ISSN 2706-6827 Vol. 6, Issue 9: 50-56, DOI: 10.25236/IJFS.2024.060908

A Study of Language Data and Its Governance Concept

Jingtao Liang[1,2,a], Tingting Su[3,b], Zhiping Zhang[4,c,*]

[1]Institute Department of Humanities and Management and Hebei Key Laboratory of Health Care with Traditional Chinese Medicine, Hebei University of Chinese Medicine, Shijiazhuang, Hebei, China
[2]Institute for Language Policies and Standards, Beijing Language and Culture University, Beijing, China
[3]School of International Education of Chinese Language, Beijing International Studies University, Beijing, China
[4]School of Journalism and Communication, Shijiazhuang University, Shijiazhuang, Hebei, China
[a]1466271026@qq.com, [b]pistacie@163.com, [c]zxpzzp2006@126.com
[*]Corresponding author

Abstract: *Language data is an interdisciplinary concept, representing an observation and research perspective on language based on Data Science and Political Economy. As Chinese government has put data into production factors, language data, as a significant type of data, also possesses the attributes of production factor. This paper explores the understanding of language data from the perspectives of its attributes and characteristics, arguing that language data can be used as production factors, living factors, and language resources, and shows the characteristics of diversity and systematicness. Based on this understanding of language data, a dynamic and collaborative concept of language data governance should be established.*

Keywords: *Language Data, Production Factors, Renewed Understanding, Language Data Governance*

1. Introduction

The 14th Five-Year Plan in 2021 emphasized the need to develop the digital economy and fully activate the role of data as production factor. This policy outlines the functions and functional domains of data, with the former referring to its role as production factor and the latter corresponding to the digital economy. In other words, data functions as a production factor within the digital economy (Li Yuming & Liang Jingtao, 2024; Liang Jingtao & Zhang Hongjie, 2024; Liang Jingtao & Tang Peilan, 2024). [13] [17][19]In this policy context, Li Yuming (2020) pioneered the academic proposition that language data, being among the most important types of data, should also belong to the production factors[6-7]. The bringing forward of the concept of language data has brought it back into the academic research. Broadly speaking, the research on language data in academia primarily focuses on:

1.1. Definition and Types of Language Data

Language data is various data composed based on the linguistic symbol system (Li Yuming & Wang Chunhui, 2022) [12], encompassing linguistic data and discourse data. Linguistic data is derived from linguistic fact research, including speech data, text data, lexical data, grammatical data, linguistic knowledge data, etc. Discourse data is formed in using linguistic data and can be categorized as monolanguage data or parallel data. As language relies on certain media for perception (Li Yuming, 2017) [4], monolanguage data can be further divided into speech data, text data, and machine language data based on the different medias. Machine language data originates from natural language processing (NLP), encompassing both machine-readable language data and language data generated by machines (Li Yuming, 2023) [10-11].

1.2. Proposition of Language Data as Production Factor

Drawing upon the policy backdrop of data being recognized as production factor, Li Yuming (2020) proposed that language data is an integral component of production factors[6-7]. Subsequently, he further elaborated on this topic in the contexts of the language industry (Li Yuming, 2020) [7],

ISSN 2706-6827 Vol. 6, Issue 9: 50-56, DOI: 10.25236/IJFS.2024.060908

language resource (Li Yuming, 2022) [8-9], and the era of Man-Computer Symbiosis (Li Yuming, 2023) [10-11].

1.3. Exploration of the Relationship between Language Data and Other Production Factors:

Li Yuming (2020) argued that language data significantly facilitates the functioning of other production factors, based on proficiency the relationships between language and labor, linguistic symbols and knowledge, as well as language and technology, management, and investment [7]. Mao Xianzhuang (2023) termed this effect of language data on other production factors as its pan-production factor nature [21].

Taking these explorations as starting point and guided by the development of the digital economy, the necessity and urgency of re-evaluating language data become more and more evident. Whether human's understanding of language data aligns with reality and meets public policy demands determines our ability to govern it well and fully leverage its supporting role in national development strategies. This study, based on a systematic review of language data research, observes and studies language data from the perspectives of its attributes and characteristics, aiming to establish a scientific understanding of language data and its' governance, and lay a cognitive foundation for its development and governance.

2. Attributes of Language Data

In the past, human beings primarily focused on the role of language data in social life rather than in social production. Starting from the policy context of data becoming production factor, we should broaden our perspective on language data.

2.1. Language Data as Living Factor

Language data plays an irreplaceable role in human social life and serves as living factor. It is the most significant carrier of information, providing fundamental support for human daily communication, knowledge accumulation, cultural inheritance, and interpersonal relationship coordination. Language data exists in three forms: spoken language, written text, and digital data, with each form constraining the extent of its functionality.

Language data in the form of spoken language functions exclusively at the moment of its production in human society. It spreads rapidly but has limited spatial distance, being fleeting and unable to be replayed. If the listener fails to hear or understand clearly, they can only rely on the speaker to repeat the information, for achieving the transmission of messages.

The written form of language data compensates for these shortcomings. However, it is influenced by the medium, resulting in slower transmission speeds. When the volume reaches a certain level, it becomes inconvenient to carry around.

Digital language data offers greater convenience for the aforementioned functions. As long as there is an access to the internet connection, it can be transmitted quickly or even in real-time and accessed by using language technology software.

2.2. Language Data as Production Factor

Language data serves as a fundamental raw material in human social production, constituting an importan type of production factors. In the context of global economic integration, capital and commodities flow across borders, requiring an information-barrier-free linguistic environment for international market circulation. Translators with bilingual language performance, due to their foundational support, have a vast social demand, leading to the differentiation of various language service professions and the creation of the significant economic benefits. In the language service-centric industrial chain, lexical data, speech data, grammatical data, and linguistic knowledge data serve as input materials, manifested as oral and written language data processed through linguistic competence of translators. Language data, as an object of labor, has begun to play the role of production factor in economic activity or social production (Liang Jingtao, 2023) [16]. Furthermore, as we advance towards the Information era, humans technologically process language data (both oral and written) to enable machines to understand natural language. For written language data alone, its'

ISSN 2706-6827 Vol. 6, Issue 9: 50-56, DOI: 10.25236/IJFS.2024.060908

processing involves scanning, recognition, proofreading, cleaning, storage, transmission, word segmentation, tagging, analysis, and other steps, each of which can differentiate into independent professions, forming social divisions of labor (Durkheim, 2013) [1] and an industrial chain centered on language data processing. The raw material here is written language data, clearly demonstrating the attribute of language data as production factor. As we move towards the Artificial Intelligence (AI) era, the attribute of language data as production factor will become more apparent, recognized by an increasing number of people. To implement the 14th Five-Year Plan and activate the attribute of data as production factor, we must first fully recognize the production factor attribute of language data.

2.3. Language Data as Language Resource

The data science research does not distinguish between language data and non-language data. In the history of natural language processing (NLP), when rule-based approaches failed to advance, humans innovated by adopting statistically driven methods, achieving a leap in speech recognition technology. The NLP research refers to this statistically driven approach as data-driven, where the data refers to language resource, hence also known as language data. Thus, the concepts of language data and language resource are largely similar. Language resource can be used as language data, and vice versa. Scholars have explored the functions of language resource. Li Yuming (2019) defined the primary functions of language resource as language information processing, language preservation, and language teaching [5]. Building upon previous research, Liang Jingtao (2020) constructed an 8+N functional system for language resource [14]. Given the high similarity between these two concepts, language data can naturally be used as language resource, has the primary functions and functional systems of language resource.

3. Characteristics of Language Data

One of the key characteristics of language data is the diversity, manifested in both its types and sources. In terms of types, it encompasses three primary forms: oral language data, written language data, and digital language data. Digital language data primarily serves computers and facilitates natural language understanding, leveraging linguistic technology tools to benefit human society. Meanwhile, oral and written language data directly serve human society, undertaking the missions of information recording, preservation, transmission, and inheritance. Regarding sources, any member of a linguistic community can generate oral language data through language use; literate members can further produce written language data. However, due to the costs of paper usage, not all written language data has the chance to be preserved. With the coming of the Information age, language data and its carriers have become virtualized and digitized. The improvement of network infrastructure has virtually reduced the cost of preserving language data, enabling digital language data to document both oral and written language data generated by any member of a linguistic community during his or her language use.

Another significant characteristic of language data is the systematicity, which is embodied in both internal and external systems. Language data constitutes a universally interconnected system, with different types of language data forming its internal system. Whether it's oral, written, or digital language data, or the four major extensions discussed by Li Yuming (2020) [7], they can all be interconnected through the shared structures and the same objects, collectively comprising the language data system. Due to constraints in technology and supporting resources, the construction of language data often necessitates considerations of language material selection principles, focusing on balance and representativeness, leading to the formation of distinct language data samples. Rapid advancements in storage and cloud computing technologies have facilitated the storage and processing of massive amounts of language data. The iterative improvement of algorithms and computing power has significantly reduced processing and analysis time, enabling relatively holistic language data analysis. The relative holism of language data as a foundational analytical material contributes to more comprehensive conclusions and knowledge generation, bolstering AI development strategies. To avoid duplication and waste of resources, it is preferable to achieve relative holistic language data through the circulation and aggregation of existing language data.

The functionality of language data is not isolated but relies on supporting resources such as scientific research, which constitute its external system. The level of scientific research determines humanity’s understanding of language data and its functions. Language data has existed since ancient times, but its application in information processing has hinged on the development of computer research, invention, and fundamental linguistic theories. Computer scientists and linguists are crucial

ISSN 2706-6827 Vol. 6, Issue 9: 50-56, DOI: 10.25236/IJFS.2024.060908

human resources in this endeavor; without their ingenuity and the continuous updating of technologies like word segmentation and annotation, the utilization of language data would be constrained by limited supporting resources. Additionally, computer operation necessitates electrical energy; without sufficient power, the computation and analysis of language data would remain theoretical. In conclusion, scientific research, human resources, electrical energy, and other supporting resources are essential to language data development, collectively forming its external system.

4. Language Data Governance Concept

The understanding of language data determines its governance concept, which encompasses a dynamic perspective and a systematic perspective.

4.1. Dynamic Perspective

The essential attribute of production factor lies in its circulation. The full activation of language data as production factor is based on its circulation and aggregation, implying that language data should not be static but dynamic. Therefore, the governance of language data must adhere to a dynamic perspective, which is embodied in the production factor perspective and the functional perspective.

Under the current public policy framework, data has been added to production factor, same as the traditional factors such as land, labor, and capital. Language data, as a crucial production factor, is an integral part of this system (Li Yuming, 2020; Liang Jingtao, 2022). [6-7][15] The specific manifestations of language data participating in economic activities as production factor include the commoditization of language data itself or the commoditization of productions developed through processing of language data as labor objects. However, these are merely applications driven by language data and cannot be compared to the formation of an industrial chain based on it and its impact on upgrading and transforming the existing industries. In the near future, various industries will emerge around the evaluation, collection, processing (cleaning, desensitization, computation, analysis and etc), circulation, sales, management (quality assessment, archiving, promotion, legal services, regulations), protection and development of language data, forming an industrial chain with language data as its core labor object. The integration of intelligent technology with traditional industries will also enable more precise production, sales, and services, leading to upgrading and transformation, particularly with the emergence of numerous intelligent products in human society that collaborate with humans in social production and daily life, revolutionizing the social world. This necessitates establishing production factor perspective in language data governance, which means governing language data based on the understanding of production factor. This concept is reflected in governance practices by emphasizing the ownership of language data and establishing a property rights system for language data (Li Yuming & Liang Jingtao, 2024) to define the relationship between humans and objects [13]. Specifically: First, the definition of the property rights system should adhere to the principles of consistency between rights and obligations, rights and responsibilities, respecting labor, complying with public order and good customs, and considering its demonstrative effect. Second, a shared property rights system should be established between the generating subject and the construction subject, with specific proportions determined by the proportion of responsibilities fulfilled, but with clear boundaries. Third, from an intellectual property perspective, priority should be given to assigning ownership of language data to the generating subject. Fourth, the property rights of language data should be defined based on the nature of the product: language structural unit data, language rule data, and language data without intellectual property disputes should be considered as public goods; language data still under copyright protection belongs to club goods; while unpublished language data is a kind of private goods (Liang Jingtao & Zhang Zhenda, 2023) [20].

Drawing on the country's experience in governing production factors such as land and labor, the governance of language data should also adhere to a functional perspective, which means the functional perspective. Generally speaking, the governance of production factors should focus on their functions in social production and life, and different governance approaches are required for production factors with different functions. This means that on the one hand, the functions of language data need to be explored in depth, and on the other hand, the functional system of language resource can be used for reference.

ISSN 2706-6827 Vol. 6, Issue 9: 50-56, DOI: 10.25236/IJFS.2024.060908

4.2. Systematic Perspective

The characteristics of language data determine its systematic perspective in governance, specifically embodied in a relational perspective and a collaborative perspective.

(1) Relational Perspective

Given that language data forms interconnected internal and external systems, it is imperative to adopt a relational perspective in its governance, with the circulation and aggregation of language data being the primary concerns. The concept of language data as productive factor, which involves defining the property rights of language data to facilitate its commoditization and market circulation, lays the institutional foundation for this concern.

The relational perspective in language data governance necessitates aggregation based on the generating subjects and structural units during the storage phase. At this stage, the primary form of association is at the symbolic level, which can be also called symbolic connextion. Language data arises from the use of linguistic symbols by members of language communities, whether it belongs to natural language data or translanguaging data. As long as it is generated by the same subject, it can and should be aggregated together. This relatively holistic language data can reveal individual linguistic habits, voiceprint characteristics, language strategies, and further uncover the patterns of language intelligence acquisition. Among these, the analysis of voiceprint and lip-reading features, through both verbal and translanguaging data, reveals stable phonetic characteristics in language use across different contexts, based on phonetic-level associations. Additionally, this relatively holistic language data, tailored to individuals, can be utilized in companion robots, providing emotional comfort to the elderly in their twilight years. Furthermore, during the storage phase, language data must undergo recognition, proofreading and digitization to facilitate future retrieval, screening, extraction, computation and analysis. Language data that cannot be retrieved or computed cannot be directly applied.

The relational perspective in language data governance also requires aggregation based on functional requirements during the application phase. At this stage, the focus shifts from symbolic-level associations to knowledge-level associations (Liang Jingtao&Zhang Hongjie, 2020) [18]. Suppose one aims to extract discussions on language resource from relatively holistic language data. The initial step involves keyword-based retrieval, followed by screening and outputting the results for computation and analysis. This aggregation of *language resource* data relies on both exact symbol matching and symbol co-occurrence matching. However, this approach may overlook language data that uses different symbols but conveys similar or identical content. Therefore, it is necessary to delve into the essence of the symbols, revisit the knowledge they convey, and correlate relevant data with the initial data at the knowledge level. This is based on knowledge associations. Subsequently, language data associated at both the symbolic and knowledge levels are aggregated for computation and analysis to reveal patterns.

It is crucial to note that while associations play a significant role in the aggregation of language data, we must not stop at relational associations. It is necessary to delve deeper into the relationships between data, such as causality, opposition, progression and etc. Logical reasoning based on these relationships can bring about qualitative leaps in artificial intelligence capabilities. Ceasing at associations can limit human thinking and result in missing perspectives. Establishing associations between language data describing the same subject varies in difficulty, with those involving subject symbols being relatively easier and those without being more complex and difficult to discern. For the latter, discovering associations relies on statistical and probabilistic analysis of data (Wu Jun, 2016) [22]. Regardless of whether it is at the symbolic or knowledge level, the relational perspective in language data governance should be emphasized.

(2) Collaborative Perspective

The concept of collaborative perspective includes two levels: collaboration among governance subjects, collaboration between humans and computers. The collaboration among governance subjects refers to the participation of all subjects involved in language data in the processing of language data, while the collaboration between humans and computers means that humans and computers need to cooperate in the governance of language data.

The sources of language data are diverse, with individuals, market subjects, and controllers such as governments, non-profit organizations (e.g., the Chinese Linguistic Data Consortium, CLDC), and individuals involved. Data scientists, linguists, and economists possess specialized knowledge in language data. In performing their duties, governments collect language data from market subjects and individuals, making them the largest controllers of language data. The diversity of language data

ISSN 2706-6827 Vol. 6, Issue 9: 50-56, DOI: 10.25236/IJFS.2024.060908

necessitates a collaborative approach to its governance. When language data has become production factor, the government's advantage lies in its public power, which determines its role as the steersman in language data governance. Market subjects, with their extensive experience in language data development and utilization, possess the most advanced language data development technologies and represent the most advanced trends in language data utilization, making them more suitable for the role of rowing in language data governance. Data scientists, linguists, economists, and other experts possess knowledge related to language data governance and can provide professional advice to better leverage the production factor function of language data, making them an indispensable part of language data governance. In addition, non-profit organizations and individuals can play complementary roles, for example, the government can authorize non-profit organizations and individuals to language data evaluation and uses, and supervise process steps. This establishes a five-in-one collaborative governance mechanism involving the government, market, scientists, non-profit organizations and individuals.

Technical demands in language data governance underscore the importance of Man-Computer collaboration. Depending on the degree of aggregation, language data can be classified into point-based, strip-based, and block-based types. Regardless of the type of language data presented, its processing, computation, analysis and mining cannot be separated from computers. Man-Computer collaboration is manifested in two aspects:

a) Man-Computer Collaboration in Data Extraction and Processing. Computers far surpass humans in data extraction storage and computation capacity, making them more suitable for these tasks. However, data scraping and computation are not auto processes for computers; they are carried out under human instructions. These instructions are input into computers through programming languages (artificial language data), which are then read, executed and produce results. When scraping data, considerations of completeness and representativeness are crucial: the data should be comprehensive, covering various aspects rather than just one or a few, and the data from different aspects should be typical rather than exceptional. Establishing principles for completeness and representativeness, however, relies heavily on human judgment. Essentially, computers are an extension of human will, carrying out tasks as directed by humans. While manual data collection was the primary method before computers, personal biases in selection criteria are inevitable. Language data is not just an information or knowledge object, it carries specific emotional attitudes. Humans tend to assign emotional values to language data based on their personal relationships, leading to classification biases (Durkheim et al., 2012) [2]. Even with established selection criteria, humans inevitably introduce varying degrees of subjectivity into the data selection process, potentially compromising completeness and representativeness. This, in turn, can amplify errors in the underlying data during computation and analysis (Huang Changning et al., 2001) [3]. Computers, devoid of emotions, can eliminate such human biases in data scraping, ensuring the reliability. Given the vast volume of scraped language data, computational tasks are more efficiently handled by computers, as manual computation would be excessively time-consuming and labor-intensive.

b) Man-Computer Collaboration in Data Analysis. Due to current concepts and technological limitations, computers can only perform shallow-level data analysis, with deep-level analysis relying on human intervention. In the era of big language data, correlations between phenomena can be identified through statistical analysis, direct and indirect experience, reasoning, brainstorming and other methods. Once correlations are established, humans must actively analyze and uncover the underlying logical relationships, refining the cognitive reasoning process and enriching human knowledge.

5. Conclusion

Starting from the policy context of data as productive factor and based on a historical review of language data research, this paper advocates for a renewed understanding of language data. Language data possesses the attributes of both living factor, production factor, and language resource, characterized by diversity and systematicity. Based on this understanding, we have examined the dynamic and systematic perspectives of language data governance. However, exploring the concept of language data and its governance is not the end, but the beginning of language data research. This leads to the following topics: whether the functions of language data correspond to the living factor, production factor, and language resource, and how to govern language data. Solving these problems has important practical significance for the development and governance of language data.

ISSN 2706-6827 Vol. 6, Issue 9: 50-56, DOI: 10.25236/IJFS.2024.060908

Acknowledgement

This work was supported by China National Social Science Fund (No. 20BYY058, Title: A study of the National Governance Capacity of Language) and China National Committee for Terminology in Science and Technology Fund (No. YB2023016, Title: A study of the Production Factor Function of Term Data).

References

[1] Émile Durkheim, 2013, The Division of Labor in Society, translated by Qu Dong, Sanlian Bookstore.
[2] Émile Durkheim, Marcel Mauss, 2012, Primitive Classification, translated by Ji Zhe, The Commercial Press.
[3] Huang Changning, Li Juanzi, 2001, Corpus Linguistics, The Commercial Press.
[4] Li Yuming, 2017, The Impact of Language Technology on Language Situation and Social Development, Social Sciences in China, Vol.2: 145-158.
[5] Li Yuming, 2019, Theories and Practices of China's Language Resources, Chinese Journal of Language Policy and Planning, Vol.3: 16-28.
[6] Li Yuming, 2020, Language Data as Production Factor in the Information Age, Guangming Daily, July the forth (012).
[7] Li Yuming, 2020, Language Industry in the Information Age, Journal of Shandong Normal University (Social Science Edition), Vol. 5: 87-98.
[8] Li Yuming, 2022, A Brief Introduction to Language Planning Theory, Lexicographical Studies, Vol.1: 1-17+125.
[9] Li Yuming, 2022, Language Resource and Language Resource Studies, Language Teaching and Linguistic Studies, Vol.2: 1-4.
[10] Li Yuming, 2023, On the Issues of Language Data in the Era of Human-Machine Symbiosis, Journal of Huazhong Normal University (Humanities and Social Sciences Edition), Vol. 5: 135-143.
[11] Li Yuming, 2023, Language as a Gap and Bridge to Culture, Journal of Tianjin Normal University (Social Science Edition), Vol. 6: 34-42.
[12] Li Yuming, Wang Chunhui, 2022, Host Words: From Data to Language Data, Chinese Journal of Language Policy and Planning, Vol. 4: 13-14.
[13] Li Yuming, Liang Jingtao, 2024, On the Function as a Factor of Production and Property Rights Systems Construction of Language Data, Language Teaching and Linguistic Studies, Vol. 2: 1-11.
[14] Liang Jingtao, 2020, A Functional Research on Language Resources, a doctoral thesis of Beijing Language and Culture University.
[15] Liang Jingtao, 2022, On How Language Data Performs as Production Factor, Journal of Institutional Economics Studies, Vol. 4: 222-233.
[16] Liang Jingtao, 2023, On Language Data, a Report on Postdoctoral Research of Ministry of Education Institute of Language Application and Beijing Normal University.
[17] Liang Jingtao, Tang Peilan, 2024, On why language is a resource, Journal of Tianjin Normal University (Social Science Edition), Vol. 4: 151-160.
[18] Liang Jingtao, Zhang Hongjie, 2020, On the Concept of Language as Knowledge in the Study of Language Resources, Journal of Language Planning, Vol. 2: 82-87.
[19] Liang Jingtao, Zhang Hongjie, 2024, On Practical Path of Language Resource from the Perspective of Digital Economy, China Language Strategies, Vol. 1: 189-198.
[20] Liang Jingtao, Zhang Zhenda, 2023, On Property Rights of Language Data from the Perspective of Intellectual Property, Journal of Institutional Economics Studies, Vol. 4: 216-231.
[21] Mao Xianzhuang, 2023, A New Model of Language Industry Development under the Background of Digital Economy: Language Data Industry, Journal of Beijing City University, Vol. 2: 74-80.
[22] Wu Jun, 2016, The Age of Intelligence: Big Data and Smart Revolution Redefine the Future, CITIC Press.