Methodology and Trends of Linguistic Research in the Era of Big Data

Liu Haitao and Lin Yanni*

Zhejiang University

Abstract: This paper presents methodology and trends of linguistic research in the era of big data. We begin with a discussion of the role of linguists in the information society and illustrate the opportunities and challenges linguists are currently facing. After highlighting the significance of authentic data on linguistic research, we argue that language is a complex adaptive system driven by humans. Then, from the perspective of philosophy of science, we introduce the research paradigm of quantitative linguistics through several cases. Finally, we discuss how China's linguistic research will benefit from the data-intensive approach in terms of scientification and internationalization.

Keywords: linguistics, big data, the data-intensive approach, scientific research paradigm

Introduction

The step from the industrial age towards the information era started in the second half of the 20th century. Today, the issue of information explosion is increasingly standing out amid the global information wave as we live in a world surrounded by an unprecedented amount of information. The

* Liu Haitao, School of International Studies, Zhejiang University; Lin Yanni, School of International Studies, Zhejiang University.

This paper is a phased achievement of "A Study on Quantitative Linguistics: Contemporary Chinese Language" (11&ZD188), a major project sponsored by the National Social Science Fund of China and implemented by Zhejiang University's "Big Data + Language Laws and Cognition" innovation team under the auspices of the Fundamental Research Funds for the Central Universities.

Correspondence concerning this article should be addressed to Liu Haitao, School of International Studies, Zhejiang University, Zhejiang. E-mail: htliu@163.com

urgent need to process massive amounts of information drives us to think how computers can help undertake these burdensome tasks, such as information extraction and machine translation, thereby allowing people to focus on more important things. In this context, computational linguistics and natural language processing have emerged as booming disciplines.

Yet, doubts about linguists are often heard in these promising disciplines. For example, Frederick Jelinek, National Academy of Engineering (NAE) fellow and natural language processing expert, purportedly said, "Every time I fire a linguist, the performance of the speech recognizer goes up." (Hirschberg, 1998; Jelinek, 2005)[①] While his statement may be joking in some sense, there is a non-ignorable fact: few linguists can be found in the computational linguistics and natural language processing community. Linguistics, a discipline fundamental to language studies, is supposed to be helpful and instructive to linguistic practice and applications in such a community that looks to process languages. But why do linguists suffer such harsh treatment? How to make linguists play a role today? These questions have driven us to reflect on the origin of the relationship between language studies and the information era. As the information era evolves, big data, which is characterized by "4Vs" (Volume, Variety, Velocity and Value) (Chen & Xu, 2015), has contributed to changing the ways of social life and thinking, developed into a new research paradigm (Li, 2015) and led to many new findings in the natural science and humanities and the social science community. That is to say, the information era presents both challenges and opportunities to language studies.

This study seeks to address the following questions focusing on big data-based language studies. How have language studies changed in the information era? Can data-based approaches provide inspiration for language studies? What is the view of quantitative linguistics as a data-based branch of linguistics? What provides a scientific basis for the research paradigm of quantitative linguistics? How can language studies be conducted using data-intensive approaches? As the "Double First-class" initiative (the World First Class University and First Class Academic Discipline Construction) is underway, what role can data-intensive approaches play in developing the discipline of linguistics?

Among the follow-up sections,

Section I describes the shift of language studies in the information era.

Section II discusses data-intensive approaches and related issues.

Section III introduces a few data-based language studies.

Section IV addresses studies in and the development of the discipline of linguistics.

Section V is titled "What's More."

The Shift of Language Studies in the Information Era

This section describes the shift of language studies in the information era. The first part explains how a world-renowned linguist moved from "garden" to "bush" to highlight that contemporary

① For the origin of this statement, see https://en.wikipedia.org/wiki/Frederick_Jelinek#cite_note-6.

language studies must focus on authenticity of linguistic data and go beyond traditional research methods. The second part demonstrates that big data will present new opportunities to language studies. The third part describes the definition and views of quantitative linguistics, a data-based branch of linguistics.

The Shift of Language Studies: From "Garden" to "Bush"

In August 2016, Joan Bresnan, who proposed Lexical Functional Grammar (Bresnan, Asudeh & Toivonen, 2015), was granted the Lifetime Achievement Award by the Association for Computational Linguistics. Her testimonials were subsequently published in *Computational Linguistics* (2016 (4)) under the title "Linguistics: The Garden and the Bush" (Bresnan, 2017). In the article, Bresnan recalled how she moved from "garden" to "bush" to argue that most traditional linguistic theories essentially deviate from social requirements for linguistic theories. As "the garden", traditional linguistics, which encompasses generative grammars, focuses on linguistic phenomena where linguists carefully select, or cultivate through introspection, and qualitatively generalize them using symbols like syntax trees and phrases. By comparison, "the bush", also called "linguistics in the field", focuses on the language that people actually use in daily communication, featuring quantitative analysis based on conditional probability and information content. The tools and methods used in the garden are very likely to fail when trimmed and delicate flowers are replaced by a dense wild bush.

Bresnan also recalled how, as a doctoral student at MIT in the 1960s, she learned from Noam Chomsky when the doctoral advisor's ideas attracted the whole world. As language is viewed as a set of formal patterns, it is certainly exciting to analyze the structures specific to human language and explore human language and mentality. Excitement inspired many people at that time. For example, an engineering doctoral student who entered MIT ten years earlier than Bresnan even planned to leave his information theory major for linguistics, but failed to do that before completing his study for an information theory doctorate as his supervisor disapproved (Jelinek, 2009). This man was Jelinek, who later threatened to "fire linguists." A puzzling question is, what turned the hot-blooded youngster, enthusiastic about theoretical (formal) linguistics, into a stern-faced boss threatening to do that over dozens of years' development of linguistics? The answer may be the research data and methods used by mainstream linguists. As mentioned above, natural language processing faces authentic, diversified languages like a bush growing in nature. It is difficult to reveal patterns of authentic languages using a few selected sentences like cultivating flowers in the garden.

Both traditional linguistics and modern linguistics study human languages. No matter whether linguists are prepared, a linguistic view has arisen as the information era arrives. Language, which represents a major carrier of information, should be studied in ways that meet both human and computer demands. Natural language processing needs to deal with authentic linguistic data, which has a most distinctive characteristic: probability. That is to say, instead of being either "grammatical" or "ungrammatical", an authentic language falls in between." In general, scientific research involves abstract modeling. Features of a model represent observable attributes of what is modeled. A theory

interprets the real world in an indirect way by leveraging a model created through abstraction and what the model represents. As such, the predictive power of a theory relies on the correspondence between the model and reality. Findings from a model that ignores the essence and fails to reflect the true colors of what is modeled are hardly usable. This may be an important reason behind the fact that most linguists are abandoned by the computational linguistics community. While this should not be used as the sole criterion for evaluating the significance and value of linguistics, Bresnan's movement from "garden" to "bush" indicates that language studies may have to embrace a significant shift in the information era.

Undoubtedly, the method and theory of formal linguistics that Chomsky proposed in the 1950s led to a revolution in linguistics. However, theoretical and practical language studies in the past dozens of years have probably indicated the need for new shifts. As regards studying objects, more attention should be paid to authentic linguistic data and the relationships between humans and language systems. As to research methods, in alignment with the characteristics of authentic linguistic data, statistical technologies and research methods should be leveraged to make up for the shortcomings of introspective and qualitative methods. As concerns model selection, a model should be of cross-linguistic validity and not limited to a specific language—as what linguistics studies is human languages, linguists should focus more on how to find the universalities of human languages, thereby preventing themselves from falling behind the current era.

New Opportunities Presented by Big Data to Language Studies

While the information era challenges language studies, it also presents new opportunities for the shift discussed above, which gives higher priority to the shift from introspective methods to data-driven methods. "Data-driven" means that language studies may encompass or suit another feature of the information era as we often hear—big data. In fact, the name "big data" is inexact as big data features wide variety, rapid processing and low value density in addition to large scale (Chen & Xu, 2015). No matter what it is called (for example, the name "thick data" was given lately), big data points to the fact that we live in a period in which data is extremely easy to obtain. Linguists should put more emphasis on "data" as a characteristic of this era and focus more on where data-driven language studies are heading rather than data quantity. In other words, we should give more prominence to which linguistic issues can be addressed with data, or which language patterns and mechanisms we previously failed to take note of or were unable to study can be discovered using data. In this connection, what data offers us is a research paradigm, a method and tool for observing and studying objects.

First of all, by providing an instrument for quantitative research, data-based methods enable us to get a clearer, more precise and finer picture of the objects we are studying. When you observe something, what you see and perceive will vary with distance (zooming-in, zooming-out) and perspective (microscopic, macroscopic). More authentic linguistic data helps reveal the profile of a language more deeply and truly. Also, data-based methods can be used to reveal some essential characteristics of languages, including

probability (Bod, Hay & Jannedy, 2003). For example, in scenarios where the introspective method is applied, sentences marked with "*" are ungrammatical or unacceptable in the view of native speakers. However, such sentences are often used in daily life. As many studies suggest, instead of being either "acceptable" or "unacceptable", a language that people understand or produce falls in between." In case the reasonableness of a statement is difficult to describe, massive linguistic data makes it possible to define the grammatical acceptance of the statement more accurately. The capability of data-based methods is to reveal what a language is truly like, and all about echoes the last sentence in the preface to Bernard Comrie's *Language Universals and Linguistic Typology*: "Linguistics is about languages; and languages are spoken by people" (Bernard, 1989).

Alongside that, data allows us to more closely study the relationships between human languages and human cognition. Language is a symbol system, yet many previous studies only analyzed symbols in the purely formal ways that separate humans from language. In fact, language is a human-driven semiotic system, or more exactly, a complex human-driven adaptive system. How a language is structured and evolves is the product of a mix of internal factors (e.g. physiology, psychology, cognition) and external ones (e.g. nature, society)—the universality of internal factors leads to language universals, and the differences of external factors result in language diversity. On the one hand, language universals are partly attributed to the universality of cognition. For example, recursion is considered as an essential property of human language (Hauser, 2002), but it is not infinite—recursions over three levels are seldom used in practice (Sampson, 1997; Karlsson, 2010). It is improper to equate humans with machines as humans are restricted by cognitive factors. On the other hand, as everyone lives in a specific natural and social environment, natural, social and cultural factors may influence language in ways that help the world embrace diversified languages. As such, extensive data collected from real scenarios of language use enables us to better discover and interpret the universality and diversity of human languages.

Quantitative Linguistics: A Branch of Linguistics that Cannot Develop without Data

Since linguists study linguistic phenomena as much as possible like physicists study physical phenomena, it is the language engineers' task to figure out how to use the insights of linguists as engineers benefit from physicists' insights, Jelinek (2005) argued in a later article. In other words, physicists work to discover rules of the physical world, and linguists study how a language is structured and evolves. Then why are achievements of language studies hardly used in practice of natural language processing? The answer relates to accuracy and scientificity of language studies, as well as the issues of research resources and methods mentioned above. As a scientific approach to discovering laws of language systems, quantitative linguistics is a branch of linguistics worth advocating.

Built on quantitative methods, quantitative linguistics provides quantitative analysis and dynamic descriptions of various linguistic phenomena, language structures, structural properties and their relations in ways that reveal relations, positions, mechanisms and profiles of various linguistic phenomena. In so doing, it seeks to explore self-adaptive mechanisms of language systems and

motivations for linguistic evolution in an effort to make language studies more accurate and scientific (Liu, 2017).

What are the connections and differences between quantitative linguistics and "mainstream" linguistics in the traditional sense? As with the other branches of linguistics, quantitative linguistics aims to explore structures and patterns of languages. However, it differs from traditional linguistics in linguistic view, source of data, research methods and levels of abstraction. In many cases, driven by specific issues related to a linguistic phenomenon, traditional linguistics analyzes specific examples or uses leveraging language intuitions, and seeks to explore the rules of language structures using introspective methods, and more or less, formal methods, as a step to study how the brain processes languages. In contrast, identifying a language as a complex adaptive system, quantitative linguistics makes use of authentic linguistic data and relies on quantitative methods to explore how a language is structured and evolves. In short, it features precision, authenticity and dynamics. The difference in levels of abstraction is another noteworthy point between quantitative linguistics and traditional linguistics. Quantitative linguistics looks to build a model that enables discussing a language system in a more abstract way. To this end, while authentic texts are used, the branch focuses less on specific words, phrases or sentences. Compared to ontology-based linguistics, it discovers and reveals the laws of a language in a way closer to the way in which physicists discover the laws of the physical world.

True, starting from a specific language structure is also interesting. It makes no sense to argue whether the linguistic view of quantitative linguistics is better or worse. Both quantitative linguistics and ontology-based linguistics work to explore the patterns of a language, though they rely on different methods. As human language is a very complicated and dynamic system, we may need to leverage the strengths of different methods in ways that help extensive exploration of a language system to figure out how this system operates and evolves, thereby gaining a more comprehensive and complete picture of that language system.

Linguistic View of Quantitative Linguistics: Language is a Complex Adaptive System

Quantitative linguistics takes a language as a complex adaptive system—a linguistic view that disrupts traditional views. As linguists represented by Saussure have early put forward the view that language is a semiotic system, language has long been considered as a semiotic system that may run independently of humans. The theory of complex adaptive systems, which first appeared in Holland's *Hidden Order* (Holland, 1995), features a core idea: individuals' adaptability leads to system complexity. Guided by this theory, complex network-related approaches like genetic algorithms, neural networks and evolutionary game theories have been introduced into social system studies over time (Miller & Page, 2012). In recent years, by looking at linguistic facts, some linguists have reported that language is, in fact, a complex adaptive system (Wang, 2006; Kretzschmar, 2015; Ellis & Larsen-Freeman, 2009).

In the system science community, "system" is defined as a whole constituted by its components and their relations. As implied by the philosophy that motion is absolute, a real system must face

various disturbances from the environment or itself (Xu, 2000). That means a system normally is dynamic and runs to achieve a functional goal. Language accords with such a statement: as a dynamic system, language runs to perform its major function of serving communication. Also, it plays the roles of culture container and social status symbol. Components of a language system need to collaborate in lexicon, syntax and semantics under the least effort principle, in ways that optimize communication. Yet, language was regarded as a static system in many previous studies. In fact, there are qualitative differences between "dynamic" and "static".

"Complex" mainly means the overall behavior of a system cannot be equated to the sum of behaviors of its components. That is to say, a system is of emergence. When it comes to language systems, take a sentence comprised of five words for example. Simply piling up the lexical senses of the five words cannot always lead to the meaning of the entire sentence. The fact that the whole is unequal to the sum of its parts is a major feature of complex systems in the real world (Solé, 2008). Alongside that, a complex system features uncertainty, indeterminacy and randomness (Morin, 2008). In a sense, complexity always relates to uncertainty or probability.

"Adaptive" qualifies a goal-defined dynamic system. A language system is adaptive, which means it may create a new structure, state or function through self-organization so as to adapt to certain external environments (Xu, 2000). An adaptive system features a self-regulating mechanism that maintains system balance, which is also true for language. For example, we may abstract closely related attributes of words, such as frequency, length, polysemy and compositeness, from the lexical system of a language. Statistics show that in a balanced lexical system, a high-frequency word is short in general, but not absolutely. As mentioned above, language is a "bush" varying naturally. If the occurrence of a low-frequency word suddenly rises, subsystems of the lexical system will respond collaboratively to enable the word to spontaneously and temporarily shorten in ways that meet communicative needs. This is a good example of system adaptability.

As a complex adaptive system, language co-evolves with humans. As mentioned earlier, language is a human-driven system. Humans, as users that feature sustained development, drive the continuous evolution of language systems. Human-related internal factors (e.g. physiology, psychology) and external factors (e.g. nature, society) influence language universals and diversity. This is why we should not study linguistic phenomena separately from humans.

With language being taken as a system, it is tempting to consider studying languages using the methods for studying common systems. Therefore, such studies involve carefully observing linguistic phenomena and exploring the components, structures, processes, behaviors, functions and environments of a language system. These studies cannot be performed without authentic linguistic data or data from language behavior experiments.

Data-intensive Approaches to Language Studies and Related Issues

The previous discussion regarding the definition and linguistic view of quantitative linguistics

clearly indicates that quantitative linguistics pursues precision, which mirrors the nature of language as a branch of science. This section interprets the research paradigm of quantitative linguistics from the perspectives of the philosophy of science and addresses a few data-related issues arising from language studies.

Necessity to Adopt Scientific Research Methods

The scientific research paradigm is essential for discovering the laws of a language. Philosophy of science, a sub-field of philosophy, specially defines what science, theory and the scientific research paradigm are. Today's scientists hold that scientific research must be conducted using scientific methods. While the concept that "linguistics is a branch of science" is accepted by most linguists, linguistics has yet to be widely recognized by the science community. A reason behind that is linguistics' failure to fully recognize and comply with the scientific research paradigm. It makes no sense for linguistics to go against scientific methods while being recognized as a branch of science.

This does not mean that the traditional data-free practice is improper. Anyone who is serious about his or her research is respectable. That being said, it is a reckless waste to put on the shelf the massive data and new ways of data operation we now have access to. More importantly, data may help us make discoveries. A good example in daily life is photography. Photos of the same scene taken using different lenses (tele-photo, standard, wide-angle, fisheye) give you different feelings. When you look at the same thing, what you see through a microscope is very different than through a telescope—to those without such experience, the inspirations gotten from it are beyond imagination. So, can our perceptions of language change as we now have access to more data? Why not introduce microscopes and telescopes, which are easily available now, into language studies?

Research Paradigm of Quantitative Linguistics

As mentioned before, quantitative linguistic studies rely on a data-intensive research paradigm, featuring precision, authenticity and dynamics. "Precision" refers to using mathematical means to quantify a language; "authenticity" means focusing on authentic language used in daily communication; "dynamics" points to taking a language as a changing complex adaptive system. That is to say, the methods adopted in quantitative linguistics are closer to those employed in the natural science community.

Quantitative research on language is time-honored but has yet to grow into a systemic discipline. German academician Gabriel Altmann started to systemically study the relations between linguistics and philosophy of science in the 1960s. By analyzing many cases, Altmann developed a detailed plan that outlines the theoretical framework of modern quantitative linguistics, in full alignment with practices of philosophy of science. Alongside that, he summarized the research paradigm of quantitative linguistics into five basic steps:

1. Making an empirical, falsifiable hypothesis.
2. Expressing the hypothesis in statistical language.

3. Finding a proper statistical method to test the hypothesis.

4. Determining whether to accept the hypothesis in alignment with test results.

5. Interpreting the results.

This research paradigm is what we call "a research paradigm aligned with concepts of philosophy of science." As American academician David Eddington (2008) wrote in his article "Linguistics and the Scientific Method", authentic language can be effectively interpreted only by scientific methods. Progress in linguistics is only made to the extent that linguists adopt the scientific method that is standard in scientific endeavors—observing a phenomenon, formulating a hypothesis, collecting data, verifying the hypothesis, and drawing a conclusion. These steps constitute what we call the empirical research method.

In the current era, the first thing to consider in a data-intensive language study is which issues are to be addressed with data, or whether there is any issue that must be addressed using data. In general, this consideration involves two types of scenarios. One type is hypothesis-driven scenarios where a conclusion is drawn in steps from hypothesis formulation through data collection to hypothesis verification. The other is data-driven scenarios: while no hypothesis is formulated, the model represented by acquired massive data is analyzed in ways that enable the rules of such model to be discovered and interpreted. In fact, hypothesis verification also needs data. While introspective methods are given priority by mainstream linguists, if methods widely recognized by scientists are employed to verify a hypothesis and make up for shortcomings of introspective methods, we may draw more convincing conclusions.

As to the scientific research paradigm, Chinese Academy of Engineering (CAE) fellow Li Guojie wrote in the preface to *Uncharted: Big Data as a Lens on Human Culture*: "data-intensive scientific research has developed into 'the fourth paradigm' in parallel with scientific experiments, theoretical analysis and computational simulations…the significance of big data to the transformation of social science is as considerable as that of aiming a telescope into outer space for the first time by Galileo is to astronomy" (Aiden & Michel, 2015). So far, this data-intensive paradigm has enabled scientists to make many interesting discoveries across a number of sectors (Hey, Tansley & Tolle, 2009).

A Few Issues Concerning Data-intensive Approaches

A few issues concerning the data-intensive research paradigm are addressed below.

Characteristics of quantitative research methods for the big data era.

Quantitative methods are not new to language studies. While previous quantitative studies on language were also aimed at discovering language laws, limited linguistic examples were found through traditional technical means, such as card-based collection. Today, you can easily get linguistic data on a connected computer. Data size wise, as almost everyone among the world's billions of people speaks every day, collecting all of their words definitely means massive data to capture. Massive data and operating technologies help reflect how language is used across different scenarios and deepen our understanding of language, benefiting today's linguists. However, this does not mean

"more massive is better." When a corpus reaches a threshold, its function of revealing patterns may not increase with its size. And, it is technically difficult for academicians of liberal arts to handle massive data. As for modeling, statistical models for quantitative research feature a verification-driven approach where a hypothesis is first formulated and then data is used to verify the reasonableness of the hypothesis. In contrast, big data models are data-driven, highlighting modeling processes and model updatability (Li, 2015). This is a significant difference, but not an essential difference for language studies, as data cannot fully replace humans. What we need to think about is how to make a more scientific interpretation on the basis of data and use data to answer questions about patterns and mechanisms of linguistic structures and evolution.

Dispute between two data views: can data speak?

The big data era marks two data views. One is data can speak in a way neither reliant on nor sensitive to humans. The other is, data cannot speak for itself and it is humans that speak for and give meaning to data.

First, data is speechless but used by humans for speaking. For example, meanings of "1" and "2" vary with the scenario. Such variability is only understandable to humans. "Data can speak" means data may make your speech better founded. Quantitative and data-based methods provide more scientific means to verify previous hypotheses and more effective ways to find models hardly discoverable in the little data or data-free era. However, data will be useless if one knows nothing about what he is going to study. All processes involving data, especially advanced research activities like discovery, analysis, generalization, interpretation and prediction, entail people's active participation, which cannot be replaced by machines. Thus, what represents the true value of big data is not data. What really works is the connection of data to knowledge, society, culture, humans and their behaviours, and the use of more scientific means of mathematical statistics to discover cognitive and behavioral models and the mechanisms by which humans interact with society and nature.

Second, while data is neutral, people observe and abstract the real world in a selective way. This relates to a common issue in modeling. Take treebank annotation for example. It is inevitable that the process of annotating is either reliant on intuitive analysis or sensitive to existing linguistic theories. To analyze the syntactic structure of a sentence, you need to identify the subject, object, adverbial modifier and other elements through your brain's cognitive mechanism and language system and annotate them. The annotating process, which is basically the process that you use to transfer language knowledge to a machine, reflects your understanding of the sentence and its syntax. With sufficient amount of annotated sentences, a machine may abstract syntactic knowledge of the language. This involves an issue: different persons may analyze the same sentence in different ways. As syntax models consist of the dependency grammar that addresses relations between words, the phrase structure grammar that addresses the relations between the parts and the whole, and the syntactic framework that combines them. Every syntactic model contemplates the syntax abstraction and modeling process in human language. Just like what is done in other fields of science, linguists must build a model by abstracting the real world and study

the model. Such an abstraction, like all abstractions in scientific research, is a trade-off that cannot cover all factors. Yet, the model is only required to reflect the main features of what is modeled. The annotation process following modeling may see disputes about linguistic phenomena, as linguistic intuitions vary among individuals. So, it is tempting to argue which ways of annotation is more reasonable. In practice, in case disputes arise from the same phenomenon, patterns and trends will not be affected badly as long as one annotation scheme is defined. Alongside that, what may lead to disputes plays a small part in the annotation process and the broader language system.

A further question may be asked: If we leave the disputed data alone, will the whole study be affected? Generally speaking, no. Language, as a dynamic complex system, is naturally in balance. That means language normally enables our basic communications. However, a language that provokes disputes about structure and components is unstable, and unusable in communication. That is to say, what is disputed plays a small part in and cannot affect the whole of a language. While language, as a dynamic system, is changing. Its core is stable and underpins its role as a vessel of communication. Such stability enables us to study the core of a language system in a scientific manner. Take part-of-speech tagging for example. If there are 10 words unclear in characteristic or attribute among 10,000 words, the 10 words may be temporarily left alone as rules are most possibly implied by the remaining 9,990 words. In short, in language studies, we must treat a language as a system, and refrain from being entangled with one or two words. Such a practice may deviate from traditional practices of analysis. Another point that should be kept in mind is as language is a complex adaptive system, most laws of language are probably statistical.

A few misunderstandings about big data.

The book *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Mayer-Schönberger & Cukier, 2013) once sold well. Perhaps for the purpose of promotion, the book's core content was simplified into three short slogans: "Analyze vast amount of data rather than settle for small sets. Embrace data's real-world messiness rather than privilege exactitude. Respect for correlations rather than elusive causality." It is noteworthy that "not", "less" and "from" in the disputed slogans are not meant to call for discard, but to highlight the shift of focus and the need to transform our ways of thinking and processing.

Analyze vast amount of data rather than settle for small sets: Previously, random sampling was essential for acquiring maximum information based on minimum data as the then technological means have limited capacity of data processing. Today, random sampling seems to have been essential eliminated as machines, software, hardware and other technological conditions are increasingly sophisticated, making it possible to process all big data. That being said, random sampling still can be carried out in alignment with what is studied.

Embrace data's real-world messiness rather than privilege exactitude: Statistics seek to reveal trends, rather than pursue privilege exactitude. What statistics require is no more than to reveal data patterns and trends by efficiently and rapidly processing data on a computer. At the heart of big data is forecast. For example, built on the patterns and trends revealed by processing meteorological big

data on computers it is possible to forecast a rainfall to arrive in a region in about five hours and alert people to take an umbrella while going out. This makes it unnecessary to announce the rainfall time accurate to the second. A big data model does well in forecasting, but has no function for deductions, so it is different from physical laws that pursue certainty. That doesn't mean it is unscientific—big data models and physical laws have their respective scopes of application. Currently, exactitude is far from being identified as a strict criterion for big data models (Li, 2015).

Respect for correlations rather than elusive causality: This slogan has provoked considerable disputes. As we know, academic research, which features rationalism, pursues causality. To this end, the following questions may be asked: Can a study matter in any way if it doesn't target causality? If so, can data make any sense? As big data seeks to build models that underpin forecasts about purchasing behaviors, weather and the spread of epidemic diseases, among others, and only pursues solutions, does it totally ignore causality? No. If the relations between two factors are so simple that causality can be easily revealed, academicians will naturally look to explore the causality. However, in the many cases that involve complicated human and social factors, while big data enables us to discover correlation, it is very difficult to reveal causality. For example, the causality "smoking is harmful to your health" has been revealed by spending enormous efforts and resources. Many instances that cannot be reproduced through behavioral experiments have demonstrated that causality involved with humans and society can hardly be cleared up in a short time—most of such systems are nonlinear, but causality is more like a feature of linear systems (Solé & Goodwin, 2008). In our view, as causality is a type of correlation and the chance of correlation implies the necessity of causality, it is not a must to pursue causality when correlation is adequate. Big data helps discover causality, or at least, enables approaching causality on the basis of correlation.

As most of the patterns previous data-based language studies have discovered are reproducible, another minor question may be derived from causality: How do these patterns matter to causality-oriented language studies? As we know, most of causality-oriented language studies are curiosity-driven. Similarly, researchers using big data are also curious. Any study—whether it uses big data or little data, whether it uses data or not—cannot be conducted without curiosity. As Li Guojie wrote, the data-intensive research paradigm is a tool. When people observe outer space using a telescope, they can perceive the fantasy in the depth of the universe that they could hardly imagine with the naked eye. Such perception only makes them more curious. A tool enables us to discover patterns previously invisible to us, and these discoveries may arouse our curiosity about why they have been shaped. Serving as a driving force for every academic study, curiosity may lead us to insights into the reason behind a linguistic phenomenon, or broadly speaking, to explore causality.

A Few Data-based Language Studies

The sections above interpret methods for language studies in the information era. This section

leverages a few research findings to explain how to conduct data-based language studies.

Dependency Distance Minimization (DDM) Studies

This part focuses on DDM studies. Dependency grammar is a theory built on relations between words (Liu, 1997, 2009). In a sentence, words constitute a line array where two words involved in the same syntactic relation are next to each other or spaced apart. Dependency grammar defines the linear distance between two interdependent words as dependency distance, which is generally measured by the number of words in between. By dependency distance, we analyzed some sentences that had been studied by psycholinguists and found that sentences considered difficult in psychological experiments feature a long dependency distance. The finding suggests that dependency distances may be influenced by psychological and cognitive factors, such as working memory. This enables text analysis indicators to connect to human cognition mechanisms. In other words, this makes it possible to study human cognition through dependency syntax analysis. On the premise that dependency distances are influenced by working memory, dependency distances of all languages should be similar as language is universal among cognition and restricted by cognitive rules, as mentioned above. More than 10 years ago, we conducted a further study based on authentic materials of 20 languages in this regard (Liu, 2008). The study, which marked the world's first large-scale DDM study using cross-linguistic authentic materials, clearly showed that average dependency distances of ten-odd languages are nearly the same, and human language features a shorter dependency distance than the reconstructed non-human random language. The study results demonstrated our assumption that DDM may represent universality of human language. DDM reveals a pattern previously invisible to us, which is characterized by general features of human language and brings (big) data into play.

DDM, as a possible general feature of human language, tends to be considered mediocre. Some academicians who have less knowledge about the principles of DDM may recognize it as a proof to the universal grammar, which was proposed by Chomsky. So, it is worthwhile clarifying the differences between DDM and the universal grammar. As Chomsky argued, the universal grammar is an innate mechanism of the human brain that decides the universality of human language. Yet, as our study suggests, DDM results from limited working memory capacity, for which people seek to minimize dependency distances in the course of linearized sentence-generalization. Working memory, as part of the human cognition system, is not language-specific. In other words, the characteristics of DDM are restricted by human cognition mechanisms. This by no means proves the existence of a biological mechanism working for language or the universal grammar. That is to say, DDM in no way demonstrates the existence or non-existence of the universal grammar.

The last ten-odd years saw our team work to deepen the understanding of DDM by studying issues like "why Chinese seems to unaware of their mother tongue's difficulty though it features a long dependency distance." Using extensive authentic linguistic data across languages may help reveal general features of language we previously ignored.

Typological Studies Based on Dependency Direction

This part addresses word order typological studies based on dependency direction. Dependency grammar analysis involves three factors: head, dependent, and dependency relations. In a sentence, the head is located either before or behind the dependent, shaping two different dependency directions, head-final and head-initial. Using the indicator of dependency direction proportion, we examined the dependency directions of 20 languages. As we found by studying extensive authentic linguistic data, dependency direction can serve as an indicator of word order type. Word order types constitute a continuum where any language can find its position and cluster analysis can be conducted in alignment with dependency distance (Liu, 2010). For example, while it was once a common practice to identify a language as "SOV language" or "SVO language", every language may contain SOV elements—some contain more, and some less. The new finding built on (big) data has furthered our understanding of linguistic typology.

Language Production Mechanism Studies Based on Dependency Distance

This part highlights language production mechanism studies conducted from the perspective of system. Language, as a complex adaptive system, is concerned with adjustments that seek to allow people to communicate with each other more easily by minimizing dependency distances in sentences. For example, dependency distances in a three-word sentence are short, but those in a 30-word sentence may be long. A long sentence triggers the adaptive mechanisms of language, which minimizes its dependency distances. In the course of adjusting, the adaptive system must work in a targeted way. As such, definite settings are essential for studying language from the perspective of system. With DDM defined as the target or setting of sentence linearization, what would people do to process a long sentence? As we revealed through computer simulations based on a corpus tagged with authentic linguistic data, a dynamic linguistic unit, chunk, is very likely to be generated in the course of long sentence processing. Chunks help considerably shorten the average dependency distance of a long sentence and allow for DDM (Lu, Xu & Liu, 2016). The finding represents the result of exploring the language production mechanism from the perspective of system.

The above findings have been obtained through data-based verification and exploration, which has made us better aware of language patterns and processing mechanisms. They demonstrate that data-intensive language studies are feasible and can help us discover language patterns and rules previously invisible to, and problems previously unsolvable for, us.

Ideas About the Construction and Development of the Linguistics Discipline

This section pushes a few ideas about the development of linguistics as a discipline, focusing on the role of data-based or data-driven approaches. First, it is argued that teaching within the discipline should mirror what the current era features and society requires. The second part focuses on how

data-based or data-driven research methods enable language studies in China to be more scientific and step into the international arena in the context of the "Double First-class" initiative. Finally, interdisciplinary language studies are addressed.

Curriculum and Content of Courses Satisfying Demands of the Times

As Jelinek (2005) said about the roles of linguists, the natural language processing community has been yearning for linguists' help, or more exactly, for linguistic knowledge that combines linguistics and data-driven statistical methods to enable machines to better understand or process human language. There is a statement that echoes Jelinek's argument: "Every time you hire a well-trained linguist, your treebank will get better" (Eberhard-Karls-Universität, 2005). Today, extensive linguistic data available for training is needed in most natural language processing tasks built on statistical machine learning and deep learning tasks built on neural networks. Linguistic data given syntactic or semantic information enables machines to better learn syntactic and semantic knowledge and process human language more effectively. Such corpus annotated with syntactic or semantic information is called a "treebank", which serves as the knowledge source in machine learning. It is noteworthy that the world's earliest large-scale treebank was built with support from Jelinek (2009), who looked to generalize grammars that helped automatic speech recognition (Jelinek, 2005). As such, you may think that linguists make contributions by annotating treebanks. However, not every linguist is able to fulfill a annotating task, as it can only be performed by a well-trained linguist who is at least aware of mainstream analysis methods adopted in the natural language processing community. Take syntax for example. Given practice in the natural language processing community has well proved the limitedness of the model phrase structure grammar, current analysis for natural language processing is predominantly built on the dependency grammar. The Project of Universal Dependencies (UD), which has emerged in recent years, seeks to address human languages worldwide. Its latest edition covers 70 treebanks of 50 languages.[①] Yet, such developments have seldom been incorporated into linguistic courses, or in other words, students of linguistic majors are presented with few opportunities to learn about what the natural language processing community looks like. This prevents these students from applying what they have learned. While well-trained linguists are needed in society, teaching within the discipline falls far behind the times and makes it impossible to satisfy the need. Linguists' endless arguments about various concepts they create in ways that hardly reflect the true colors of language are as senseless as questioning how many angels can dance on a pinpoint (Percy & Samway, 1991). The analogy is not meant to deny the value of flowers planted in the garden—even plastic and silk flowers can add luster to people's lives. That being said, instead of always staying in the garden, humans should embrace the real, colorful world, which is there regardless of your ignorance. To this end, only by advancing with the times, facing authentic and natural linguistic data, and leveraging more scientific research methods can linguists enable the language rules or

① Available at: http://universaldependencies.org/.

theories they discover or propose to better serve society. At the same time, linguistic majors need to provide more courses that allow future linguists to be ready to perform tasks that reflect the distinct characteristics of the times.

Objective of Data-intensive Approaches and Language Studies: "I & S"

From 2010 onwards, we have clearly declared the objective of language studies—to enable language studies in China to step into the international arena and make them scientific ("I & S" for short)—in various occasions. Why have we done so?

One part of that objective is to enable language studies in China to step into the international arena. As its definition indicates, linguistics studies rules of language systems, so it should be of universal relevance. My interest in linguistics was aroused by the process of learning foreign languages many years ago. Later, I accidentally read an impressing sentence: "While learning language seeks to add knowledge to individuals, studying linguistics seeks to add knowledge to all human beings" (Xu, 1988). Linguistic research should have universal value. A true fact is, linguists in Mainland China have made limited contributions to the world's linguistics community since the reform and opening up or earlier, though the territory may boast the world's largest pool of linguistic researchers. This fact cannot demonstrate that Chinese linguists' studies are valueless, but strongly suggests that those studies are unknown to the rest of the world. Obviously, this falls behind China's overall economic and scientific development. The nation and society call on language studies in China to step into the international arena. In particular, in the context of the "Double First-class" initiative, China's disciplines must go out and share great achievements worldwide—to be world-class is based on making yourself known to the rest of the world. As the nation advocates developing world-class disciplines, can a discipline never heard of by other countries become world-class? Of course not. It is illogical for a man who claims himself to be a world champion of a sporting event never turns up in any international competitions. Striving to be world-class can make sense only if Chinese linguists make themselves known worldwide and compete with peers of other countries at the starting line in international competitions. This is the only way to demonstrate that Chinese linguists are also able to study interesting issues and contribute to the development of linguistics in the world.

The other part of that objective is to make language studies scientific. This represents both a task for Chinese linguists and what global linguists pursue. An effective way to achieve recognition from the science community is publishing research papers in high-profile science journals. However, such publishing is tough for linguists. If journals recognized by scientists rarely receive articles contributed from a discipline, how can the discipline become part of science or a leading scientific discipline? An important reason behind this difficulty is that scientific research entails scientific methods. In view of what linguistics now looks like, leveraging scientific methods is the only way to make language studies scientific.

Here is a pertinent question: What is the relation between the data-intensive research paradigm and "I & S"? In many cases, the difficulty can be attributed to factors other than language barriers,

including objects of studies and research methods. As regards selecting objects of studies, it is crucial to consider how to generalize special issues in Chinese from the perspective of linguistics. As to research methods, the data-intensive research paradigm is more recognizable to academics than speculative or introspective methods. Data is essential for verifying, hypothesizing or discovering patterns in the course of leveraging the data-intensive research paradigm. As part of the efforts to achieve "I & S," we must introduce Chinese linguists' productive studies to the world by combining the strengths of China's linguistics circle with common practices of the science community, thereby convincing foreign countries that Chinese people are also able to make outstanding research achievements. The data-intensive research paradigm is undoubtedly helpful to this.

Interdisciplinary Language Studies in the Big Data Era

Recent years have seen "interdisciplinary research" become a hot word in the academic world. Academic activities were not born with disciplines—history tells us there have been countless people expert in both literature and science. Later, as technological development enabled the boom, complication and diversification of means for exploration, knowledge and skills beyond any single individual's ability required academic activities to be divided into disciplines. As dozens of years of study on specific methods suggests, each specific method works like a blind man feeling an elephant. It is necessary to synergize specific methods to conclude what the entire elephant is like. As such, researchers tend to leverage different methods and means while addressing the same object of study. For example, biological, physical and mathematical methods may be adopted to study a language. This is how an interdisciplinary landscape is shaped.

There is a widespread misunderstanding that an interdisciplinary landscape takes shape whenever people from different majors work together. In fact, such co-working often produces undesired effects as the object of study is unclear. In theory, an interdisciplinary language study refers to studying a linguistic issue leveraging methods of other disciplines. For example, in case we are curious about a linguistic issue that cannot be studied adopting existing methods of the linguistics discipline, we may draw on methods of other disciplines.

An example of interdisciplinary language study is given here. As a study on child language acquisition suggests, a child makes a leap in mastered syntax of the native language at the age of 2 or 3—as language is a complex adaptive system. Although the vocabulary size is smaller, the syntactic complexity of a child's speech at the age of 2 or 3 is very close to that of the adults'. Previous observations on psycholinguistics and child language acquisition have revealed but failed to clearly display the phenomenon. A few years ago, syntactic emergence was displayed by Spanish researchers using a complex network targeting children aged 2 or so (Corominas-Murtra, Valverde & Sole, 2009). That proves "interdisciplinary" does not mean "boundless". Essentially, the "inter" is a practice of addressing an issue within a discipline by drawing on methods of other disciplines.

Recent years have witnessed achievements made by our team in interdisciplinary linguistic studies. For example, we conducted a typological study on Slavic languages using a complex network.

Today, word order typology represents the mainstream of linguistic typology. Traditional methods of word order typology cannot work well when used to analyze Slavic languages, which feature morphologically rich forms and highly flexible word orders. Given this, we drew on the means of complex networks from statistical physics—indicators of complex networks were used to study authentic texts in 12 Slavic languages (Liu & Cong, 2013). To get an insight into what interdisciplinary linguistic studies are like, you may read the two articles we published in *Physics of Life Reviews* concerning how to study laws of human language using a complex network to discover linearization models of human languages by analyzing dependency distances (Liu & Xu, 2017; Cong & Liu, 2014).

In the two examples above, "inter" does not mean "movement" into the physics community. From the perspective of physics, such interdisciplinary language studies extend applications of complex networks, provide typical application examples, and enrich the complex network theory. From the perspective of linguistics, complex networks help address linguistic issues previously difficult for us. However, as the two disciplines increasingly interact and connect more closely, an interdiscipline, or even a new research paradigm, is very likely to take shape. The broad mix may distinguish the new discipline from both physics and linguistics, making it hard to identify it as physical linguistics or linguistic physics.

Can big data also help the development of interdisciplinary linguistic studies? In practice, interdisciplinary linguistic studies require researchers to have a certain amount of knowledge about other disciplines. Traditionally in a narrow sense, linguistics is defined as a discipline that "studies how a language is structured and evolves." While linguistics actually has many other aspects, it basically seeks to process linguistic data. That means it needs knowledge of statistics, mathematics and computer science. For example, in an interdisciplinary language study, software that is used to study networks in the biology community may serve as a tool to study a network constituted of linguistic data. Alongside that, with language being taken as a complex adaptive system, rules discovered from authentic texts may be instructive to computational linguistics and natural language processing, two promising disciplines. As what we deal with is linguistic data, it is obvious that data-based approaches can help the development of interdisciplinary linguistic studies.

What is More

Linguistics is a branch of science. However, its identity makes no sense without recognition by the science community. We maintain that with effort, linguistic studies can be made scientific, on the premise that scientific methods are learned and adopted. In the long run, making more efforts to do that is helpful and essential to the development of the linguistics discipline and improvements in individuals' academic abilities. Endeavors and the courage to tackle big issues are critical to making breakthroughs. A discipline anyone can easily get access to and make achievements in can hardly be part of science. In view of declining humanities studies in China, Ge Zhaoguang (at Fudan University) argued in an article entitled "How Can Humanities Succeed in Self-rescue?" "It takes a good

blacksmith to make steel." At the end of the article, he wrote: "If humanistic knowledge imparted at university was something amateurs can acquire through imitation, it would be unnecessary to engage those doctors or professors to lecture" (Ge, 2012). Even humanities cannot develop without researchers' hard work, to say nothing of linguistics, which claims to be a branch of science.

(Acknowledgements: Thanks to all the teachers and students who contribute questions to this paper. Their interest in linguistic studies has served as a major driving force for this paper. Also, we are grateful to Chen Heng, Chen Xinying, Huang Wei, Jiang Jingyang, Liang Junying, Lu Qian, Xu Chunshan and Yu Shuiyuan who offered suggestions on the first draft.)

REFERENCES

Bod, R., Hay, J., & Jannedy, S. (2003). *Probabilistic linguistics*. Cambridge, Mass: The MIT Press.

Bresnan, J., Asudeh, A., & Toivonen, I., et al. (2015). *Lexical-functional syntax*. 2nd Edition. John Wiley & Sons.

Bresnan, J. (2017). Linguistics: The garden and the bush. *Computational Linguistics*, 42(4), 599-617.

Chen, G., & Xu, C. (2015). *Introduction to big data: key technologies and best practices of industry application*. Beijing: Tsinghua University Press.

Comrie, B. (1989). *Language universals and linguistic typology*. In (Shen Jiaxuan, Trans.). Beijing: Huaxia Publishing House.

Cong, J., & Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews*, 11(4), 598-618.

Corominas-Murtra, B., Valverde, S., & Sole, R. V. (2009). The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. *Advances in Complex Systems*, 12(3), 371-392.

Eberhard-Karls-Universität, Tübingen. (2005). Linguistic treebanks and data-intensive parsing (ESSLLI 2005). *Treebanks: An Overview*. Retrieved from http://www.sfs.uni- tuebingen.de/~kuebler/esslli05/treebank-intro.pdf.

Eddington, D. (2008). Linguistics and the scientific method. *Southwest Journal of Linguistics*, 27(2), 1-17.

Ellis, N. C., & Larsen-Freeman, D. (2009). *Language as a complex adaptive system*. New Jersey: Wiley-Blackwell.

Erez, A., & Jean-Baptiste, M. (2015). *Uncharted: big data as a lens on human culture*. In (Wang, Tongtong, Shen, Huawei, & Cheng, Xueqi, Trans.) Hangzhou: Zhejiang People's Publishing House.

Ge, Z. (2012). How can humanities succeed in self-rescue? *Shanghai Wave*, (9), 96.

Hauser, M., Chomsky, N., & Fitch, T. (2002). The faculty of language: "what is it, who has it, and how did it evolve?." *Science*, 298(5598), 1569-1579.

Hey, T., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data- intensive scientific discovery*. US: Microsoft research Redmond, WA.

Hirschberg, J. (1998). "Every time I fire a linguist, my performance goes up," and other myths of the statistical natural language processing revolution (Invited speech). *15th National Conference on Artificial Intelligence* (AAAI-98), Madison, Wisconsin.

Holland, J. H. (1995). *Hidden order: how adaptation builds complexity*. NY: Basic Books.

Jelinek, F. (2009). The dawn of statistical ASR and MT. *Computational Linguistics*, 35(4), 483-494.

Jelinek, F. (2005). Some of my best friends are linguists. *Language Resources and Evaluation*, 39(1), 25-34.

Jonny, H. M., & Scott E. P. (2012). *Complex adaptive systems: an introduction to computational models of social life*. In (Long Yuntao, Trans.). (309), Shanghai: Shanghai People's Publishing House.

Karlsson, F. (2010). "Syntactic recursion and iteration," Hulst H V D, editor, *Recursion and human language*. New York and Berlin: Mouton de Gruyter, 43-67.

Kretzschmar, W. (2015). *Language and complex systems*. Cambridge: Cambridge University Press.

Li, G. (2015). Further understanding of big data. *Big Data Research*, (1), 8-16.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: a new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*. Retrieved from http://doi. org/10.1016/j. plrev. (03), 002.

Liu, H. (2010). Dependency direction as a means of word-order typology: a method based on dependency treebanks. *Lingua*, 120(6), 1567-1578.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), 159-191.

Liu, H. (1997). Dependency grammar and machine translation. *Applied Linguistics*, (3), 87-93.

Liu, H. (2009). *Dependency grammar: from theory to practice*. Beijing: China Science Publishing & Media Ltd.

Liu, H. (2017). *An introduction to quantitative linguistics*. Beijing: The Commercial Press.

Liu, H., & Cong J. (2013). Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58 (5), 432-437.

Lu, Q., Xu, C., & Liu, H. (2016). Can chunking reduce syntactic complexity of natural languages? *Complexity*.

Morin, E. (2008). *Introduction à la pensée complexe*. In (Chen Yizhuang, Trans.). Shanghai: East China Normal University Press.

Percy, W., & Samway, P. (1991). *Signposts in a strange land*. New York: Farrar, Straus, and Giroux, xv, 428 p.

Sampson, G. (1997). Depth in English grammar. *Journal of Linguistics*, 33(1), 131-151.

Solé, R. V., & Goodwin, B. (2008). *Signs of life how complexity pervades biology: how complexity pervades biology*. New York: Basic Books.

Viktor, M., & Kenneth, C. (2013). *Big data: a revolution that will transform how we live, work, and think*. In (Sheng Yangyan & Zhou Tao, Trans.). Hangzhou: Zhejiang People's Publishing House.

Wang, S. (2006). Language is a complex adaptive system. *Journal of Tsinghua University*, 6 (21), 5-13.

Xu, G. (2000). *System science*. Shanghai: Shanghai Scientific & Technological Education Publishing House.

Xu, L. (1988). *Theory of generative grammar*. Shanghai: Shanghai Foreign Language Education Press.

(Translator: Huang Chaozheng ; Editor: Gerald)