Journal of Literature and Art Studies, June 2022, Vol. 12, No. 6, 653-660
doi: 10.17265/2159-5836/2022.06.010

The New Transformations and Prospects of Speech Act Research from the Multimodal Perspective

LUN Xin-yu

Macau University of Science and Technology, Macau, China

This article sorts out the research changes on the speech act from a multimodal perspective. After collecting relevant documents from the two major retrieval systems of Web of Science and China National Knowledge Infrastructure (CNKI) databases, the research direction of speech behavior in the linguistics field has shown a shift from monomodal to multimodal, and the research object has shown a shift from written language to on-site speech. The research method has changed from formalization to corpus construction. Through these changes, this article conducts a forward-looking discussion on potential research topics in the future, intending to provide assistance for future research on speech acts.

Keywords: multimodality, speech act, new transformations, foresight

Introduction

In the middle and late 20th century, interdisciplinary research paradigms gradually emerged, and the degree of mutual learning and exchange among various disciplines deepened. Among them, the concept of "multimodal" borrowed from biological sciences is often used in linguistics as "human beings interact with the external environment (such as people, machines, objects, animals, etc.) through a variety of senses (such as vision, hearing, etc.) or meaningful resource symbols (such as text, image, sound, etc.)." From this perspective, the research trend of speech act theory has shown a new change. To better demonstrate these changes and to control the new direction of speech act research, this article sorts out the literature on the Web of Science and China National Knowledge Infrastructure (CNKI) databases, whose topics and keywords include "speech act" and "multimodal." All the literature was analyzed from the three perspectives of research direction, research object and research methods. Based on the results, the potential research topics in the future will also be prospectively discussed.

Research Direction from Monomodal to Multimodal

In the 1950s, the speech act theory proposed by the British scholar J. L. Austin (1955) under the trend of analytic philosophy broke the traditional linguistics understanding of "language is a symbol system for describing the world" and regarded the essence of language as "a peculiar human behavior." This theory has

LUN Xin-yu, Ph.D., Senior Teaching Assistant, International College, Macau University of Science and Technology.

constructive significance for the development of linguistics, which has now become an indispensable part of the field of pragmatics after continuous revision and supplementation by scholars.

Since the speech act theory was put forward, the related research has mostly focused on a monomodal perspective, that is, starting from the language level, dividing different speech acts and investigating their illocutionary force. Among them, Austin (1955) focused on the study of distinguishing speech acts. He tried to clarify "the entire speech act in the total speech context" and divided the entire speech act into "locational act," "illocational act" and "prelocational act," which mean "speaking verbal and meaningful words," "giving a kind of illocutionary force to the words in a particular context" and "a certain effect of speaking or acting on the listener," respectively. Among these three types, illocational acts are considered to best reflect the speaker's intentions. Each type of speech act can be distinguished from other types because the illocutionary force contained is different. Therefore, the research on speech acts mostly refers to the investigation of illocutionary forces.

In the early days of speech act theory, Austin divided verbs into agentive verbs and perlocutionary verbs and classified five types of speech acts based on the actions performed by more than 3,000 agent verbs in English: verdictives, exerctives, commissives, behabitives and expositives. Although Austin is not very satisfied with the initial classification, this research fills a gap in the philosophy of language. Since then, Searle (1969), based on criticizing and inheriting the Austin speech act theory, summarized the effective conditions for the smooth implementation of all speech acts into preparatory conditions, sincerity conditions, essential conditions and propositional content conditions and divided speech acts into stated, directive, commissive, expressive and declaration by typical English acting verbs. Compared with Austin's classification, this classification is considered more reasonable and rigorous. However, with the expansion of research, scholars have gradually realized that speech acts in daily life are not equivalent to the simple use of agent verbs. Starting with agent verbs in the language cannot fully describe the illocutionary force of each type of speech act. As a result, research on speech acts from different perspectives has been carried out. During this period, Martin (1981) abandoned the approach of starting with acting verbs and divided speech act into 17 categories from the perspective of function. His classification is based on a completely different classification standard. In addition, scholars such as Sinclair (1975) and Hancher (1979) combined the syntactic and grammatical characteristics of different situational factors to make Searle's classification framework more detailed. Wunderlieh (1980) proposed four criteria for the classification of speech act: classification according to the main grammatical markers in a specific language, classification according to the content of the proposition and the consequences of speech, classification according to the function of speech act, and classification according to the origin of speech act, and at the same time sorted out the problems that may be encountered in the classification process and provided solutions. Some scholars believe that the classification of speech act should be based on semantic logic rather than acting verbs. For example, Leech (1983) summarized the original classification into statement, instruction, promise, question and expression according to illocutionary predictions to make a new attempt on the classification of speech act. Since then, the research on the classification of speech acts did not level off until the 1990s, and most of the ways of distinguishing have a certain degree of rationality and promoted the development of speech act research. However, whether from the perspective of agent verbs or from the perspectives of semantics and function, the research perspective on speech act still remains on the monomodality of the language. This kind of research perspective has the following two major shortcomings: on the one hand, in the early days of the theory, scholars

equated the classification of agent verbs with the classification of speech act, which made the discussion of speech act remain at the syntactic level, and language is not regarded as a kind of human behavior. At the same time, this also caused later research to still be unable to escape the limitation of "researching and describing speech from the perspective of language rather than behavior". On the other hand, the illocutionary force of speech act is not only reflected in the vocabulary and syntax level but may also be reflected in the situation, context, etc., it is difficult to make a complete, prescriptive descriptive classification of speech acts based on a single-modal view within the language. Faced with this situation, the research perspective of exploring speech act from the multimodal perspective is added to the agenda.

In fact, as early as the 1960s, Austin (1969), Searle (1985) and others pointed out that the "illocutionary force display item includes not only the agent verb but also the mood, punctuation, word order, intonation contour, stress, etc". This shows that in addition to acting verbs, intonation, rhythm and stress, and other prosodic features, as well as the gestures that accompany the utterance, are all means to achieve speech act. Among them, there are few investigations from the perspectives of tone, labeling, and word order. Only a few scholars (such as Willman, 2009; Recanati, 2013) discuss the relationship between tone and illocutionary force through the analysis of discourses. The prosodic perspective, gestures, expressions and other action perspectives have led to a boom in exploring the characteristics of illocutionary force. For example, from the perspective of prosody, Chang (2005) gave a formal description of the relationship between prosody and pragmatics and believed that the characteristics of speech acts such as assertion, inquiry and request could be explored from the perspective of intonation patterns. Meyer (2006) agreed that prosodic marks the illocutionary force of statements in many languages. He marked the commonness and difference between illocutionary force statement and questioning illocutionary force in pitch stress peaks in Russian through perceptual experiments. Veaux (2011) used the Gaussian Mixture Model (GMM) to outline the prosodic features of neutral speech and expressive speech in French and believed that intonation is an important factor in distinguishing the two speech acts. De Silva (2020) described three mandatory speech acts in Brazilian Portuguese and Mexican Spanish by means of a voice perception test: the melody contours of order, request, and supplication. The results showed that intonation would be one of the mechanisms that distinguished the three indicative speech acts. Repp (2020) studied the prosodic differences between WH-exclamation and WH-interrogative speech acts in German through two experiments and found that the prosodic contour of the Exclamatory speech act in German was very stiff, while the prosodic expression of the interrogative speech act was more flexible. From the perspective of action, Watanuki (1995) used video and audio recording equipment to capture verbal and nonverbal information in interpersonal communication and analyzed the role of voice, gaze, facial expressions and gestures in expressing illocutionary force in a task-oriented dialog. Yoshikawa (1996) discussed the correlation between facial expressions and illocutionary force expression. Bucciarelli (2004) explored the role of gestures in helping agent speech act. Xiong (2006) extracted the features of gesture swings and concussion in different speech acts. Landragin (2006) derived the "multimodal reference domain" model of gestures and speech. Church (2014) summarized the synchronization characteristics of gestures and language in expressing illocutionary force. In general, it was not until the 1990s that the study of speech act from the multimodal perspective was gradually launched. However, to date, the relevant research content is still relatively scattered, and the research on various influencing factors is

not balanced. There are few comprehensive works and books focusing on speech acts from an all-around mode, which leaves a great space for scholars to study in the future.

Research Objects from Written Language to On-site Speech

Since the inception of the theory, Austin (1955) has regarded the agent verbs in the language as the entry point to explore speech act. Since then, Searle (1969), Sinclari (1975), Hancher (1979), and Leech (1983) have explored translating the spoken language in daily speech act into written language. The fixed form or prescriptive applied verbs, syntactic structures and semantic features in written language are the research objects of this period. Specifically, Searle (1969) regarded the unit of language communication as the effect of symbols, words or sentences in the speech act, and so common symbols, words or sentences in the same speech act are used as research objects. Sinclari (1975) and Hancher (1979) also supplemented Searle's research by collecting linguistic features such as acting verbs and syntactic features in different situations and contexts. Leech (1983) started from semantics and classified speech acts by inducing the semantic characteristics of acting verbs. It is undeniable that the use of agent verbs, syntax and semantics and other linguistic features as the research object to explore speech act research could extract the characteristics of speech act to a certain extent. However, the use of written language as a textual corpus also does not break out of the research limitation of equating linguistic form with speech act. Although many linguists have incorporated other language forms and contexts into the exploration of illocutionary force, the labeling, retrieval and analysis of language forms themselves are still ranked first in this period.

In the 1990s, with the change in research perspectives and the development of corpus collection and processing technology, resources such as video and audio recording on-site speech acts gradually became the corpus favored by researchers. This is because the recording of live speech could not only examine the expression of language but also analyze nonverbal forms that cannot be observed from written words. Most importantly, the common expression of illocutionary force by multimodal resources could be researched by investigating on-site speech. Based on this, an increasing number of linguists, behaviorists and psychologists have joined the related research work. For example, Kendon (1995) collected several live video materials to discuss the illocutionary force function of four common gestures in Italian. Mubenga (2009) explored the similarities and differences between English and French speech acts with film and video as corpora. Dohen (2009) took the speech fragments in adult verbal communication as objects and elaborated how the hand and mouth express the illocutionary force in speech act in harmony. Lee (2017) analyzed the role of gaze, gesture, touch and body behavior in speech act by recording live speaking practice videos of English as a second language students. Using live speech as a research object can help scholars conduct a comprehensive and multidimensional exploration of speech acts. Moreover, Gu (2013) believes that the transition from written language to live speech can help analyze the interaction between different language symbols in live discourse. Based on this, he made conceptual interpretation and model construction from the perspectives of speech, thought, emotion and appearance and answered the question of how to study live speech as an object, which is considered to be a real way to study speech as a universal behavior of people. Specifically, starting from eight case studies, he referred to the participants in the daily speech act as "living whole people" and called the implemented illocutionary force "living illocutionary force." Meanwhile, Gu (2002) defines on-site extemporaneous discourse as "words spoken (or written) by some (some)

speakers of a language at a certain time and place without prior preparation." He believes that illocutionary force should be inspected from the three dimensions of language structure, prosodic feature and physical appearance and proposes a research framework of true-to-truth modeling. Later, Huang (2017, 2018) inherited Gu's research ideas and studied the audiovisual corpus of 12 illiterate people implementing the speech act, which is the first attempt to bring together multimodal sources in Chinese to explore the expression of the illocutionary force. It truly fulfills the goal of examining live speech acts from a multimodal perspective and provides a new way to explore the emotion of illocutionary force. At present, although the study of live speech is not as rich as that of written language, it represents the frontier of speech act research and is worthy of further development in the future.

Research Methods from Formal Inspection to Corpus Construction

Since the field of modern linguistics began to develop at the beginning of the last century, formal methods have influenced the investigation of speech as one of the important research methods in this field. Scholars have tried to use this method to convert the structural rules and illocutionary force characteristics of speech act at the language level into a set of finite symbolic formulas with statutes. For example, Austin (1955) regarded agent behavior as a conventional act, which could be clearly expressed through the agent sentence pattern, and the agent verb is the core component of the agent sentence pattern. Therefore, the characteristics of the agent behavior could be extracted by exploring the usage of the agent verb. Searle (1969) took the act of promise as an example and summarized the rules of this speech act into nine formal principles. He argued that each specific speech act should have sufficient, necessary conditions and implications. Since then, methods such as multiple-choice question answering, ranking question answering, role playing, and conversation analysis have been widely used in the formal investigation of speech acts. However, speech activities as human behaviors are not mechanical and fixed, and not all speech acts can be refined into linguistically regulated finite symbolic formulas. At levels other than language levels, therefore, this approach can only summarize the characteristics of the speech act to a certain extent through certain illocutionary force markers; that is, when scholars abstract the language form characteristics of speech acts, many factors such as context and nonverbal elements that are discarded and discarded are ignored, which shows that this method is not yet able to fully investigate speech acts.

In the 1990s, corpus-based linguistic research methods began to flourish. On the one hand, due to the progress of pragmatic annotation tools, the development of computer software, such as AntConc, Tmxmall, CUC_ParaConc and Elan, has greatly improved the efficiency of labeling speech acts and other pragmatics. On the other hand, the limitations of formal methods urge researchers to find ways to better explore speech acts. Scholars gradually discovered that starting from the corpus, the collected corpus could be labeled and integrated from multiple angles and could be freely processed, modified and stored with the help of a computer. Especially when investigating live speech corpora from a multimodal perspective, this new corpus could help to maximize the freshness and authenticity of the original corpus and help explore more dimensions embodied by illocutionary force. Therefore, corpus construction has gradually become one of the important methods to explore speech acts. Through the inspection of the research results of speech acts in the past 30 years, it can be seen that the early scholars who used the corpus research method are mostly based on the relatively mature corpus that has been built. For example, Aijmer (1996) studied several speech acts in English based on the London-Lund Corpus of

Spoken English. Shriberg (1998) and Taylor (1998) investigated prosodic features such as length, pause, energy and speed in several speech acts based on the more than 1,000 dialogs in the Switchboard corpus and the utterance in the DCIEM Maptask corpus, respectively. Since then, more targeted self-built corpus methods have become the mainstream of research. For example, Colletta (2009) studied how people narrate in speech acts from speech, gesture, gaze, and facial expressions through a self-built video corpus. Noris (2011) established a Dutch multimodal video corpus through equipment such as head-mounted scene cameras and eye trackers and explored the synchronization relationship between gaze and speech. Palacios (2013) processed the collected comedy discourse into an illocutionary force corpus of humor from a multimodal perspective and discussed how illocutionary force is generated through multimodal resources such as language font color composition. Branco (2014) constructed a multimodal corpus based on two Brazilian films and explored the narrative means of speech acts in the two films. Lubis (2016) established an audio-visual emotional corpus containing 100 minutes of annotations and transcribed materials based on the speech acts of 14 Japanese native speakers and explored the emotional expression in Japanese speech acts. The corpus construction method can clearly show the interaction between language and other factors in speech act, and its usage is more flexible. Compared with traditional formal methods, corpora could also increase the objectivity of quantitative analysis for pragmatic research. Although there has not been broad agreement on the corpus marking methods of speech acts, it has become an important research project in the field of pragmatics to investigate the multimodal dimensions of speech acts from the construction of corpora.

Discussion

Speech act theory has become an important part of pragmatics. After nearly 70 years of development, relevant research results have been fruitful, reflecting the research perspective from monomodal to multimodal, the research object from written language to on-site speech, and the research method from formal inspection to corpus construction. This research trend from a two-dimensional single dimension to a three-dimensional dimension reflects that the investigation of speech act has gradually returned to the path of Austin's "Language theory as part of the theory of behavior," which is more in line with real speech. This change provides inspiration and thinking in the following aspects for future research.

Both language and nonverbal factors are an indispensable part of speech act to express illocutionary force. According to the existing research on multimodal factors, the three dimensions of language structure, prosody characteristics, and physical appearance could be used as entry points to explore the expression of illocutionary force, and their specific performance could be analyzed in more detail.

Using live speech as a research corpus could more accurately, truly and comprehensively explore speech acts. Combined with existing computer-aided markup tools, speech acts could be explored in a more comprehensive and in-depth way by collecting audio and video materials of live speech acts.

At present, there are many studies based on multimodal corpora in English, but the construction of other languages' multimodal corpora is still in its infancy. Therefore, it is an important direction of future research to find a construction method suitable for exploring multilanguage speech acts based on existing research results.

Conclusion

This article sorts out the research changes of speech acts based on a multimodal perspective. After collecting the relevant documents of the two major retrieval systems, it could be seen that the research direction of speech act has changed from monomodal to multimodal, the research object has shown a change from written language to on-site speech, and the research method has shown a change from formal inspection to corpus construction. These new changes reflect that those scholars are more detailed in the dimensions of speech act, the objects are more precise, and the methods are more diversified. At the same time, this article also carried out a forward-looking discussion on potential research topics through these new changes and provided several research directions worthy of further evaluation, intending to provide assistance for future research on speech acts.

References

Austin, J. L. (1975). *How to do things with words*. Oxford university press.

Hancher, M. (1979). The classification of cooperative illocutionary acts1. *Language in society, 8*(1), 1-14.

Huang, L. H. (2017). Speech act theory and multimodal research——Also on the logic of multimodal (corpus) pragmatics. *Journal of Beijing International Studies University*, (03), 12-30+133.

Huang, L. H. (2018). *Research on language power based on multimodal corpus: a new exploration of multimodal pragmatics*. Shanghai: Shanghai Foreign Language Education Press.

Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of pragmatics, 23*(3), 247-279.

Landragin, F. (2006). Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems. *Signal Processing, 86*(12), 3578-3595.

Leech, G. (1981). *Semantics* (second edition). Harmondsworth: Penguin Books.

Leech, G. N. (1983). *Principles of pragmatics*. London and New York: Longman

Lubis, N., Gomez, R., Sakti, S., Nakamura, K., Yoshino, K., Nakamura, S., & Nakadai, K. (2016, May). Construction of Japanese audio-visual emotion database and its application in emotion recognition. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2180-2184).

Martin, J. R. (1981). How many speech acts? *UEA Papers in Linguistics*, (14-15), 52-77.

Meyer, R., & Mleinek, I. (2006). How prosody signals force and focus—A study of pitch accents in Russian yes–no questions. *Journal of Pragmatics, 38*(10), 1615-1635.

Mubenga, K. (2009). Towards a multimodal pragmatic analysis of film discourse in audiovisual translation. *Meta: journal des traducteurs/Meta: Translators' Journal, 54*(3), 466-484.

Noris, B., Barker, M., Nadel, J., Hentsch, F., Ansermet, F., & Billard, A. (2011, August). Measuring gaze of children with autism spectrum disorders in naturalistic interactions. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5356-5359). IEEE.

Palacios, C. (2013). Some scopes of the multimodal perspective for the study of comics and humour. *Signo Y Sena-revista Del Instituto De Linguistica, 23*, 257-278.

Recanati, F. (2013). Content, mood, and force. *Philosophy Compass, 8*(7), 622-632.

Repp, S. (2020). The prosody of wh-exclamatives and wh-questions in German: Speech act differences, information structure, and sex of speaker. *Language and speech, 63*(2), 306-361.

Searle, J. R., & Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., ... & Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech, 41*(3-4), 443-492.

Silva, C. G. D., Carnaval, M., & Moraes, J. A. D. (2020). Atos de fala diretivos em português e em espanhol: uma análise acústica comparativa. *Entrepalavras, 10*(1), 10.

Sinclair, J. M., & Coulthard, R. M. (1975). *Towards an analysis of discourse*. Oxford: Oxford University Press.

Taylor, P., King, S., Isard, S., & Wright, H. (1998). Intonation and dialog context as constraints for speech recognition. *Language and Speech, 41*(3-4), 493-512.

Veaux, C., & Rodet, X. (2011). Intonation conversion from neutral to expressive speech. In *Twelfth Annual Conference of the International Speech Communication Association*.

Watanuki, K., Sakamoto, K., & Togawa, F. (1995). Multimodal interaction in human communication. *IEICE TRANSACTIONS on Information and Systems*, *78*(6), 609-615.

Willman, M. D. (2009). Illocutionary force and its relation to mood: Comparative methodology reconsidered. *Dao*, *8*(4), 439-455.

Xiong, Y., & Quek, F. (2006). Hand motion oscillatory gestures and multimodal discourse analysis. *International Journal of Human-Computer Interaction*, *21*(3), 285-312.

Yoshikawa, S., & Nakamura, M. (1996, January). Facial expressions and speech acts. *International Journal of Psychology*, *31*(3-4), 18429-18429.