FUTURITY OF SOCIAL SCIENCE

DOI: https://doi.org/10.57125/FS.2023.06.20.02

How to cite: Aliyeva, G. B. (2023). Text Linguistics and the Use of Linguistic Data in Modern Technologies: Prospects for Development. *Futurity of Social Science, 1*(2),. 18-29. https://doi.org/10.57125/FS.2023.06.20.02

Text Linguistics and the Use of Linguistic Data in Modern Technologies: Prospects for Development

Gulchohra Babali Aliyeva

Doctor of Sciences, Professor, Azerbaijan State Marine Academy, Baku, Azerbaijan, https://orcid.org/0000-0003-2266-947X

***Correspondence email:** aliyevagulchohra59@gmail.com.*

Received: March 11, 2023 | **Accepted:** May 23, 2023 | **Published:** June 20, 2023

Abstract: This study explores text linguistics as a branch of linguistics that analyses the structure, meaning, and other aspects of texts. The paper aims to analyse the use of linguistic data in modern technologies, in particular, to develop applications that can understand and generate linguistic content. To achieve this goal, the research uses the methodology of qualitative analysis and is based on the work of textual theorists. The study is based on a qualitative analysis of examples collected from the BNC corpus. The instructions for obtaining data are to use textual forms located at specific positions in the co-text. The use of qualitative analysis for data acquisition is useful for studying phonology, morphology, morphosyntax, and general linguistic facts. The results confirm the promising possibilities of text linguistics and the use of language data in technology, in particular the growth in the volume and availability of language data and the application of new methods of text analysis. The conclusions point to the connection of text linguistics with other fields and the expansion of the capabilities of semantic search engines and automatic text processing systems. In addition, the article emphasises the importance of digital resources and the study of phrases as effective tools for linguistic analysis. The analysis of three illustrative examples taken from the British linguistic corpus is carried out with the aim of studying textual and discourse phenomena in the use of language data. A certain inconsistency between the factors has been revealed, which leads to a blurring of the linguistic facts that the authors of the article tried to investigate. It should be noted

that the selected examples, which were created to support the theory, are artificial, which makes it difficult to analyse the context. This makes it difficult for native English speakers to evaluate these examples. This is especially true for the first example, where the syntactic constructions do not match the context. This deviation was observed at the level of the sentence's information structure. As for the second example, the discomfort observed can be explained by the lack of motivation to use the referent that was introduced into the context as a reference marker that is independently associated with it in the place where the phrase appears in the text. There is also a clear polyphonic inconsistency in the origin of the point of view that characterises each segment of the sentence. This inconsistency is related to the choice of the reference marker.

Keywords: linguistic text analysis, corpus linguistics, lexical models, syntactic models, unstressed pronoun.

Introduction

For a long time, the study of text was inseparable from rhetoric. Starting with ancient rhetoric, text was considered exclusively a literary field. It was only in the mid-twentieth century, after the development of semantic theories and research in phonetics and pragmatics, that text became an object of linguistic research. Thus, compared to other linguistic studies, textual analysis is a very young field of research. Text analysis begins with the search for rules that can explain the textual skills of native speakers. This approach led to the development of generative grammar (Chomsky, Roberts & Watumull, 2023). Proponents of this approach tried to explain native speakers' textual competence by using well-defined rules to predict speakers' intentions for text construction. This approach soon proved to be ineffective because a text could not be approached with the same tools as a sentence. Subsequently, discourse and its context narrowed down the linguistic research of scientists. The development of cognitive psychology has led to the emergence of several partial text theories for automatic test analysis (Fox, Jackson & Crawford, 2023). In any case, and from the perspective of different theories, text linguistics has always been closely related to didactics, which is the study of text that aims to shed light on certain complex issues that go beyond the scope of traditional grammar. Thus, before it became a scientific discipline, textual linguistics first appeared as a didactic method. This dual use of textual linguistics continues to this day.

Research Problem

The field of language learning has been witnessing the development of new technologies for several decades, which are increasingly replacing traditional methods for two main reasons: on the one hand, the use of multimedia tools provides endless possibilities for integrating different media (texts, images, photography, audio, video or 3D, etc.); on the other hand, interactivity develops human-machine interaction, thus providing a certain autonomy in choosing a language learning path (Gillings & Mautner, 2023).

One of the most important distinctions for the present work is that between a sentence and a statement. A sentence is a unit of writing (i.e. text). It can be simple or complex, composed of several syntactic clauses. The linguists Raković, Iqbal, Li, Fan, Singh, Surendrannair & Gašević (2023) proved that spoken language does not manifest sentences. An oral utterance contains syntactic clauses but is not a sentence. The minimum units of spoken text are intonational units and periods. Pan (2022) defines a period as an integrative unit of a sentence characterised by a final intonation. As for the language as a whole, from a functionalist point of view, it is a “treasure” (Raphael, 2023), thus a structured set of linguistic resources that have been stored over the centuries of use by its speakers, allowing them to express themselves and communicate with others in a given situation.

Thus, the contribution of new technologies has enabled the development of corpus linguistics. From the very beginning, corpus linguistics was seen not only as a tool for making observations of linguistic phenomena but also as a methodology to promote an operational approach to teaching and learning.

Research Focus

The present work focuses on the problem of the choice and nature of language data, as well as their use, which is certainly central to the linguistic approach. Linguistics is perhaps the only science that positions language as both an object and a means of research (the latter being created as a meta-language). The paper presents the type of linguistic data, the choice, and the nature of the data used mainly in studies of discourse reference.

Research Aim and Research Questions

The main goal of this paper is to show that a basic linguistic phenomenon (such as anaphora) is often misidentified based on only one type of data (e.g., invented examples), and another type of data (natural, attested data) is not taken into account. In this context, the following issues of the presented work are highlighted:

1. How often is the linguistic phenomenon of anaphora misidentified in linguistic research based on invented examples?

2. To what extent can other types of data, such as natural, attested data, help to determine the correct characterisation of anaphora?

3. What factors influence the incorrect definition of anaphora based on the examples you have made up?

Literature Review

Language is distinguished as a system, as a resource for use that exists independently of how it is used. However, a language without morphosyntax loses its schematicity and accuracy (ZAHLER, 2023). Except in formal models, this type of judgement cannot be made independently, according to functionalist models of language, but this is not consistent with formal approaches that treat syntax as "autonomous" from semantics, phonology, or pragmatics (Schleppegrell & Oteíza, 2023). Morphosyntax, on the other hand, is the expressive shell of language for functionalist approaches (Yang & Liu, 2023). Similarly, syntactic constructions carry meaning and are not semantically neutral (Schneider, 2022). As for lexical items, their interpretation in the context of use places them at the boundary between language as a system and its actual use in a particular situation. In fact, the generally accepted meaning of these units in a language is always modified according to the context that affects the utterance (Laske, 2022). When it comes to the relationship between the meanings of lexical items in an expression, it can also lead to misunderstandings caused by the incompatibility of the meanings of words that appear in the same grammatical construction (Liu, Jhang & Lee). However, at the level of translation co-authorship, the subject can provide a more imaginative interpretation of this meaning (synecdoche, personification, oxymoron, etc.) according to the context and the speaker's or author's own perception of the expression that includes a priori incompatible elements (Mautner, 2016). However, it is worth remembering that such an interpretation may lead to incompatibility of semantic levels at the literal level, depending on the specific situation.

The characteristic that is most used in modern linguistics in terms of analysed data is its authenticity, not that it was created by the linguist himself.

The disadvantages of this approach are well known: circularity can occur when a linguist (even unconsciously) ensures that his theory or hypothesis about the functioning of a particular aspect of language or discourse is reflected in the very material created for demonstration.

Research Methodology

The study uses qualitative analysis of examples collected from the BNC (British National Corpus). The instructions for extracting data from the corpus are to use textual forms located at specific positions in the co-text. The results of the study are subject to qualitative analysis on a case-by-case basis, and many of the forms identified are irrelevant to the research objectives. Obtaining data in this way is useful for studying phonology, morphology, morphosyntax, and general linguistic facts. The relative frequencies of occurrence of micro-opposition elements, for example, the use of the temporal adverb "now" in English discourse, are used to determine which element is marked and which is unmarked. In order to show the different types of data in the text, data belonging to the first type are represented by Latin characters, and data belonging to the second type are represented by italics. However, the main task at the data level, apart from selecting the data, is to represent, process, and evaluate the data by the linguist. Seven other aspects of context need to be taken into account: the context of measurement, the pre-established discourse, the type of speech event, the socio-cultural framework established by the text, interchange, exchangeability, and the situation of utterance. The situation of utterance is the most important of these aspects, and it serves as the starting point for the construction of discourse. The context is constantly constructed and revised by the user as the discourse unfolds. Text, in the sense of the act of speaking or writing, reflects the character of the speaker and non-verbal cues such as gaze and gestures. Text has a linear structure, unlike discourse. It functions as a sequence of perceived signals created by the speaker or author for the perceiving reader or listener to understand the discourse associated with a particular piece of text and context. However, the text itself does not determine a particular interpretation that can be made later. This role is played by the discourse.

Results

In the field of linguistics, it is common to create corpora of various types of utterances and use them in research. In particular, this approach is a particularly desirable tool for gaining a more detailed and realistic understanding of the functioning of language and its possible uses. Compared to the traditional method based on fictional examples, statistically valid data obtained through corpus analysis lends much more credibility to the phenomena under study. Although other research methods exist, the use of corpora is one of the most effective approaches. However, in approaches aimed at identifying discursive and functional phenomena, the interest of such access may be limited.

Lin & Adolphs (2023) point out that the use of indexical reference in discourse takes precedence over quantitative reference. Meaningful evaluation of the reference to any expression according to its context is a priority. It is difficult or even impossible to obtain data of each type or subtype with the help of instructions for a specific corpus of texts. However, modern corpus analysis includes accessible annotated corpora that capture facts at higher levels than just morphosyntax.

Corpora are useful for studying discourse phenomena. These texts are created and processed by speakers and interlocutors to achieve their personal communicative goals. However, the use of corpus data can have a certain drawback: if in a study or discourse analysis, extracts from the corpus are presented without any indication other than their source (which is already the beginning of recontextualisation), they can lose their original context. This can lead to the fact that, although

proven, they become artificial, similar to fictional data. Therefore, in order to obtain the most correct results, the recipient of the research should try to recontextualise them in accordance with the co-text, which will allow access to the discourse associated with these textual fragments.

For example: "*Tamar had no alternative but to do as he said, and they rode from the yard without another word. How bitterly she regretted now that she had not revealed the episode with Davis before her marriage to Stephen" (http://www.natcorp.ox.ac.uk/). The example is about evaluating the discourse that may be associated with this sequence, not about determining the correctness of the form of the passage as a piece of text that can have any meaning. This implies that it is necessary to consider the context of use that the reader of the passage has deduced in order to create a discourse. An example that illustrates this fact is the use of the temporal adverb "now" in English discourse. In the example taken from the BNC corpus, boldface is used to highlight the words "went" and "regretted now", indicating that it is when the trip happened (preterite) that the reference to the adverb "now" is explained.

In this regard, it is noted that there are a significant number of elements in the above passage that require a certain context for their interpretation. This requires the reader to make several assumptions and establish connections between different situations and individuals. For example, the proper names Tamar, Davis, and Stephen are mentioned, indicating that their relationships are relevant to further understanding of the text. Also noteworthy are specific references to the courthouse and "court", the episode with Davis and her marriage to Stephen, and the use of ellipsis, pronouns (he, she, they, her), and the possessive pronoun.

The reference in the first line probably refers to "Tamar" as the default. The subsequent reference to "he" may be predictable (but it is not the only possible interpretation here). The pronoun "she" and the possessive pronoun "her" may refer to "Tamar", but not every reader will necessarily understand that it is a feminine noun. In any case, the reader will make the anaphoric connection indirectly, leading to the conclusion that "Tamar felt regret when she rode a horse", but it can also be understood as riding a bicycle, motorbike, etc. The author does not pay attention to the two "local" inferences: one of the persons spoken of is "Tamar" and she is part of a group of persons referred to by this pronoun.

In fact, it is difficult to determine the exact contextual meaning of the word "ride". These people can be riding together or separately. If the vehicles are bicycles or horses, then this can be considered a typical situation. But without more information, it is impossible to say for sure that everyone is riding their own bicycle, horse, etc. Of course, interpreting the adverb reduces the immediate problems, but the word "rode" cannot be contextualised without more information.

It is obvious that in this case, we are dealing with an example of "represented, personal thinking and perception" (Original, Cox & Udell, 2023): these thoughts are associated with Tamar, who perceives them as a subject of consciousness, and not (directly) as the narrator's thoughts. This suggests that in the second sentence, the adverb "now", with its deictic function, refers not only to the external temporal context indicating the past tense expressed by the preterite in the verb "went" (according to Lee) but also to the idea that Tamar has just articulated: namely, "that she felt helpless and unable to do anything other than what he (and therefore, by this assumption, "Davis") told her to do". It was this thought that most likely made her regret now that she had not disclosed (to Stephen, no doubt) the Davis episode before her marriage to him. Thus, the situation is much more complicated. So, we can see here the extent to which the reader is forced to make a number of inferences when reading this passage, and how important it is to use the linguistic data in the corpus

to make the necessary connections to fully interpret this passage, presented completely out of context.

In the second example, "*The guy in my story picked [Ø]...got [Ø]...and picked up all those tools" (http://www.natcorp.ox.ac.uk/), the phrase challenges the expected distinction between "configurational" and "non-configurational" languages in generative grammar. To this end, it is demonstrated that empty squares are often found in both subject and non-subject positions in spoken English. This contradicts what is claimed in generative grammar, where English is classified as a prototypical "configurational" language (with the help of main and secondary clauses, which allows the linguist to represent the syntactic structure by means of a tree). After analysing this passage, we can conclude that the speaker hesitates in choosing the optimal verb. This is probably due to successive hesitations that result from considerations of the verb's meaning. At the end of this process, the speaker decides to choose the same verb form as at the beginning. However, this is followed by a particle that indicates that the verb can have the meaning of "choose", "pick up" or "take hold of".

In Example Three, "*The guy in my story picked them [...] got them [...] and picked up all those tools" (http://www.natcorp.ox.ac.uk/) is definitely a case of over-analysis, where more is visible in the user's textual production than can be reasonably assumed, according to the reader's own point of view. In other words, natural data can reveal facts about language and discourse that are real and cannot be created artificially. However, linguists who use these data do not always provide the context in which they were collected, although this context is important for a complete and accurate analysis of the language data. In addition, the reader may make mistakes in their analysis because they may be artificially driven to demonstrate a particular theoretical conclusion.

Thus, the natural data that researchers collect to study the indexical procedures of using language data in corpora (anaphora, anadeixis, and deixis) allow us to postulate theories of the functioning of reference in discourse that differ significantly from traditional theories based on scripts and recorded language data. These examples, as they arise in the spontaneous use of a speaker or writer when they do not pay attention to the conformity of their statements to modern linguistic standards, can more accurately reveal the reality of language as a biological system.

These examples, or so-called "asterisk" examples, are often not fully considered or stigmatised in some linguistic approaches. Some linguists apply strict restrictions on the use of unstressed pronouns with indirect anaphoric references. Indirect anaphora requires the interpreter to draw conclusions from the context, often based on knowledge about the world, the type of text, etc. Interpreting indirect anaphora requires a semi-automatic transition from what is explicitly mentioned or focused on to a referent that has some connection to it. In some linguistic approaches, comments on language data are marked with symbols that indicate their status, such as an asterisk to indicate ungrammaticality, an exclamation point for a semantic anomaly, etc. However, it is sometimes difficult to clearly distinguish between the different statuses of speech data. Pronouns can be used to confirm the obviousness of the referent to the speaker, but not to take into account the point of view of the interlocutor or reader. Such examples can be used to minimise the effort of finding the referent in the context. Indirect anaphora can also be expressed through pronouns or demonstrative phrases, although this is considered marginal. Research shows that different linguistic data can have different statuses and the use of unstressed pronouns for indirect anaphors depends on the context and linguistic situation. However, such examples are sometimes rejected or stigmatised by some linguists as "marginal" or defective in some aspects (and therefore not worthy of serious consideration).

Discussion

The study has shown that the frequency of incorrect definition of the linguistic phenomenon of anaphora based on invented examples in linguistic research may vary depending on the specific study and researchers. However, in general, several factors should be taken into account that contribute to the correct definition of anaphora: collecting real examples (the best method for defining anaphora is to analyse real texts and speech acts rather than inventing examples. This provides the right context and thus, it is possible to evaluate the use of anaphora in real communicative situations), attention to context (researchers should be attentive to the context in which the examples are used. It is worth considering the structure and connections between sentences to correctly identify cases of anaphora), studying relevant theoretical approaches (the existence of theoretical approaches to anaphora can help researchers analyse this phenomenon more thoroughly and avoid misidentifications). These findings are in line with Dai (2023), who argues that despite the factors that contribute to the correct definition of anaphora, the creation of invented examples can be useful for illustrative purposes, but their use should not be the basis for a general definition and generalisation of anaphora. In the same spirit, some linguists apply strict restrictions on the use of unstressed pronouns with indirect anaphoric references. In particular, they distinguish between direct anaphora, where the anaphoricity is inherent, and indirect anaphora, which requires the interpreter to draw conclusions from the context, often based on their knowledge of the world, the type of text, etc. (Cooper, 2023). More precisely, the interpretation of indirect anaphora requires a semi-automatic transition from what is explicitly mentioned or focused on to a referent that has some connection to it, whether through a part-to-whole, case-to-type, or metonymic relation. For linguistic data retrieval to be effective, a potential relevance or adequacy relationship must exist between the indirect referent and the subsequent discourse.

In the light of the second research question, it should be noted that data, such as natural, attested data, can help to determine the correct characterisation of anaphora to a certain extent. For example, if there is statistical data on the use of anaphora in a particular language or on its distribution among different speaking groups, this can give an idea of its prevalence and adequacy in different contexts. In the same context, Udomphol (2023) notes that counting natural data can also help in studying the relationship between anaphora use and the sociocultural context, for example, how age, social or geographical factors influence anaphora choice. However, it should be added that it is also important to take into account that the use of anaphora can be highly individual and depend on one's own preferences, speech style, and other personal factors that may be difficult to reproduce in natural data, hence the impossibility of assessing the status of anaphora language data. The status of linguistic data is closely related to its choice and nature. This status is usually expressed by a prefixed symbol immediately preceding the data. In the corpora, an asterisk is used to indicate ungrammaticality, an exclamation point is used to indicate a semantic anomaly in the interpretation of an example, and a pound sign indicates that its use as a (potential) utterance is (or would be) unnatural in this context. An additional symbol used for corpus-validated statements is the percentage symbol (%). It indicates that the type of example with the prefix occurs for many speakers in a given corpus of utterances, but not for all.

In practice, however, it is sometimes difficult to distinguish between the first three statuses. For example, in formalist work (formal syntax and semantics), a linguist often uses an asterisk to indicate any defect noticed in an example, whether syntactic, semantic, or discourse. However, this generalised practice remains neutral as to the value and function of asterisked examples for those who use them.

Regarding the third research question, the results of the study showed that there are several factors that can cause the misidentification of anaphora based on invented examples. For example, insufficient or inadequate linguistic competence. In this regard, Lin & Adolphs (2023) argue that faulty knowledge or understanding of the language can lead to misidentification of anaphora. For example, if a user does not understand the correct use of pronouns, distortions or misunderstandings of anaphoric relations may occur. The next factor is incorrect contextual understanding. According to Original, Cox & Udell (2023), the correct definition of anaphora requires an understanding of the context in which it is used. If the user misjudges the context or is not aware of additional information, anaphora misidentification may occur. But the lack of appropriate logical thinking skills is quite significant. After all, identifying anaphora also requires the ability to think logically and understand the surrounding text. If the user does not have these skills or they are not sufficiently developed, this can lead to an incorrect definition of anaphora. Other factors include insufficient contextual content and personal biases or unconscious biases. The examples that the user uses to define anaphora may be insufficient or not contain enough information for correct understanding. This can lead to misidentification of anaphora, and personal preconceptions or unconscious biases can also lead to misidentification of anaphora. If the user has certain beliefs or perceptions, this can distort their understanding and contribute to the misidentification of anaphora.

However, according to Dai (2023), anaphoric reference to implicit referents is more often the exception than the rule. According to Hashimoto (2023), indirect anaphora cannot be expressed with unstressed pronouns. On the contrary, Jalilbayli (2022) points out that indirect pronoun and demonstrative anaphoric data are less frequent than indirect anaphoric data preceding an indefinite article and appear more often in unplanned texts. The authors also argue that indirect anaphora cannot usually be expressed through pronoun or demonstrative phrases (Espada, Martínez, Rico & Sánchez, 2023).

In this context, McCoy, Smolensky, Linzen, Gao & Celikyilmaz (2023) similarly present attested passages from different types of work in American English as data for analysis. They provide examples of spoken and written language. However, the authors concluded that the use of pronouns or demonstratives to create indirect anaphors is marginal and uninteresting, and probably does not take into account the "hierarchy of givenness". Thus, unstressed pronouns occupy the top position, where they express only the cognitive status of "in focus", not the "activated" status. The status of "activated" is less active, occurring on the periphery of the participants' consciousness during speech.

Regarding the examples proposed by Meyer (2023), pronouns can be reformulated as follows: they can be seen as examples used to confirm that the referent is very obvious to the speaker or author, but that the interlocutor's or reader's point of view is not taken into account to better justify the reference in the context; either because they are common uses (so-called "collective" examples) that do not imply a specific referent and are therefore not considered important for understanding the purpose, or because the sender believes that in a working context, the interlocutor or reader will be able to easily find the target referent with a sufficiently vague statement based on the principle of least effort.

In Zinn & Mülle's (2022) study, which is in fact in line with Schneider, Hundt & Schreier (2019) in their view of unplanned spoken and written genres as "free" discourse, where freedom with respect to the standard of correctness is made "explicit" in the written word. Other authors in this regard consider passages of this type (as cited above) to be only "minor violations" of the hierarchy of givenness (Vogel & Hamann, 2023). This can be an example of "scientism" (Udomphol, 2023).

In a related study on pronoun usage in English and French, Sarvazyan, González, Franco-Salvador, Rangel, Chulvi & Rosso (2023) clearly show that when a distinction is made between central and peripheral implicit referents, the former type of pronouns are erased almost as quickly as the latter type of pronouns, both in the "implicit antecedent" and "explicit" condition. These results were also confirmed by the work of Piantadosi, S. (2023) on the presence of indirect pronouns that reveal the presence or absence of certain internal properties of the structure or type of utterance that cannot be represented exclusively by "standard" or "canonical" linguistic data.

The authors, by using an asterisk to indicate unacceptable examples of various types, express a position where syntax has a higher priority than other aspects of the utterance. This is not a typical approach among other linguistic frameworks, except generative linguistics. In generative grammar, the phonological and semantic components are "interpretive" for the syntactic representation of the sentence being generated.

Conclusions and Implications

The analytical analysis of three vivid examples from the British linguistic corpus has shown and described textual and discourse phenomena in the use of language data. Some confusion was found among them, in fact, between factors that should be considered separately. We can see that this leads to a blurring of the linguistic facts that we have tried to establish in the present work.

At the outset, it should be noted that the selected examples, which were constructed to substantiate the theory, have an artificial appearance, which makes it difficult to find context and analyse them. This makes them difficult to evaluate even for native English speakers. This is especially true of the first example. For this example, the syntactic constructions used are not appropriate for the context. This deviation was observed at the level of the sentence's information structure. As for the second example, the discomfort felt can be explained by the lack of motivation to use a referent that has just been introduced into the context with a reference marker that independently refers to it in the place where the phrase appears in the text. There is also an apparent polyphonic confusion about the origin of the point of view that characterises each segment of the sentence. This confusion is directly related to the choice of this reference marker.

In the third example, there was no indication of any problematic aspect of its internal syntactic or semantic construction, nor of its potential use in a particular context.

However, in all of these examples, the asterisk is preceded by a symbol. It was found that the problem is not in the syntax as such, but in factors related to perception: the intersection of perspective threads (a case of "polyphonic dissonance"), transformed into a potential utterance where, despite the presence of an asterisk, the problem is related to the consistent integration of the content of the two main constituent sentences, linked by anaphoric relations. The reader will need to "re-contextualise" the proposed passages, as far as possible, according to the co-text presented, which is problematic in such circumstances.

Another important aspect of language data mining is related to prosody. In fact, the first step to successfully putting an example into context, even in written form, should be the possibility and probability of imposing a certain prosodic structure on the textual elements involved. This will be a function of the information structure of each component of the complex example and, in turn, will allow us to determine the discourse units with which it can correspond. This was evident, first of all, in the analysis of the second mentioned relational structure used in Example 2, as repetition, in terms of possible prosodic realisation; it was also evident in Example 3, where two types of direct object production in spoken language, which depend on intonation, are considered.

Unfortunately, linguists who are far from the active norm reject formulations that are not particularly interesting to them on a theoretical level and are therefore considered impractical in describing such a phenomenon, even if they allow for a more accurate understanding of the mechanism of language data selection for a corpus.

Suggestions for Future Research

This research can focus on understanding the use of different sources of language data, such as social media texts or specialised domain corpora, and how they affect language understanding and generation. From the perspective of text features, it is possible to delve into specific linguistic features such as syntax, semantics, pragmatics, or discourse. By gaining a better understanding of these features across languages and contexts, researchers could improve the accuracy and reliability of natural language understanding systems. Research into the role of context in language understanding will allow us to focus on incorporating contextual information, such as discourse cohesion or speaker intent, into language models. This could help improve their ability to generate more coherent and contextually relevant responses. Furthermore, language data in modern technologies often reflects societal biases. Research could be conducted to identify and reduce biases in training data to ensure fair and unbiased results from language technology applications. Language data can be invaluable for documenting and preserving endangered languages. Future research could examine methodologies and techniques for collecting, processing, and analysing language data on endangered languages to aid in revitalisation efforts. Many languages and regions lack sufficient linguistic data due to limited resources or technological access. Future research could focus on developing innovative methods for collecting language data in resource-limited settings, such as using crowdsourcing, speech recognition technologies, or mobile apps. In addition, future research could focus on analysing the quality and diversity of language data used for machine translation, exploring methods for collecting multilingual parallel corpora, and developing methods for adapting machine translation models to resource-limited languages.

References

Chomsky, N., Roberts, I., & Watumull, J. (2023). Noam Chomsky: The False Promise of ChatGPT. The New York Times, 8. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Cooper, C. R. (2023). The identification of YouTube videos that feature the linguistic features of English informal speech. Applied Corpus Linguistics, 3(3), 100068. https://doi.org/10.1016/j.acorp.2023.100068

Dai, Z. (2023). Statistics in Corpus Linguistics: A New Approach by Sean Wallis. New York/Oxon: Routledge, 2021. ISBN 9781138589384 (PB: 44.95), ISBN 9781138589377 (HB: 160.00), ISBN 9780429491696 (eBook: 44.95), xxvi+ 382 pages. Natural Language Engineering, 29(1), 177-180. https://doi.org/10.1017/s1351324921000413

Espada, J. P., Martínez, J. S., Rico, I. C., & Sánchez, L. E. V. (2023). Extracting keywords of educational texts using a novel mechanism based on linguistic approaches and evolutive graphs. Expert Systems with Applications, 213, 118842. https://doi.org/10.1016/j.eswa.2022.118842

Fox, J. M., Jackson, R. A., & Crawford, K. R. (2023). News of Noam: Unpacking Media Coverage of Chomsky. International Journal of Languages, Literature and Linguistics, 9, 318-325. https://doi.org/10.18178/ijlll.2023.9.5.425

Friginal, E., Cox, A., & Udell, R. (2023). Corpus Linguistics and Writing Instruction. In Demystifying Corpus Linguistics for English Language Teaching (pp. 79-97). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-11220-1_5

Gillings, M., & Mautner, G. (2023). Concordancing for CADS: Practical challenges and theoretical implications. International Journal of Corpus Linguistics. https://doi.org/10.1075/ijcl.21168.gil

Hashimoto, B. (2023). Corpus of Founding Era American English: designing a corpus for interpreting the United States Constitution. Corpora, 18(1), 1-14. https://www.euppublishing.com/doi/abs/10.3366/cor.2023.0270

Jalilbayli, O. B. (2022). Forecasting the prospects for innovative changes in the development of future linguistic education for the XXI century: the choice of optimal strategies. Futurity Education, 2(4), 36-43. https://doi.org/10.57125/FED.2022.25.12.0.4

Laske, C. (2022). Corpus linguistics: the digital tool kit for analysing language and the law. Comparative Legal History, 10(1), 3-32. https://doi.org/10.1080/2049677X.2022.2063510

Lin, P., & Adolphs, S. (2023). Corpus linguistics. In The Routledge Handbook of Applied Linguistics (pp. 296-308). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003082644-25/corpus-linguistics-phoebe-lin-svenja-adolphs

Liu, C., Jhang, S., & Lee, S. From Gender-Biased to Gender-Specific and Gender-Inclusive Words: A Corpus-Based Study. https://doi.org/10.24303/lakdoi.2023.31.1.85

Mautner, G. (2016). Checks and balances: How corpus linguistics can contribute to CDA. Methods of critical discourse studies, 154-179. https://www.torrossa.com/en/resources/an/5018239#page=165

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. Transactions of the Association for Computational Linguistics, 11, 652-670. https://doi.org/10.1162/tacl_a_00567

Meyer, C. F. (2023). English corpus linguistics: An introduction. Cambridge University Press. https://assets.cambridge.org/052180/8790/sample/0521808790ws.pdf

Pan, Y. (2022). Intensification for discursive evaluation: a corpus-pragmatic view. Text & Talk, 42(3), 391-417. https://doi.org/10.1515/text-2020-0046

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. Lingbuzz Preprint, lingbuzz, 7180. https://lingbuzz.net/lingbuzz/007180

Raković, M., Iqbal, S., Li, T., Fan, Y., Singh, S., Surendrannair, S., ... & Gašević, D. (2023). Harnessing the potential of trace data and linguistic analysis to predict learner performance in a multi-text writing task. Journal of Computer Assisted Learning, 39(3), 703-718. https://doi.org/10.1111/jcal.12769

Raphael, E. B. (2023). Gendered Representations in Language: A Corpus-Based Comparative Study of Adjective-Noun Collocations for Marital Relationships. Theory and Practice in Language Studies, 13(5), 1191-1196. https://doi.org/10.17507/tpls.1305.12

Römer, U. (2004). Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. Corpora and language learners, 151-168. https://www.torrossa.com/en/resources/an/5001951#page=158

Sarvazyan, A. M., González, J. Á., Franco-Salvador, M., Rangel, F., Chulvi, B., & Rosso, P. (2023). Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. arXiv preprint arXiv:2309.11285. https://doi.org/10.48550/arXiv.2309.11285

Schleppegrell, M. J., & Oteíza, T. (2023). Systemic functional linguistics: Exploring meaning in language. In The Routledge handbook of discourse analysis (pp. 156-169). Routledge. https://www.researchgate.net/publication/348994754_The_Routledge_Handbook_of_Systemic_Functional_Linguistics

Schneider, G. (2022). Recent changes in spoken British English in verbal and nominal constructions. Broadening the Spectrum of Corpus Linguistics: New approaches to variability and change, 105, 173. https://www.torrossa.com/en/resources/an/5495945#page=180

Schneider, G., Hundt, M., & Schreier, D. (2019). Pluralized non-count nouns across Englishes: A corpus-linguistic approach to variety types. Corpus Linguistics and Linguistic Theory, 16(3), 515-546. https://doi.org/10.1515/cllt-2018-0068

Tomas, F., Dodier, O., & Demarchi, S. (2022). Computational measures of deceptive language: prospects and issues. Frontiers in Communication, 7, 792378. https://doi.org/10.3389/fcomm.2022.792378

Udomphol, P. (2023). A Comparative Corpus-Based Analysis of the Cross-Cultural Lexico-Grammatical Differences Between Master's Level Academic Writing in New Zealand and the United States (Doctoral dissertation, Auckland University of Technology). https://openrepository.aut.ac.nz/handle/10292/15834

Vogel, F., & Hamann, H. (2023). Legal Linguistics in times of language models and text automation: JLL Call for Abstracts (Deadline 31 March 2023). International Journal of Language & Law (JLL), 12, 1-7. https://doi.org/10.14762/jll.2023.001

Yang, Y., & Liu, X. (2023). Review of Multifunctionality in English: Corpora, Language and Academic Literacy Pedagogy: Zihan Yin and Elaine Vine (Eds.), Routledge, London and New York, 2022 (Paperback), ISBN: 978-0-367-72512-9. https://link.springer.com/article/10.1007/s41701-023-00156-9

ZAHLER, S. (2023). Some Issues in Usage-Based Methods: Contributions from Corpus Linguistics, Psycholinguistics, and Variationist Sociolinguistics. The Handbook of Usage-Based Linguistics, 73-90. https://doi.org/10.1002/9781119839859.ch4

Zinn, J. O., & Müller, M. (2022). Understanding discourse and language of risk. Journal of Risk Research, 25(3), 271-284. https://doi.org/10.1080/13669877.2021.2020883