P-ISSN: 2709-6254	Journal of Development and Social Sciences	April-June 2025, Vol. 6, No. 2
O-ISSN:2709-6262	https://doi.org/10.47205/jdss.2025(6-II)12	[132-140]

Journal of Development and Social Sciences
www.jdss.org.pk

RESEARCH PAPER

Exploring the Linguistic Capabilities and Limitations of AI for Endangered Language preservation

[1]Majid Ali*, [2]Dr. Zaffar Iqbal Bhatti and [3]Tanzila Abbas

1. Ph D Scholar, Department of English Linguistics, Minhaj University Lahore, Punjab, Pakistan.
2. Associate Professor, Department of English Linguistics, Minhaj University Lahore, Punjab, Pakistan.
3. Ph D Scholar, Department of English Linguistics, Minhaj University Lahore, Punjab, Pakistan.

***Corresponding Author:** majidnuml@gmail.com

ABSTRACT

The main purpose of the present study was to investigate the crossway of semantics and computational linguistics, aiming how methods of computational linguistics support to understand the process of modeling meaning in natural language. This study opts a qualitative and exploratory approach to examine the capability of AI language models to cater the dimensions of semantics and pragmatics in the context of language documentation. For the current study, the data was collected through the different sources such as real world linguistics examples to understand the nature of the data. The published documentation of under sourced languages, such as digital archive (ELAR, DOBES) grammar and fields notes were also became the source for collecting the data. The data was analyze through thematic linguistic analysis. The study opted semantic and evaluation criteria to check the lexical accuracy in the form of synonyms and antonyms use. For pragmatic, the pragmatic evaluation was done to check the contextual interpretation and appropriateness of speech act. The study finds gap between semantic and pragmatic modeling in AI language system, having a greater disparity evident in under documented languages. On the basis of these findings researchers presented some recommendations for future researchers and scholars.

KEYWORDS Linguistics, Semantics, Pragmatics, Computational Linguistics, AI Models

Introduction

The sole purpose of this study is to investigate the crossway of semantics and computational linguistics, aiming how methods of computational linguistics support to understand the process of modeling meaning in natural language. It studies the different basic theories like computational techniques, its applications and challenges along with the future directions. It also focuses to understand the dynamic role of computational linguistic in the field of semantics. As we know that human language is a complex phenomenon. It encompasses the different layers of meaning which makes it more crucial for human to understand and analyze it. The field of semantics is the branch of linguistics which deals in how to understand the meaning, or sometimes it raises the question of what is meaning of meaning? On the other hand, the computational linguistics apply computational methods like BERT and GPT to evaluate the language (Farhat, 2019; Ahmad et al., 2021; Younus et al., 2023; Maitlo et al., 2024). This collaboration led to the emergence of new field of computational semantics. The purpose of this field is to modeling and mechanizing semantics understanding. (Church and Mercer 1993)This assignment digs deep into how these methods help to understand semantics and the use of its applications in natural language processing(NLP).

Literature Review"

Review of the literature summarize and evaluate the text of writing of the definite theme, and provide frame work to think about the possible consequence of innovative study" (Ahmad, Rao & Rao , 2024, p.3944). In another statement Ahmad, Sanober &

Cheema, (2024) pointed out that "A review of literature may only be a clear overview of the sources, in an organizational pattern, and its function is to estimate and summarize the previous writings linked to current topic" (p.3) .

Semantics in Linguistics

Semantics is the systematic study of meaning that provide the proper order and sequence to study meaning. In this field each discipline has a different approach and a particular interest, yet each discipline contributes to others. The field of semantics can be divided further into linguistics semantics that has the tendency to deal in the linguistics meaning, concerning to the various kinds of meaning, including lexical meaning and grammatical meaning (Cheema et al., 2023; Abbas et al., 2024; Ahmad et al., 2025). Philosophical semantics concerns with how we know, how the particular fact that we know or accept as true is related to the other facts? It focuses the nature of meaning as true and referenced. Gottlob Frege proposed the notion about realty of meaning. According to him meaning is determined by only in the condition in which it would be occur.

Key Theories in Semantics

The main purpose of this section is to place out the major approaches to construct the relationship between the semantic theory and the natural language. This thing come to an understanding that these theories aiming to assign the meaning to the constituents of the sentences out of which the sentence meaning can derivate. The only way to derive the meaning from the sentences is what the particular theory will provide? So there is a need to find out the true semantic theory. The theory of semantics which proposed the notion to derive the meaning of the sentences out of the meaning of the words or the linguistics units in which they are combined is known as compositional semantics theory. (Katz and Fodor 1963)

Reference Theory

It is a kind of theory in which pairs of expression occur with the combination of neighboring expression which determine the truth value of sentences. The proper understanding of the theory of reference is noticeable evidence of Gottlob Frege's attempt to construct a logic sufficient for the formalization of mathematical inferences. This theory can be best understanding with the illustration of proper names, by considering the following sentences.

- Barack Obama was the 44th president of the United States.
- John McCain was the 44th president of the United States.

The first one is true and the second one is false. Clearly the difference of these two sentences can be trace out by the condition of truth value of the sentences which is determining by the difference of the expressions.

Computational Linguistics

Computational linguistics is the interdisciplinary field which deals with the language and the computer. It involves the use of computational methods to process and analyze the natural language data. The ultimate purpose of this field is to understand, interpret, and generate human language with the help of computer. By bridging the gap between the field of computational linguistics and the natural language, it plays a pivotal role in introducing the advanced technologies for example machine translation, voice assistant and sentiment analysis. Computational linguistics covers many areas of language like it addresses its core areas or fundamental aspects of language such as syntax, semantics, phonology and pragmatics with the help of algorithm and its models (Grishman 1986; Jalbani et al., 2023; Rasheed et al., 2024)

Relevance Of Computational Linguistics And Semantic Analysis

Computational linguistics is very much relevant to the semantics analysis, it introduced the different tools and perfect methodologies to understand the meaning of words, phrases and sentences in an efficient way. Semantics analysis focuses on the understanding of meaning behind the text and speech which enable machine to understand process and interpret human language effectively. This field contributed a lot to make this phenomenon possible to analyze the context, resolving ambiguity and locating the relationship between words and concepts. (Benjamin 2018; Abbas et al., 2025)

Key Challenges In Semantics Analysis

A key challenge in semantics analysis in the form of word sense disambiguation which determines the correct meaning of a word based on its context. The ambiguity will be resolved by applying the computational technique like use of algorithm. Here is the fine example of disambiguation of sentence.

- "She went to the bank to withdraw her amount,"

As we know the word "Bank "refers to the financial building. On the other hand, another example in a sentence,

- "Aslam sat by the bank of the river"

It obviously refers to a landform. Computational linguistics will be interpreted this situation after reading the context which is embedded to the sentence. Models like BERT excel at WSD will analyze the structure of the sentence and words surrounded by context (Goddard 2011)

Machine Translation"

It is not easy to translate text of one language to another word by word. It requires the capturing of the meaning of words accurately in semantics field. When it comes to the computational linguistics it will ensure translation system like google translation not only try to covert the words but also save their meaning with the context which is embedded to it. More specifically the idiomatic expression is very difficult to translate because there is no one to one connection between the words combination and their meaning.

For example; English idiom;

- **In English:** "I am feeling blue" (expressing the sadness.)
- **In Punjabi:** "Mera Dil Udaas ay" (capture sense of emotional sadness)
- **In Pashto:** "Za Khafa Yam" (express emotional sadness)
- **In Arabic: "**Qalbi Maksour" (deeper emotional sadness)

By computational linguistics these expressions can be translated semantically by conveying the same emotional meaning in another language rather than a literal translation.

Computational Methods in Natural Language Processing

There is certain system that computational models used to explore the various aspects of language. Like finite state machine, context free grammar, and regular expression to identify the set patterns in language. It works in such a unique way of tokenizing the sentence, parsing sentence tree, or analyze the field of morphology. It can be understood by analyzing the rule based morphological system to decomposing the word "unhappiness" mainly into its root "happy" and affixes ("un"- "ness") these systematical approaches are very helpful to provide the valuable insights into structured language(Clark, Fox et al. 2012).

Current Trends in NLP

The field of NLP is progressing quickly, determined by advancing bounds in AI, especially in Deep learning. Probably the most prominent patterns incorporate multilingual NLP, the strength of transformer-based models, and expanded center and ethical focus on low-resources language.

Relationship Between Semantics and Computational Linguistics

Semantics, as a systematic study of meaning tries to make an investigation and the significant importance of meaning in a language, assumes an important part in computational linguistics as it includes understanding and addressing the importance of words and their meaning phrases, sentences, and larger data computationally. Computational linguistics depends on semantic examination to empower machines to decipher human language in a manner that lines up with human comprehension. For example, in natural language processing (NLP) applications, for instance, sentiment analysis, representation of semantic assist frameworks with deciding if a sentence communicates a positive, negative, or natural sentiment.

For instance, the sentence

"The movie was a work of art"
Conveys positive sense, while "The movie was a calamity" passes negativity task feasible just on through semantic comprehension..(Gliozzo and Strapparava 2009)

Modeling Meaning and Computational Linguistics

As we know that human language is a complex phenomenon because it encompasses different layers of meaning. Even a single sentence can be complicated due to its linguistics unit constructions. There are different theories and the researches exploring the different constructions of meaning. different discussions have been generated on the meaning making process. It differs on the basis of theories and notions. This section deals with the modeling meaning in computational linguistics. It faces various challenges just because of the complexity of human language. There are several challenges as it is suggested but ambiguity is one of the major difficulties out of them. According to the context demand words and sentence have multiple meaning which make it problematic for computational linguistics to understand the exact interpretation. For instance, the word Bank can have dual interpretation according to its contextual demand. Either it is financial institution or the side of the river. It is difficult for machine to read the word embeddedness and the broader context attachment due to lack in real world experience. (Bolshakov and Gelbukh 2004)

Influential Model in Natural Language Processing (NLP) BERT

It is the model that introduced by google which is a language model optimization tool to understand the meaning of language text and context. It has the tendency to use the bidirectional transformer architecture meaning. It has the ability to read the text in both ways like it reads both directions of the text simultaneously. This make the model able to read the text from preceding and succeeding words in a sentence. BERT, the model has the specialty of reading deep comprehension, such as question answering, sentiment analysis, and named entity recognition. The model innovation lies in its pre trained objectives. Like masked language modeling(MLM). During the training procedures words randomly masked in a sentence. The model tries to predict these masked words on the bases of context provided by the surrounded words. Another objective of this model is next sentence prediction(NSP). BERT is capable to capture the relationship between pairs of sentences by predicting a next sentence follows the first one naturally. (Bates 1995)

Enhancing Multilingual and Cross Lingual in Computational Linguistics

Enhancing the understanding of multilingual and cross lingual under the shadow of computational linguistics is difficult for developing language technologies which serve to the global population. There is an important model that serve to categories the language text is multilingual BERT(Mbert) and another cross lingual language model(XLM-R). These are highly trained to read the text of different languages simultaneously. The shared feature of representation allows the particular model to transfer the knowledge from high resource language like English to low resource language like Punjabi language. There is an ability to provide the task through machine translation, cross lingual search and multilingual question answering of different languages. (Tsai and Roth 2016) Understanding of cross lingual is beneficial technique for languages like zero shot and few shot learning, where a particular model has the ability to perform task in other language on the behalf of trained data of first language. For example, a sentiment analysis model is trained in English; language and able to analyze in Punjabi language.

Material and Methods

The research methodology is the procedure which is used by the researchers to gather data for resolving problems of investigation and design of the research comprises of the whole procedure which is conducted research" (Ahmad, Farhat & Choudhary, 2022, p.524). This study opts a qualitative and exploratory approach to examine the capability of AI language models to cater the dimensions of semantics and pragmatics in the context of language documentation. The research methodology is a process of collecting and analyzing the data gathered by researcher from the different yet relative sources (Rao et al., 2023; Sadaf et al., 2024).

Data Collection

For the current study, the data was collected through the different sources such as real world linguistics examples were very helpful to understand the nature of the data. The published documentation of under sourced languages, such as digital archive (ELAR, DOBES) grammar and fields notes were also became the source for collecting the data.

Data Analysis

After collecting the data, there is a need to analyze it. For this purpose, thematic linguistic analysis was applied to analyze the data. The researcher has opted semantic and evaluation criteria to check the lexical accuracy in the form of synonyms and antonyms use. For pragmatic, the pragmatic evaluation was done to check the contextual interpretation and appropriateness of speech act.

Semantic Evaluation

There are different AI models having s strong capacity for semantic interpretations, most importantly in high resource language like English. When providing the synonym for "happy" respons would be "joyeful" or "cheerful"which is an accurate understanding lexical relations and semantic field. In low resource language such as Punjabi and Yoruba. The performance of model would be mixed.

- **Prompt:** what does "khush" mean inn Punjabi?
- **AI output:** it means happy or pleased.
- **Evaluation:** Accurate.
- **Prompt:** while giving a synonym for "egbon" in Yoruba language.
- **AI output:** younger sibling.
- **Evaluation:** incorrect, "egbon" actually refer to older sibling, suggesting a semantic reversal.

These results shown that models can provide the basic meaning in many languages, semantic accuracy still weakens in representation of linguistic systems.

Pragmatic Evaluation:

The accurate understanding of pragmatic seemed bit limited, more specifically in in low context culture setting.

- **Prompt:** "Can you open the window?"
- **AI Output:** yes, I can.
- **Evaluation:** The model interpreted the utterance at literal level by missing its polite functions rather than a question about ability.
- **Prompt:** "He stole my heart."
- **AI Output:** "He is a thief."
- **Evaluation**: This shows a failure in the interpretation of the idiomatic expression.

Table 1
Summary of Findings

Evaluation area	High resources languages
Semantic accuracy	High
Pragmatic appropriateness	Moderate
Idiomatic interpretation	Context dependent
Cultural context awareness	Limited

Description

These findings detail a sharp gap between semantic and pragmatic modeling in AI language system, having a greater disparity evident in under documented languages.

Ethical Considerations

The human participation is not directly take part of this study as it solely relies on publically accessible platform and linguistic data. However, it is important to take care of the respect and integrity of endangered data of languages. It is important in particular to acknowledge the source, communities and avoiding the misuse of sensitive and cultural information. For the sake of referencing, the sources of indigenous languages are cited according to ethical considerations.

Discussion:

It reflects through the analysis of the intersection of semantics and pragmatics in AI language model in language documentation that how important the competence of pragmatics for the accurate language documentation. This fact is considerable that AI can assist in understanding of linguistic capabilities but it can be analyzed that human language interpretation cannot be replaced specially in language documentation.

Conclusion

As the global demand of preserving the endangered languages is increasing, artificial intelligence or AI models are offering an unprecedented opportunity to assist the task in this critical work. This research study tries to investigate hoe AI language models navigate the domains of language and help to assess their contribution in language documentation. The findings show a clear picture how AI language models are strengthening in semantic capabilities but still lacking in the domain of pragmatic. This vary gap even in the modern era possess a substantial challenge, as pragmatic is central to understand how language functions within a social and cultural environment.

Recommendations

Based on the findings of this research study, it proved that while AI language models show promise in semantic task but it remains underdeveloped in pragmatic competence especially in the context of endangered and culturally embedded languages. In order to enhance the role of AI in language documentation, the researcher proposed given recommendations:

I. AI models should not operate in isolation. They should be integrated with human linguistic expertise.
II. AI language models should be trained on culturally diverse and low resources data.
III. AI language models must include pragmatic benchmark.

References

Abbas, T., Farhat, P. A., & Rasheed, B. (2024). Conversational Analysis of Political Talk Shows by Pakistani Politicians using Discourse Markers. *Pakistan Languages and Humanities Review, 8*(2), 701–711

Abbas, T., Soomro, A. R., & Abbasi, I. A. (2025). The Impact of Shortened Words Usage on Communication Effectiveness among Pakistan's University Students. *Journal of Arts and Linguistics Studies, 3*(1), 49-66. https://doi.org/10.71281/jals.v3i1.207

Ahmad, A., Abbasi, I. A., Abbasi, R. H., & Rasheed, B. (2025). Exploring the Intricate Relationship between Semantics and Computational Linguistics. *Liberal Journal of Language & Literature Review, 3*(1), 164-181.

Ahmad, A., Farhat, P. A., & Choudhary, S. M. (2022). Students' Insights about the Influence of Text Messaging on Academic Writing Skills. *Journal of Development and Social Sciences, 3*(4), 522-533. https://doi.org/10.47205/jdss.2022(3-IV)49

Ahmad, A., Maitlo, S. K., Rasheed & Soomro, A. R., Ahmed, A. (2021). Impact of Phonological Instructions in the Enhancement of ESL Learners' Pronunciation. *Remittances Review, 6*(1), 94-109. https://doi.org/10.33182/rr.v6i1.125

Ahmad, A., Rao, I. S., & Rao, M. S. (2023). ESL Students Anxiety in English as a Second Language Learning from The Perspective of Linguistic Skills. *Pakistan Journal of Humanities and Social Sciences, 11*(4), 3943-3951.

Ahmad, A., Sanober, R. S., & Cheema, M. I. (2024). ESL Learners Attitude towards Metacognition Approach for Learning Creative Writing at University Level. *Journal of Development and Social Sciences, 5*(1), 01-14. https://doi.org/10.47205/jdss.2024(5-I)01

Bates, M. (1995). Models of natural language understanding. *Proceedings of the National Academy of Sciences*, *92*(22), 9977-9982. https://doi.org/10.1073/pnas.92.22.9977

Benjamin, M. (2018). Hard Numbers: Language Exclusion in Computational Linguistics and Natural Language Processing. Proceedings of the LREC 2018 Workshop "CCURL2018–Sustaining Knowledge Diversity in the Digital Age.

Bolshakov, I. and A. Gelbukh (2004). Computational linguistics models, resources, applications, Serie Ciencia de la Computación.

Cheema, M. I., Maitlo, S. K., Ahmad, A., & Jalbani, A. N. (2023). Analyzing the Portrayal of The Characters in Cathrine Mansfield's Literary Novel Bliss by Using Critical Discourse Analysis. *International Journal of Contemporary Issues in Social Sciences (IJCISS), 2*(4), 225-231.

Church, K., & Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational linguistics*, *19*(1), 1-24. https://aclanthology.org/J93-1001.pdf

Clark, A., et al. (2012). The handbook of computational linguistics and natural language processing, John Wiley & Sons.

Farhat, P. A. (2019). The Effect of Computer Assisted Language Learning (CALL) on English Language Learners' Pronunciation in Secondary School in Pakistan.

Gliozzo, A. and C. Strapparava (2009). Semantic domains in computational linguistics, Springer Science & Business Media.

Goddard, C. (2011). Semantic analysis: A practical introduction, Oxford University Press, USA.

Grishman, R. (1986). Computational linguistics: an introduction, Cambridge University Press.

Jalbani, A. N., Ahmad, A., & Maitlo, S. K. (2023). A Comparative Study to Evaluate ESL Learners' Proficiency and Attitudes towards English Language. *Global Language Review, VIII*(II), 446-455. https://doi.org/10.31703/glr.2023(VIII-II).36

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *Language*, *39*(2), 170-210. https://doi.org/10.2307/411200

Maitlo, S. K., Kalhoro, I. A., Soomro, A. R., & Ahmad, A. (2024). Exploring the Negative Impact of Short Message Service (SMS) Texting on Academic Writing Skills at University Level. *Policy Research Journal, 2*(4), 2327-2333.

Rao, I. S., Jeevan, S., & Ahmad, A. (2023). Impact of Metacognitive Strategies on Creative Writing of ESL Students at College Level in District Lahore. *Global Language Review, VIII*(I), 315-324. https://doi.org/10.31703/glr.2023(VIII-I).29

Rasheed, H. R., Zafar, J. M., & Munawar, N. (2024). Emerging Trends of Assessment and Evaluation toward Students' Learning in Early Childhood Education: An Analysis. *Remittances Review, 9*(3), 442-456.

Sadaf, H., Rasheed, B., & Ahmad, A. (2024). Exploring the Role of YouTube Lectures, Vlogs, and Videos in Enhancing ESL Learning. *Journal of Asian Development Studies, 13*(2), 657-670. https://doi.org/10.62345/jads.2024.13.2.52

Tsai, C. T., & Roth, D. (2016, June). Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 589-598). https://aclanthology.org/N16-1072.pdf

Younus, J., Farhat, P. A., & Ahmad, A. (2023). Analyzing The Factors Involvement in Declining Kalasha Language. *Pakistan Journal of Humanities and Social Sciences, 11*(3), 3520-3529.