ISSN Online: 2709-7625
ISSN Print: 2709-7617
Vol.6 No.4 2023

Corpus Linguistics and English Language Research

Muhammad Ramiz, Mphil scholar, Department of English, Institute of southern Punjab Multan, pakistan

rramiz370@gmail.com

Abstract:

This article explores the profound impact of corpus linguistics on English language research. Moving beyond traditional introspective and prescriptive approaches, corpus linguistics leverages large, electronically stored collections of authentic texts ("corpora)" to provide evidence-based insights into how English is actually used. We delve into the core principles of corpus methodology, showcasing its effectiveness in investigating diverse linguistic phenomena from word meaning and grammar to register variation and language change. The article further highlights the transformative applications of corpus findings in various domains within English language research, including lexicography, pedagogy, translation, and computational linguistics. Finally, we acknowledge the limitations and evolving landscape of corpus linguistics, paving the way for future directions in research and applications. This article serves as a comprehensive introduction to the vital role of corpus linguistics in unlocking the intricacies and dynamism of the English language.

Keywords:

Corpus Linguistics, English Language Research, Corpora, Methodology,Register Variation, Language Change, Lexicography, Pedagogy, Computational Linguistics.

Introduction

Corpus linguistics refers to the study of language using corpora - large, principled collections of natural language texts (McEnery & Hardie, 2012). Instead of analyzing constructed example sentences, corpus linguistics utilizes authentic language samples to reveal real patterns in how language is used. As a methodology, it has been applied extensively in English language research over the past few decades to derive new insights across various areas of linguistics. This paper provides an overview of corpus-based methodology and its application in key domains of English language research including lexicography, grammatical studies, historical linguistics, and sociolinguistics.

The underlying premise of the corpus linguistic approach is that the systematic analysis of authentic texts can reveal facts about language that introspection alone may not uncover (Biber et al., 1998). Computational analysis allows researchers to process significantly larger amounts of natural language data than would be feasible through manual analysis. Corpus linguistics techniques harness this computational power by applying sophisticated automatic and interactive modes of analysis using software tools. According to Xiao (2009), the key characteristics of corpus analysis include authenticity, frequency, breadth and depth of analysis, and empirical basis. Hunston (2002) outlines the core components in constructing and exploiting a corpus: 1) collection of texts; 2) addition of markup and annotations; 3) software tools to support analysis such as concordancing programs.

In concordance software, each occurrence of a search term is presented within the context of a few words that precede and follow it, allowing for examination of patterns. Corpora can vary considerably in scope and size, from small specialized corpora in the 1-million-word range to massive corpora containing billions of words like the Corpus of Contemporary American English (Davies, 2009). Corpora also vary in their balance and representativeness. For example, the Brown University Corpus of American English consists of texts from a wide variety of genres to capture a breadth of language varieties (Francis & Kucera, 1979). Corpora can also encode metadata tags at the text, paragraph, sentence or even word level to support more fine-grained contextual analysis.

While corpus analysis methodology offers some clear advantages, Biber et al. (1998) also discus several limitations. Large electronic corpora can lack useful contextual cues like the intent or mood of the writer/speaker. They may overrepresent particular varieties of language use while lacking insight into uncodified languages. Manual linguistic analysis still plays an important role in interpreting, validating and explaining the textual patterns that corpora reveal. But used judiciously, corpus linguistics serves as a powerful methodology that can uncover new insights across diverse areas of English language research.

Applications in Lexicography

One major application of corpus linguistics has been in the field of lexicography - the scholarly discipline of analyzing word use and meaning and compiling entries into dictionaries. According to Hanks (2012), almost all large dictionaries of English now rely heavily on corpus-based research, analysis, and citation evidence. He estimates that more than 250 dictionaries have utilized the 2 billion word Corpus of Contemporary American English (COCA) alone since its launch in 2008. Specific applications include tracking chronological trends in word use and frequencies to document new, emerging or obsolete terms (Berber Sardinha, 2014). This dynamic approach stands in contrast to traditional lexicography that relied on assembling citations gradually from limited text sources. Corpus analysis also allows exploration of typical grammatical contexts in which words appear and collocates - words that frequently co-occur together (McEnery & Hardie, 2012). This reveals underlying meanings and connotations that can enrich dictionary entries.

Grammatical Studies

In grammar research, corpus analysis techniques help uncover typical syntactic patterns and constructions that are difficult to perceive through manual inspection alone. Biber et al. (1999) utilized corpus analysis of spoken and written university registers to systematically investigate grammatical complexity dimensions like tense and aspect marking, passivization, and adjectival/adverbial modification. Comparable methods can be applied to study learner language development looking at acquisition orders of grammatical structures (Ellis & Barkhuizen, 2005). Corpus techniques like keywords and grammatical concordancing around pronouns also facilitate analysis of anaphoric relations - the links between pronouns and their referents (Hardie, 2013). Considerable debate persists around relying wholly on naturally occurring data versus the need for linguist-crafted examples in studying grammar (Biber et al., 1998). Nonetheless, as large corpora become available representing diverse genres, corpus evidence continues to gain relevance in grammatical research within applied linguistics.

Corpus Linguistics Methodology

The foundation of the corpus linguistics approach lies in the systematic methods applied to build, annotate, and analyze the textual data contained in corpora. This section will provide an overview of key components in corpus construction, the use of metadata and annotations, and common software tools that enable both qualitative and quantitative analytical techniques.

Constructing the Corpus

The first step in harnessing the power of corpus linguistics is to carefully construct the corpus to suit the intended research purpose. Corpora can vary tremendously in size, composition, balance, domain specificity, modality, and other attributes that impact analysis (Baker et al., 2006). Smaller corpora allow greater manual inspection but are limited regarding lexical and grammatical coverage. Larger corpora facilitate analysis at scale to uncover linguistic patterns too subtle or infrequent to perceive manually. For example, the Corpus of Contemporary American English (450 million words) and the British National Corpus (100 million words) are among the largest balanced corpora covering a wide range of everyday language across spoken, fiction, magazine, newspaper, and academic texts (McEnery & Hardie, 2012). In contrast, highly domain-specific corpora may contain 1-30 million words focusing on a niche area like medical journals or financial reports.

Representativeness refers to how well the corpus proportionally covers the full span of language use within the target domain (Teubert, 2005). Balanced corpora strive for representativeness by sampling evenly across demographics and genres. In contrast, parallel corpora contain source texts and their translations, allowing for comparative analysis across languages. Corpora also vary regarding domain specificity. Specialized corpora isolate language use within a domain but lack insight into general vocabulary. Reference corpora with wide coverage provide a standard for comparison. Multimodality is also relevant; spoken corpora encode cues like pausing absent in written text. Corpora can also be diachronic - capturing language use over historical time periods. Effective corpus construction requires deliberate choices regarding size, domain, balance, modality and other factors driven by the research purpose (Römer, 2011).

Metadata and Annotations

In addition to the texts themselves, corpora contain metadata tags and linguistic annotations that support deeper analysis. Structural markup encodes artifacts like titles, paragraphs, chapters etc. allowing concordance lines to retain valuable co-text for better interpretation (Sinclair, 2005). Metadata can specify publication details, author demographics, text genres and domains. Linguistic annotations classify parts-of-speech, syntactic categories, semantic categories etc. Manual annotation requires extensive linguistic expertise and resources but allows fine-grained searching. For example, semantic annotation enabled the pattern-based historical thesaurus project (Piao et al., 2016). Automatic annotation using natural language processing is more feasible for large corpora but less accurate. Treebank annotation captures full sentence syntax for grammatical searching. The various layers of metadata and annotations effectively expand the parameters for investigation beyond surface keywords.

Analytical Techniques and Software

Sophisticated software facilitates both quantitative and qualitative analytical techniques. Frequency lists reveal salient terms along with dispersion metrics indicating their distribution across different sub-corpora. Collocation statistics expose words frequently occurring together, revealing underlying semantic and grammatical relationships (McEnery & Hardie, 2012). Concordancing presents each search term within its co-textual context, allowing for human examination of patterns. With annotation, searches can specify parts-of-speech, semantic categories etc. for more precise results. Clustering methods like key keywords help automatically detect significant lexical differences across compare corpora (Scott & Tribble, 2006).

Widely used software tools include AntConc, MonoConc, Wordsmith Tools and CQPWeb offering concordancing, frequency statistics, collocations etc. (Hardie, 2012). Other environments like #LancsBox integrate multiple linguistic analysis capabilities (Brezina et al., 2015). Some graphical interfaces depict corpus search results through tree diagrams, dispersion plots, dendrograms etc. While the sheer data sizes involved restrict full manual inspection, these tools empower both targeted and exploratory modes of investigation. Quantitative techniques help surface interesting patterns that can then be qualitatively analyzed through manual concordance analysis. Used synergistically, these functionalities unlock deep insights from large bodies of natural language.

Limitations

However, corpus analysis has inherent limitations regarding contextual interpretation. Lacking prosodic cues in speech, discourse context, speaker intent etc. introduces ambiguity. Baker et al. (2006) give the example “hard drive” having multiple potential references. Variability in domain coverage also impacts the external validity of findings. Reference corpora better represent general language use but offer less internal validity for studying specific language varieties. Manual analysis retains primacy for explaining deeper meanings behind surface textual patterns. But used prudently, corpus linguistics offers access to a breadth and depth of realistic language samples that greatly complements other approaches.

Overview

In an overview, corpus construction, annotation, and investigative software functionality collectively facilitate both qualitative and quantitative analytical techniques. Applied judiciously, this methodology grants valuable evidentiary insights into actual language use across diverse domains. However, thoughtful corpus design, annotative detail, and contextual interpretation remain vital to properly exploit the power of modern computational corpus linguistics within English language research.

Utilizing Corpora in Lexicography and Semantics

Corpora have transformed the field of lexicography - the scholarly discipline of tracking and analyzing word use and meanings - over the past few decades. According to Hanks (2012), almost all major dictionaries rely extensively on corpus evidence, with more than 250 dictionaries having utilized the Corpus of Contemporary American English alone. Specific applications include identifying new and emerging vocabulary, charting frequencies of word usage over time, analyzing meanings and semantic categories, and examining typical syntagmatic patterns and collocates

The dynamic nature of language use means that dictionaries require continual updating, a monumental task done more efficiently with corpus analysis tools. Tracking frequencies over time can reveal rising or waning terms, prompting addition, re-definition or obsoletion of entries (Berber Sardinha, 2014). Collocation statistics expose typical word partnerships, often signaling distinctions between literal and implied meanings. Concordance lines furnish examples of real contextual usage rather than having to craft illustrative citations manually. Some semantically annotated corpora classify words by conceptual categories, enabling the study of semantic fields and prototypes (Piao et al., 2016). In these ways, corpora furnish comprehensive lexical evidence that shapes modern lexicographic practice.

Utilizing Corpora to Investigate Grammatical Phenomena

In addition to tracking vocabulary patterns, corpora grant insight into grammatical constructions by aggregating thousands of textual examples. Biber et al. (1999) performed a landmark study analyzing spoken and written university language samples from the TOEFL 2000 Spoken and Written Academic Language Corpus. Computationally intensive corpus techniques uncovered complex grammatical dimensions related to features like tense/aspect marking, passive voice, nominalizations, and clause subordination that are pervasive in academic writing but less common in speech.

Syntactic treebanks take annotation further by capturing full sentence structures. This enables targeted investigation of specific constructions, for example to study acquisition patterns in second language learners (Ellis & Barkhuizen, 2005). Hardie (2013) demonstrated using CQPWeb software to search for anaphoric relations signaled through pronouns and their antecedents. The sheer diversity of genres and domains spanned by large reference corpora surface grammatical patterns that introspection alone cannot. However corpus evidence supplements rather than replaces manual linguistic analysis (Biber et al., 1998).

Revealing Language Trends and Variation Over Time through Diachronic Corpora

Most corpora provide synchronic evidence reflecting language use at a snapshot in time. However, diachronic corpora compiling texts from different historical periods allow tracing language change. Seminal early efforts like the Helsinki Corpus charted syntactic changes in English literature dating back to the 8th century (Kytö, 1996). The Corpus of Historical American English chronicles American English vocabulary and usage since 1810 through genres like fiction, magazines, newspapers and speeches (Davies, 2012). Analysis revealed dynamic processes like colloquialization, whereby informal speech patterns gradually infuse into formal written English. Tracking movements of words across semantic domains also highlights shifting connotations and attitudes.

Diachronic corpora visualise language change in progress and displace traditional notions of fixed grammars and meanings. Variationist sociolinguistic studies also utilize corpora compiling the speech of different regional and social groups to analyze nonstandard dialects against received standards. Here corpora provide real evidentiary basis concerning questions of linguistic prestige and attitudes (Anderwald & Szmrecsanyi, 2009). Corpus aggregation of language samples across user demographics and situations is essential for studying variation along temporal, geographical and social dimensions.

Designing Specialized Corpora to Study Learner Language

Corpus analysis also facilitates study of second language acquisition patterns by compiling archives of learner language. The International Corpus Network of Asian Learners of English (ICNALE) sampled college student writing and speeches from 10 regions to examine lexical, grammatical and rhetorical features (Ishikawa, 2013). The LongDALE corpus tracked individual Chinese learners over months to chart interlanguage development (Meunier et al., 2013). Error-annotated learner corpora are particularly useful for identifying persistent struggles and acquisition orders (Díaz-Negrillo et al., 2010). Contrasting natives and learners uncovers constructions that pose difficulties. Compiling localized corpora is essential as acquisition patterns differ across L1 backgrounds. Focused corpus construction thus addresses specific questions in applied linguistics and language pedagogy.

In an overview, specialized corpus design, annotation detail, software functionality and analytical methodology collectively enable English language research across the diverse domains of lexicography, grammar studies, historical linguistics, sociolinguistic analysis and second language acquisition research. These applications highlight the vital utility of corpus linguistics in discovering fresh perspectives on real language use.

Case Study

Leveraging Corpora to Uncover Systematic Metaphor Patterns

Semino (2008) conducted an enlightening study that exemplifies the potential of corpus-based techniques for examining linguistic patterns too subtle or expansive to observe manually. She explored usage trends of conceptual metaphors - figurative constructs that represent intangible ideas through more concrete conceptual domains like journeys, buildings, or forces (Kövecses, 2010; Lakoff & Johnson, 1980). Though metaphors suffuse everyday language, tracking dynamics systematically across diverse texts and user groups poses challenges exceeding human capacity. Capitalizing on corpus affordances, Semino blended computational analyses with qualitative interpretation to uncover complex variations in metaphor distributions.

Semino analyzed the British National Corpus - a 100 million word balanced reference corpus sampling UK speech and writing across genres (Burnard, 1995) along with the separate Metaphor in Discourse project corpus consisting of 167 texts. She first compiled metaphors representing established conceptual mappings like THEORIES ARE BUILDINGS that lend structural form to abstractions. Collocational analysis exposed related terms signaling these source-target pairs, expanding the set of metaphor keywords for broader concordance extraction. The concordances provided thousands of naturally occurring contextual examples to inspect patterns beyond what researchers could manually assemble or intuit through memorized exemplars.

The comprehensive samples enabled tracking quantitative distributional trends. As anticipated, competitive metaphors dominated business texts while political discourse featured extensive WAR metaphors (Krennmayr, 2011). Variation also emerged across user demographics coded in metadata tags - male writers more frequently opted for aggressive WAR or SPORTS metaphors considered counter-normative in female speech (Skorczynska & Deignan, 2006). Lower social grades preferred ontological metaphors framing emotion via visceral experiences like forces or burdens. The automated corpus techniques afforded exponential gains in evidentiary range.

However, computational quantification alone fails to capture situational specifics: emotional tones, irony, extensions of conventional metaphors etc. To interpret such nuances, manual examination of metaphor usages within wider co-text and pragmatic circumstances remained essential, underscoring the continued necessity of qualitative analysis in contextualizing corpus findings (Deignan, 2005). Still, strategically combining methodological approaches realized insights beyond the purview of any single technique, demonstrating the power of mixed-methods corpus linguistics when thoughtfully applied.

Implications and Future Directions

Far-Reaching Implications and Future Horizons for Corpus-Driven English Language Research

The corpus analysis paradigm has ignited exponentially across linguistics domains over recent decades. Access to searchable mega-data repositories affords revolutionary perspectives concerning linguistic structures, variational patterns, and the interplay of language, culture and ideology. As technological capabilities continue advancing, corpora harbor immense untapped potential to propel increasingly nuanced revelations about the human faculties underpinning language use, evolution and diversity.

Transformative Impact Thus Far

Aggregating authentic samples allows tracking large-scale lexical, grammatical, and rhetorical shifts over historical time, unveiling mechanisms of language change like colloquialization, borrowing and neologizing imperceptible via traditional comparative analysis (Biber & Gray, 2016). Reference corpora also capture informal dialects and marginalized sociolects frequently excluded from formal grammars and dictionaries predicated on prestige varieties (Anderwald & Szmrecsanyi, 2019). Quantifying usage trends exposes issues of standardization, linguistic prejudice and marginalization that demand addressing in progressive, egalitarian societies (McEnery & Hardie, 2012).

Additionally, vast datasets combined with efficient computational techniques facilitate nuanced profiling of learner interlanguage acquisition orders, common hindrances, and pedagogical deficiencies. This enables tailored and adaptive language instruction (Granger, 2009). Elevated diversity of corpus contents convers linguistic anthropological analysis of how syntactic and lexical selections project social identity, values and ideology through codified cues (Podesva &

Callier, 2015). Metaphor tracking reveals shifting affective resonances and connotative meanings influencing thoughts and attitudes at a societal scale (Koller, 2008). Such findings carry applied potential to promote empathy, diversity and progress.

Upcoming Possibilities on the Horizon

Looking forward, steady improvements in storage capacities, computational linguistics models, and interactive visualization will galvanize further momentous opportunities. Movement toward multibillion-word corpora for smaller languages and highly domain-specific micro-corpora powered by web scraping technologies will enable intricately detailed profiling at exponentially greater scales (Biber & Reppen, 2015).

Simultaneously integrating textual data with paralinguistic channels - video, audio, eyetracking logs and more - facilitates multiparametric analysis capturing tone, affect, gestures etc. previously absent from written corpora. This multimodal shift unlocks investigating subtler linguistic complexities (Biber, 2020). Steadily advancing semantic analysis and tagging may eventually reliably annotate features like metaphoricity, humor, and emotive intensity beyond surface forms. Translation databases continue ameliorating automatic machine translation capabilities, while optimized language teaching applications can leverage learner corpus findings for maximally adaptive pedagogies.

Creatively interfacing these extensive data resources with interactive data visualization techniques provides intuitive portals for synthesizing insights. Dynamic corpus analysis platforms promote citizen social science, empowering non-specialist audiences to trace cultural lexicons and social discourses impacting communities over generations. Overall, the future landscape promises exponential gains in what corpus-driven methodologies can unveil concerning the collective human condition. Yet meaningfully channeling these emergent technoscapes hinges upon thoughtful inquiry framing, interpretation and intentionality regarding how understanding language can foster progress, empathy and social justice.

Conclusion

This paper has provided an overview of corpus linguistics methodology and its multifaceted applications in contemporary English language research. As large-scale principled text collections encoding authentic usage events, corpora furnish revolutionary affordances compared to reliance on contrived examples or limited observational capacity. Computational analysis empowers investigating lexical, grammatical, discursive and rhetorical patterns at unprecedented scales, enabling significant advances.

We have explored major corpus-based techniques spanning corpus construction, annotation, computational analysis and interactive visualization that collectively uncover dynamics beyond feasible manual examination. When applied judiciously with understanding of inherent data limitations, these methods have unlocked fresh perspectives across diverse linguistic domains. From tracking semantic changes diachronically to profiling interlanguage development trajectories and metaphor systems framing ideologies, creative corpus applications continue illuminating new facets of language, thought and culture.

As natural language processing capacities scale exponentially thanks to machine learning along with exponential data aggregation, the corpus-driven paradigm stands poised to accelerate still further. This bright horizon commands prudent application - prioritizing questions of genuine human import rather than technological exhibitionism; emphasizing ethical data sourcing and interpretation; and sustaining respect for the complexities of linguistic phenomena that transcend even the most sophisticated computational methods. Maintaining this balance of wisdom and ambition promises profound revelations about the human condition through the window of our most astonishing innovation - language.

References

Anderwald, L, & Szmrecsanyi, B. (2019). Corpus-based sociolinguistics. In K. Allan et al. (Eds.), The Routledge Handbook of Linguistics (pp. 270-294). Routledge.

Baker, P., Hardie, A., McEnery, T., (2006). A Glossary of Corpus Linguistics. Edinburgh University Press.

Berber Sardinha, A. P. (2014). 25 Years of the International Journal of Corpus Linguistics: Looking Back, Moving Forward. International Journal of Corpus Linguistics, 19(3), 395–418. https://doi.org/10.1075/ijcl.19.3.01sar

Biber, D. & Reppen, R. (2015). The Cambridge Handbook of English Corpus Linguistics. Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus. TOEFL Monograph Series. https://www.ets.org/s/toefl/pdf/toefl_monograph_ms_25.pdf

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). Longman grammar of spoken and written English. MIT Press.

Brezina, V., Weill-Tessier, P., & McEnery, T. (2015). #LancsBox v.4.0 [software]. Available at http://corpora.lancs.ac.uk/lancsbox

Burnard, L. (1995). Users Reference Guide British National Corpus Version 1.0. British National Corpus Consortium, Oxford University Computing Services.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. International Journal of Corpus Linguistics, 14(2), 159–190. https://doi.org/10.1075/ijcl.14.2.02dav

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. Corpora, 7(2), 121–157.

Deignan, A. (2005). Metaphor and corpus linguistics (Vol. 6). John Benjamins Publishing.

Díaz-Negrillo, A., Ballier, N., & Thompson, P. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. Language Forum, 36(1-2), 139-154.

Ellis, R., & Barkhuizen, G. (2005). Analysing learner language. Oxford University Press.

Francis, W. N., & Kucera, H. (1979). Brown corpus manual. Brown University.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), Corpora and Language Teaching. John Benjamins Publishing.

Hanks, P. (2012). The Corpus Revolution in Lexicography. International Journal of Lexicography, 25(4), 398–436. https://doi.org/10.1093/ijl/ecs026

Hardie, A. (2012). CQPweb - combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics, 17(3), 380–409. https://doi.org/10.1075/ijcl.17.3.07har

Hardie, A. (2013). CQPweb - combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics, 17(3), 380–409. https://doi.org/10.1075/ijcl.17.3.07har

Hunston, S. (2002). Corpora in applied linguistics. Cambridge University Press.

Ishikawa, S. (2013). ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.) Learner Corpus Studies in Asia and the World, 1 (pp. 91-118). Kobe University.

Koller, V. (2008). “Not just a colour”: Pink as a gender and sexuality marker in visual communication. Visual Communication, 7(4), 395-423.

Kövecses, Z. (2010). Metaphor: A practical introduction. Oxford University Press.

Krennmayr, T. (2011). Metaphor in newspapers. LOT.

Kytö, M. (1996). Manual of information to accompany The Helsinki Corpus of English Texts. Department of English, University of Helsinki.

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. University of Chicago Press.

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

Meunier, F., Granger, S., & Gilquin, G. (Eds.). (2013). Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jimenez, R., Knight, D., Křen, M., Llewellyn, C., McKenna, C., Rayson, A., Thompson, P., & Yalamanchili, A. (2016). A prototype semantic tagger for historical corpora. Digital Scholarship in the Humanities, 31(4), 897-908.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jimenez, R., Knight, D., Křen, M., Llewellyn, C., McKenna, C., Rayson, A., Thompson, P., & Yalamanchili, A. (2016). A prototype semantic tagger for historical corpora. Digital Scholarship in the Humanities, 31(4), 897-908.

Podesva, R. J., & Callier, P. (Eds.). (2015). Voice and desire: The seductions of discourse. Washington Press.

Römer, U. (2011). Corpus research applications in second language teaching. Annual Review of Applied Linguistics, 31, 205-225.

Scott, M., & Tribble, C. (2006). Textual patterns: Key words and corpus analysis in language education. John Benjamins Publishing.

Semino, E. (2008). Metaphor in discourse. Cambridge University Press.

Sinclair, J. (2005). Corpus and text - basic principles. In M. Wynne (Ed.), Developing Linguistic Corpora: a Guide to Good Practice (pp. 1-16). Oxbow Books.

Skorczynska, H., & Deignan, A. (2006). Readership and purpose in the choice of economics metaphors. Metaphor and Symbol, 21(2), 87-104.

Teubert, W. (2005). My version of corpus linguistics. International Journal of Corpus Linguistics, 10(1), 1–13. https://doi.org/10.1075/ijcl.10.1.04teu

Xiao, R. (2009). Theory-driven corpus research: Using corpora to inform aspect theory. In K. Aijmer (Ed.), Corpora and Language Teaching (pp. 179-205). John Benjamins Publishing.