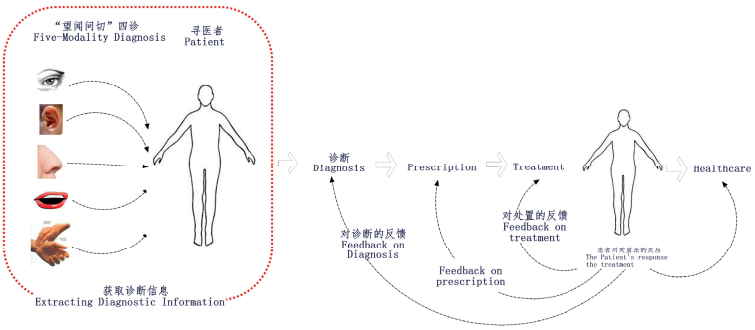
The conceptual corpus linguistics sketched above centers itself around the speaker as a whole person, which is the only "concrete, real thing" (in Bunge's terminology) that factual sciences deal with, thus treating speech as facts derived from real thing, and language as a construct.

Our position is that if corpus linguistics wants to be part of factual sciences, it must abide by this fundamental ontology.

Assuming that we endorse the ontology, we proceed to our next task, viz. conceptualizing ideal corpus linguistics. Ideal corpus linguistics, admitting the inadequacy of technology and limitations of human knowledge, attempts to think big and try to reach out to the bottom of iceberg. The central concern is how to compile a corpus that approximates, as faithfully, as accurately, as fully as possible, the human speaker's total saturated experience with total saturated signification. Eating a Roasted Peking Duck in a restaurant is a sequenced series of TSE-TSS states. An ideal corpus linguist, attempting to compile a corpus capturing this type of activity, faces a twofold task: (1) Make use of all sorts of data-capturing tools, currently available ones as well as imagined ones, (2) design a scheme for segmenting and annotating the collected data, manually or automatically or both. The twofold task is examined critically in depth in Gu (2009).

The practice of traditional Chinese medicine (TCM for short) provides us a real-life case study. A feeling-ill person comes to see a TCM doctor. The two are engaged in a TSE-TSS interaction empowered by natural multimodalities. The encounter however is unequal in terms of knowledge power. What is pertinent to our discussion here is the diagnostic method and the tools of data collection. The diagnostic method is called in Chinese sizhen (四诊), i.e., four ways of diagnosing, which are in our terminology five ways: by looking at, hearing, smelling, touching, and questioning, as visualized in Figure 7.

Figure 7 - *Diagnosing in TCM*



In view of ideal corpus linguistics, the practicing doctor is collecting multimodal data through natural multimodalities. Over three thousand years, such data collection and reflected practices have yielded thousand volumes of documents – TCM knowledge for short, and have saved a countless number of lives.

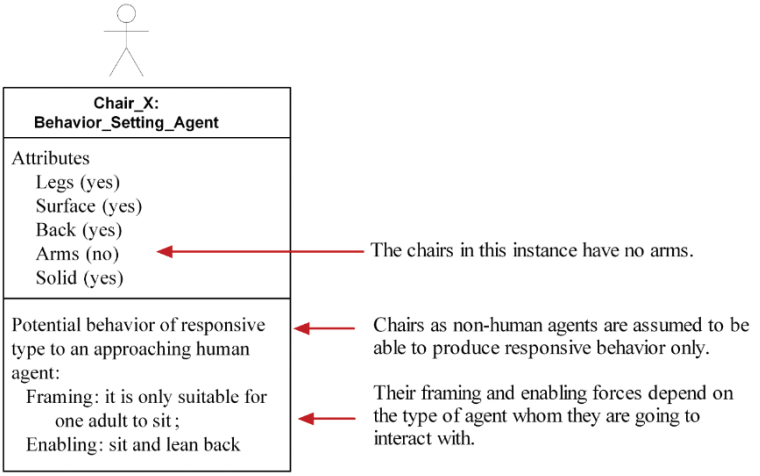
An agent-oriented modeling language (AOML) has been proposed in Gu (2006b, 2009), and can be used as an empowering tool for ideal corpus linguistics. As practicing corpus linguists well know, tools currently available, are technology-driven, i.e., empowered by the techniques, algorithms available to tool developers. There are programming languages (e.g. Java), mark-up languages (e.g. XML, TML, RDF), and modeling languages (e.g. UML). These languages are designed to talk to machines, and are not designed for corpus linguists to deal with raw data. XML, TML and RDF are in a way designed to deal with raw data, to mark them up so as to give them a machine-processable structure. But they do not offer much help to corpus linguists who face the problem of how to get the raw data first. AOML, in contrast, is designed to empower the practicing corpus linguist so that s/he can formalize human-interpreted TSE-TSS interactions. A telling advantage of doing so is that the formalization is quite intuitive to the corpus linguist on the one hand, and is easily convertible to software programming on the other.

The notion of agent has recently been given a great deal of attention in artificial intelligence (see, e.g., Hexmoor et al. 2003, Alonso et al. 2003, Ye & Churchill 2003, Russell & Norvig 2003, Wagner 2004). The AI agent is an automaton, and it is a part of a programming metalanguage. The agent of AOM is intended to be part of a modeling metalanguage.

The AOM conceives of the world, real or non-real, as consisting of agents of various kinds who interact with one another by way of (i) exchanging attributes, and (ii) exercising framing and enabling behaviors. This way of conception works quite naturally with modeling human agents. It takes some imagination to apply it to the interaction between, say, a human agent and a chair. In AOM, a chair, once engaged in interaction with an approaching human agent, becomes an agent as well. It has its own attributes, and exerts framing and enabling behaviors on the human agent. For instance, it enables the human agent to sit on it, and at the same time frames the human

agent's behavior, e.g. it does not allow the latter to stand on its arm (see Figure 8 for demonstration). When the same chair interacts with a fly, on the other hand, the framing behavior it exercises on the human agent disappears, and instead it enables the fly to perch on its arm very happily.[4]

Figure 8 - *Chair_Agent construct (quoted from Gu 2009: 447)*



It is obvious to corpus linguists that this notion of agent is quite intuitive. It is based on Gibson's (1986) ecological approach to perception, and is closely associated with his notion of affordance.

7. *Practical corpus linguistics*

To pre-empt some potential misunderstanding, it is important to bear it in mind that the term practical corpus linguistics is not meant to characterize some current practices. It is a relative concept circumscribed by conceptual corpus linguistics, and ideal corpus linguistics, as discussed above. Viewed from the perspective of software engineering (Chang, 2001, 2002, 2005), conceptual corpus linguistics represents the *stage of conceptual design*, and ideal corpus linguistics the *stage of data modeling*. Practical corpus linguistics, on the other hand,

[4] This paragraph is quoted from Gu (2009: 445-446).

deals with the third *stage of implementation/ application/evaluation*. The three-stage design and development ensure consistency, coherence, efficiency and effectiveness. This practice is rarely followed in corpus linguistics for several reasons. One is that it often takes years to compile a corpus; another is that it often takes cross-disciplinary, cross-department teams to complete a whole project, and thirdly, it is left to end users to decide how to use a compiled corpus. Thanks for these reasons, corpus linguistics may result in an incoherent, disintegrated state of affairs, as McEnery and Hardie observe:

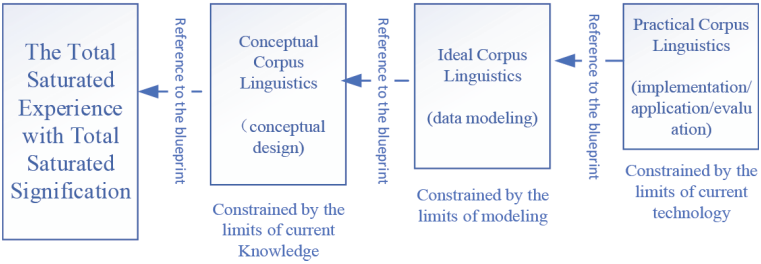
> "it is very important to realise that corpus linguistics is a *heterogeneous* field. Differences exist within corpus linguistics which separate out and subcategorise varying approaches to the use of corpus data." (McEnery and Hardie, 2012: 1; italics added)

Our proposal for separating practical corpus linguistics from the other two, meanwhile acknowledging the constraints imposed by practicalities such as the three reasons above, is intended to send a message that corpus linguistics is not merely practical and methodological, as it is held by many practitioners. McEnery and Hardie (2012: 1) capture this general spirit quite well:

> "What is corpus linguistics? It is certainly quite distinct from most other topics you might study in linguistics, as it is not directly about the study of any particular aspect of language. Rather, it is an area which focuses upon *a set of procedures, or methods*, for studying language". (italics added)

Practical corpus linguistics, in our conception, is both practical in the sense of constructing actual corpora for various purposes and theory-motivated in the sense of checking if it meets the blueprints laid down by ideal corpus linguistics and conceptual corpus linguistics, the blueprint of which is the TSE-TSS. Figure 9 visually demonstrates these relations.

Figure 9 - *Blueprint and reference relations in corpus linguistics*



The visualization makes it transparent the self-evaluation mechanism as an intrinsic design feature. The real-life TSE-TSS checks and evaluates conceptual corpus linguistics which in turn checks and evaluates ideal corpus linguistics, which in turn checks and evaluates practical corpus linguistics. Practical corpus linguistics, on the other hand, has an extra dimension of check-evaluation, namely by practical uses of corpus as well as by the demands of stakeholders. Such check-evaluation, surely very important, is an external one in our conceptualization. McEnery and Hardie (2012: 14-16) address the issues of "total accountability", "falsifiability" and "replicability". These three assessments are mainly targeted at the use or misuse of a corpus. It is also external to our design.

One may note that rectangles in Figure 9 are different in size and arranged in an order of from the largest on the right to the smallest on the left. This is intended to signal the underlying systems thinking as advocated by Bunge (synopsized in 5.2 above).

8. *Final summative reflections*

It is time to round up our journey and take stock. We started with a review of four positions represented respectively by Leech, Chafe, Sinclair and Gu: We witnessed a broadening scope of theorizing corpus linguistics, from about the *language performance*, to about the *mind* of language user, to about the *experiencing participant*, and finally to about the *whole person*.

The whole person, in Bunge's theory of ontology, is the only concrete materialistic thing that can be studied according to the principles and methodologies of factual sciences. As Bunge has argued, logically, linguistics aiming at idealized and abstract phenomenon violates the postulates of factual sciences, hence it is pseudoscientific at best.

Consequently, this paper argues that the fundamental ontological foundation of corpus linguistics is the postulate that there is no such linguistic fact in itself, and that there are only linguistic facts as states of a speaker. This point is worth reiterating: *What is recorded when a corpus compiler switches on a tape recorder is not a linguistic fact unless it is taken as a state or a sequence of states of the speaker being recorded*.

What are the logical conclusions to be drawn from the ontological foundation of corpus linguistics above?

8.1 The Postulate of Speaker Changeability or the Postulate of Language Pseudo-Changeability

The statement that language always changes is pseudo-scientifically valid and misleading. To construct a corpus to map out the state of the art of a language is equally pseudo-scientific and misleading. It is the live speaker that always changes from one state to another.

> "A change is an event or a process, whether quantitative or qualitative or both. Whatever its nature, a change is a modification in or of some thing or things: more precisely, it consists in a variation of the state of an entity. To put it negatively, there is no change separate from things – nor, indeed, are there changeless things even though some change slowly or only in certain limited respects. The world, then, consists of things that do not remain in the same state forever. This metaphysical hypothesis is an extrapolation from both ordinary experience and scientific knowledge." (Bunge, 1977: 215)

This speaker changeability postulate is not difficult to comprehend. Corpus linguists however would find it hard to buy in practice. For it requires that the speaker's state of affairs, physical and mental, be recorded as the primary data source. One may counterargue that this is practically infeasible and unnecessary for most purposes. The counterargument holds water but at its own peril: The corpus data is pseudo-scientific, incomplete and limited in use.

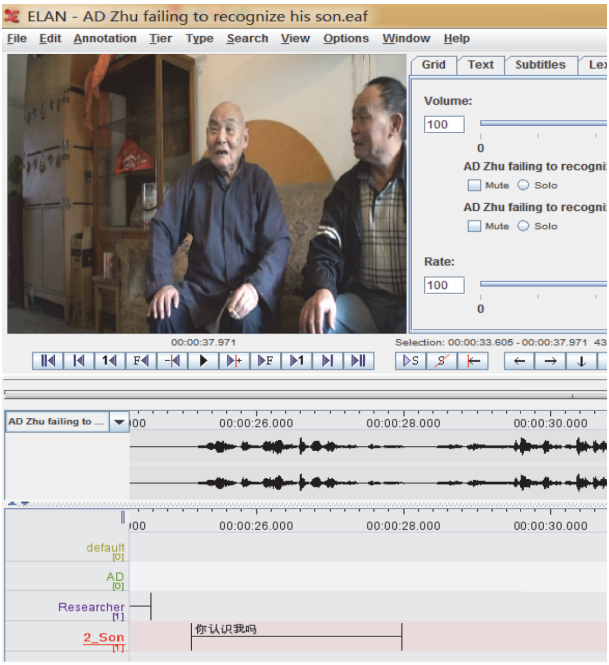
The present author used to hold this counter-argument himself until he embarked upon compiling two specialized corpora, one for autism spectrum disorder children, viz. multimodal corpus of ASD child discourse (MC-ASD-CD), and the other for ageing population, viz. multimodal corpus of gerontological discourse (MCGD). Both corpora have been under construction for a decade now. One big challenge is the issue of privacy protection. Permission to record

is hard to come by. Our solution to the challenge is to form research partnerships with ASD therapy centres, neurology departments of hospitals. In this way trust from patients as well as caregivers becomes much easier to attain. Furthermore, ethics concerning data collection is also secured since it has to gain approval from the institutional ethics committee before the recording starts.

Note that to segment and annotate audio-video streams of two discourse types, one crucial information that must be annotated is the contemporary states of the child or the ageing person. Without this crucial information the corpus linguist would study everything but the child or the ageing person, which is the very purpose of compiling the corpora. To illustrate the point, here is a talk exchange between an eighty-one year old AD (AD), the AD's second son (SS) and a corpus compiler (CC).

– CC: (pointing at SS) Do you know him?
– AD: (looking at his second son) No.
– CC: (pointing at himself) Do you know me?
– AD: (looking at CC) No.

Figure 10 - *A talk exchange between AD and others (quoted from Gu 2015: 466-7)*



One can no doubt do conversation analysis, pragmatic analysis, etc. All these descriptive analyses will only have some bearing on the theories of CA or pragmatics unless we treat talk exchanges substantiating properties of real-life speakers, and ongoing changing states of the mind in particular. Perkins, studying ASD child discourse samples, rightly points out:

> "the child could be described as breaking Grice's Maxims of Quantity, Relevance and possibly Manner ('be brief'), but such descriptive labels do not get us very far when trying to design a remedial programme. One can hardly tell the child to 'stop breaking Grice's maxims'!" (2007: 31)

To tackle the speaker changeability problem, we have collaborated with two major hospitals and turned a walk-in clinic into a research lab. In this way the states of the speaker also include biodata and medical images data.

Finally, we must pre-empt a mis-construal of our conceptualization of corpus linguistics: we may be charged for methodological individualism. This charge bears no tooth, for Bunge's theory of systemism is consistently followed in our whole enterprise. The charge however is understandable for we have not dealt with a crucial property of a system, namely emergence for lack of space.

In a word, the tripartite scheme of conceptualizing corpus linguistics, as outlined above, is hopefully able to produce a linguistic theory bearing its own trade mark. Let us wait and see. Fingers across!

Acknowledge

This paper is part of the projects (Codes 20AYY011, 21&ZD294) funded by the National Social Science Foundation of China.

References

Alonso, Eduardo & Kudenko, Daniel & Kazakov, Dimitar. 2003. *Adaptive Agents and Multi-Agent Systems*. Berlin: Springer.

Biber, Douglas & Johansson, Stig & Leech, Geoffrey & Conrad, Susan & Fingegan, Edward. 2000. *Longman Grammar of Spoken and Written English*. Beijing: Foreign Language Teaching and Research Press.

Bunge, Mario Augusto. 1984. Philosophical problems in linguistics. *Erkenntnis* 21. 107-173.

Bunge, Mario Augusto. 1977. *Treatise on Basic Philosophy Volume 3: Ontology I: The Furniture of the World*. Dordrecht: D. Reidel Publishing Company.

Bunge, Mario Augusto. 1979. *Treatise on Basic Philosophy Volume 4: Ontology II: A World of Systems*. Dordrecht: D. Reidel Publishing Company.

Bunge, Mario Augusto. 2000. Systemism: the alternative to individualism and holism. *Journal of Socio-Economics* 29. 147–157

Bunge, Mario Augusto. 2016. *Between Two Worlds: Memoirs of a Philosopher-Scientist*. Switzerland: Springer International Publishing.

Carter, Ronald & McCarthy, Michael. 2006. *Cambridge Grammar of English*. Cambridge: Cambridge University Press

Chafe, Wallace. 1992. The importance of corpus linguistics to understanding the nature of language. In Svartvik, Jan (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium* 82, 79-97. Berlin and New York: Mouton de Gruyter.

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time*. Chicago: The University of Chicago Press.

Chafe, Wallace. 2018. *Thought-Based Linguistics*. Cambridge: Cambridge University Press.

Chang, Shi-kuo. 2001. *Handbook of Software Engineering & Knowledge Engineering, Vol. 1: Fundamentals*. New Jersey: World Scientific Publishing Co. Pte. Ltd.

Chang, Shi-kuo. 2002. *Handbook of Software Engineering & Knowledge Engineering, Vol. 2: Emerging Technologies*. New Jersey: World Scientific Publishing Co. Pte. Ltd.

Chang, Shi-kuo. 2005. *Handbook of Software Engineering & Knowledge Engineering, Vol. 3: Recent Advances.* New Jersey: World Scientific Publishing Co. Pte. Ltd.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Westport: Praeger.

Dillinger, Mike. 1990. On the concept of 'a language'. In Weingartner, Paul & Dorn, Georg J.W. (eds.), *Studies on Mario Bunger's Treatise*, 10-38. Amsterdam: Rodopi.

Fodor, Jerry A.. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: A Bradford Book.

Gibson, J.J.. 1986. *The Ecological Approach to Visual Perception*. Hillsdale, New Jersey: Lawrence Erlbaum Associations, Inc., Publishers.

Goffman, E.. 1963. *Behavior in Public Places*. New York: The Free Press.

Greenough, William T. & Black, James E. & Wallace, Christopher S.. 1987. Experience and Brain Development. *Child Development*, 58. 539-559.

Gu, Yueguo. 2006a. Multimodal text analysis: a corpus linguistic approach to situated discourse. *Text and Talk*, 26-2. 127-167.

Gu, Yueguo. 2006b. "Agent-oriented modeling language, Part 1: Modeling dynamic behavior" (基于角色的建模语言(AML)一：动态行为建模). In *Proceedings of the 20th International CODATA Conference, Beijing*, 21–47. Published by the Information Centre, the Chinese Academy of Social Sciences.

Gu, Yueguo. 2009. From real-life situation to video stream data-mining. *International Journal of Corpus Linguistics* 14:4, 433-466.

Gu, Yueguo. 2013. A conceptual model of Chinese illocution, emotion and prosody. In Tseng, Chiu-yu (ed.), *Human Language Resources and Linguistic Typology*, 309-362. Taibei: Academia Sinica. Pp.

Gu, Yueguo. 2015. Multimodality and linguistic research. (多模态感官系统与语言研究). *Journal of Contemporary Linguistics*《当代语言学》Vol. 17, No. 4, 448-469.

Gu, Yueguo. 2016. Multimodal experiencing, situated cognition and big data with a demonstrative analysis of a newborn baby. 当下亲历与认知、多模态感官系统与大数据研究模型。*Contemporary Linguistics*,《当代语言学》第18 卷第4 期475－513 页。

Gu, Yueguo & Xu, Xunfeng. 2013. Alzheimer's disease patient discourse: A multimodal corpus linguistics approach. *Plenary speech at the 5th Symposium on Functional Linguistics and Multimodality*. The Polytechnic University of Hong Kong.

Hexmoor, Henry & Castelfranchi, Cristiano & Falcone, Rino. 2003. *Agent Autonomy*. Boston: Kluwer Academic Publishers.

Leech, G.N.. 1992. Corpora and theories of linguistic performance. In Svartvlk, Jan (ed.), *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991,* 105-22. Mouton de Gruyter, Berlin/New York.

Legge, James. 2008 [1891]. *Tao Te Ching* by Lao Tse. The Floating Press.

Linell, Per. 2005. *The Written Language Bias in Linguistics: Its Nature, origins and Transformations*. London: Routledge Taylor & Francis Group.

McEnery, Tony & Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Perkins, Michael. 2007. *Pragmatic Impairment*. Cambridge: Cambridge University Press.

Russell, Stuart J. & Norvig, Peter. 2003. *Artificial Intelligence*. New Jersey: Pearson Education, Inc.

Sinclair, John. 2004. Carter, Ronald (ed.), *Trust the Text: Language, Corpus and Discourse*. London: Routledge Taylor & Francis Group.

Skyttner, Lars. 2001. *General Systems Theory: Ideas and Applications*. Singapore: World Scientific Publishing Co. Pte. Ltd.

Wagner, Thomas A.. 2004. *Application Science for Multi-Agent Systems*. Boston: Kluwer Academic Publishers.

Ye, Yiming & Churchill, Elizabeth. 2003. *Agent Supported Cooperative Work*. Dordrecht: Kluwer Academic Publishers Group.