Corpus-Based Discourse Analysis Tools: A Comparative Study of AntConc and Praaline

Asrorova Nargiza Isomitdinovna
nibotova@gmail.com
Independent researcher,
Uzbek state world languages university

Annotation *Corpus-based discourse analysis has recently garnered importance as a methodological approach in applied linguistics, especially, in discourse studies. The integration of specialized software tools significantly enhances the systematic investigation of discourse in large corpora. This paper provides a comparative analysis of two widely used corpus analysis tools – AntConc and Praaline (an extension of Praat), and gives a descriptive overview of both programs. The comparison aims to shed light into the affordances of both tools for researchers engaged in CBDA of speech. The comparative study focuses on the functions, applications, strengths, and limitations of AntConc and Praaline in the context of spoken discourse analysis. Through an in-depth investigation, it is argued that while AntConc excels in textual concordance and keyword analysis, Praaline enables more nuanced multimodal analysis, particularly of phonetic and prosodic features. Moreover, the paper explores how these tools support diverse research needs, from pedagogical applications to sociolinguistic and pragmatic investigations, making them essential assets in modern linguistic inquiry.*

Keywords *Corpus, discourse analysis, automated annotation, AntConc, Praaline, speech prosody*

Дискурс-анализ на основе корпуса: Сравнительное исследование программ AntConc и Praaline

Асророва Наргиза Исомитдиновна
nibotova@gmail.com
Независимый исследователь,
Узбекский государственный университет мировых языков

Аннотация *Дискурс-анализ на основе корпуса стал одним из ключевых методологических подходов в прикладной лингвистике, особенно в рамках дискурсивных исследований. Интеграция специализированных программных средств значительно расширяет возможности систематического изучения дискурса на больших корпусах. В данной статье представлен сравнительный анализ двух широко используемых инструментов корпусного анализа – AntConc и Praaline (дополнение программы Praat), а также содержится описательный обзор функционала обеих программ. Сравнительное исследование сосредоточено на функциях, областях применения, сильных сторонах и ограничениях AntConc и Praaline в контексте анализа устной речи. В результате углублённого анализа автор приходит к выводу, что AntConc наиболее эффективен в работе с текстовой конкорданцией и анализом ключевых слов, тогда как связка Praaline предоставляет более гибкие возможности для мультимодального анализа, особенно в части фонетических и просодических характеристик. Кроме того, в статье рассматривается, как эти инструменты*

поддерживают разнообразные исследовательские задачи — от педагогических до социолингвистических и прагматических, что делает их незаменимыми в современной лингвистической науке.

Ключевые слова *Корпус, дискурс-анализ, автоматическая аннотация, AntConc, Praaline, просодия речи*

Korpusga asoslangan diskurs tahlili: AntConc va Praaline dasturlarining qiyosiy tadqiqi

Husenova Asrorova Nargiza Isomitdinovna
nibotova@gmail.com
Mustaqil izlanuvchi,
O'zbekiston davlat jahon tillari universiteti

Annotatsiya *Korpusga asoslangan diskurs tahlili amaliy tilshunoslik, ayniqsa diskurs tadqiqotlarida nihoyatda muhim metodologik yondashuvga aylanib ulgurgan. Maxsus dasturiy vositalarning mazkur jarayonga integratsiyasi esa katta hajmdagi korpuslarda diskursni tizimli tahlil qilish samaradorligini sezilarli darajada oshiradi. Ushbu maqolada korpusga asoslangan diskurs tahlilida keng qo'llaniladigan ikki korpus tahlil dasturi – AntConc va Praaline (Praat dasturining qo'shimchasi) qiyosiy tahlil qilinib, har ikki dasturiy ta'minotga tavsifiy sharh beriladi. Qiyosiy yondashuv orqali og'zaki so'zlashuvlarning korpusga asoslangan diskurs tahlili (KADT) bilan shug'ullanuvchi tadqiqotchilar uchun AntConc va Praaline dasturlarining funksiyalari, qo'llash tavsiya etiladigan sohalari, ustunliklari va cheklovlarini og'zaki diskurs tahlili misolida o'rganib chiqiladi. Tadqiqot natijasi shuni ko'rsatdiki, AntConc matnli konkordans va kalit so'zlar tahlilida ustunlikka ega bo'lib, Praaline esa og'zaki diskursning fonetik va prosodik xususiyatlarini yanada aniqroq multimodal tahlil qilish imkonini beradi. Bundan tashqari, maqolada bu vositalarning pedagogik, sotsiolingvistik va pragmatik tadqiqotlarda qanday qo'llanilishi mumkinligi muhokama qilinadi, bu esa ularni zamonaviy lingvistik tadqiqotlar uchun muhim vosita sifatida belgilaydi.*

Kalit so'zlar *So'zlar: korpus, diskurs tahlili, avtomatlashtirilgan annotatsiya, AntConc, Praaline, nutq prozodiyasi*

Introduction

Corpus-based discourse analysis (CBDA) investigates how language is used in context through the lens of large datasets called corpora. CBDA integrates the principles of discourse analysis with corpus linguistics. Unlike traditional discourse analysis, which often relies on manual and qualitative methods, CBDA incorporates quantitative techniques such as frequency counts, concordances, and collocations to identify patterns of discourse features across large datasets (Baker, 2006; Partington et al., 2013). Corpus, the plural of which is corpora, is defined as "a body of written or transcribed speech which can serve as a basis for linguistic analysis and description" (Kennedy, 1998, p.1). Two major types of corpora are used in CBDA: written corpora (e.g., newspapers, academic texts) and spoken corpora (e.g., interviews, conversations, classroom discourse). The analysis of spoken corpora presents unique challenges as

elements of prosody, pauses, overlaps, and non-verbal cues are present. Researchers dealing with speech data often use diverse tools, data formats and metadata which require collaborative workflows. Therefore, tools compatible with both text-based and audio-based analysis are essential in this field. With the increasing availability of spoken corpora and the need for multimodal analysis, researchers now rely on specialized tools to annotate, analyze, and visualize discourse features. Among these tools, AntConc and Praaline have gained prominence.

Literature Review

AntConc, a freeware corpus analysis toolkit aimed for text-based corpora, was created by Laurence Anthony in 2005. It enables users to examine written or transcribed corpora using data-driven analysis. The application is suitable for N-gram analysis, lexical bundles, collocation analysis, concordancing, and keyword and frequency analysis. Made freely available to the public, AntConc can be downloaded directly from the official website as a single executable file. Creating a full list of every word in the corpus is one of the tasks a researcher can accomplish with AntConc. AntConc's wordlist can be arranged either by frequency or alphabetically, just like other corpus analysis tools. AntConc's extra features like word analysis and reverse ordering are what makes it stand out among other available tools.

The primary function of AntConc is the Concordancer, much like other corpus analysis tools. As stated by Sun and Wang (2003) concordancers aid in tracking the development of a second language by boosting vocabulary acquisition, grammatical accuracy, and writing proficiency. Users can use this tool to look at how particular terms are used across the corpus. Simple words, phrases, or even entire regular expressions (REGEX) can be used as search terms, allowing for more flexible and in-depth queries. By choosing the N-gram option, researchers can also examine terms based on their stems. The Concordance Plot tool helps identify reoccurring patterns or keywords by graphically displaying a term's distribution throughout a text (Anthony, 2005).

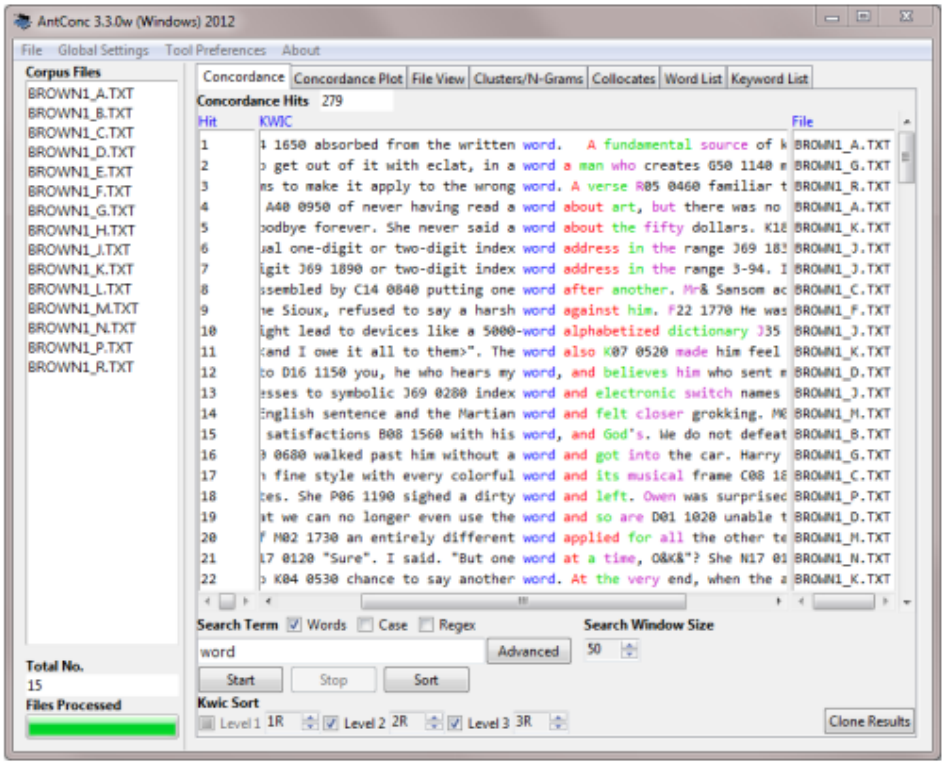

Figure 1. *Concordancing interface of AntConc*

The View Files tool, which can be used in conjunction with or independently of the Concordancer, is another helpful feature. It displays the original context of a search term when used with Concordancer results. When used independently, it provides information about lexical usage in context by displaying instances of a specific substring, word, phrase, or REGEX pattern in the original file (Anthony, 2004, 2005). Additionally, AntConc has Word Clusters or Bundles tool which is used to study multiword expressions like idioms, collocations, and phrasal verbs. This feature finds repeating word patterns that revolve around a specific term and enables them to be arranged either by frequency or alphabetically. Similar to the View Files, Plot, and Concordancer tools, the tool takes search inputs in the form of individual words, phrases, substrings, or. Like the Concordancer, Plot, and View Files tools, it accepts search inputs defined as substrings, individual words, phrases, or regular expressions. AntConc is particularly effective in examining lexical and grammatical features, making it a strong tool for linguistic profiling, discourse markers, stance-taking expressions, and keyword-based thematic analysis in transcribed data.

The second tool examined by this study is Praaline, which is an open-source platform designed for the management, annotation, analysis, and visualization of spoken language corpora based on Praat ((Boersma & Weenink, 2014). Praaline is compatible with established tools such as Praat, Sonic Visualiser, and provides connectivity to the R statistical environment – all within a modular and cohesive framework that supports automation and reusability. The software offers a unified, intuitive interface, allowing users to seamlessly access various tools. Both corpus metadata and annotation layers can be stored in a local or remote database, with full flexibility for users to define their own metadata and annotation schemas. It also enables users to execute customizable sequences of analytical processes through scripts and plug-ins, with automatic updating of the database based on the analysis outcomes. Furthermore, the corpus database can be queried to generate summarized datasets for further analysis. Praaline supports extensibility through plug-ins written in Python or C++, and also allows users to run Praat and R scripts directly on corpus data. It comes equipped with a variety of visualization tools, editing interfaces, and additional plug-ins. Distributed under the GPL license, Praaline is freely available to the research community.

One of the featured aspects of Praaline is its ability to import and export annotations in variety of formats ((Barras et al., 1998) and it also offers a prominent advantage over other tools by "defining structural links between the annotation tiers, such as hierarchy, containment, attachment, controlled vocabularies" (Christodoulides, 2018).

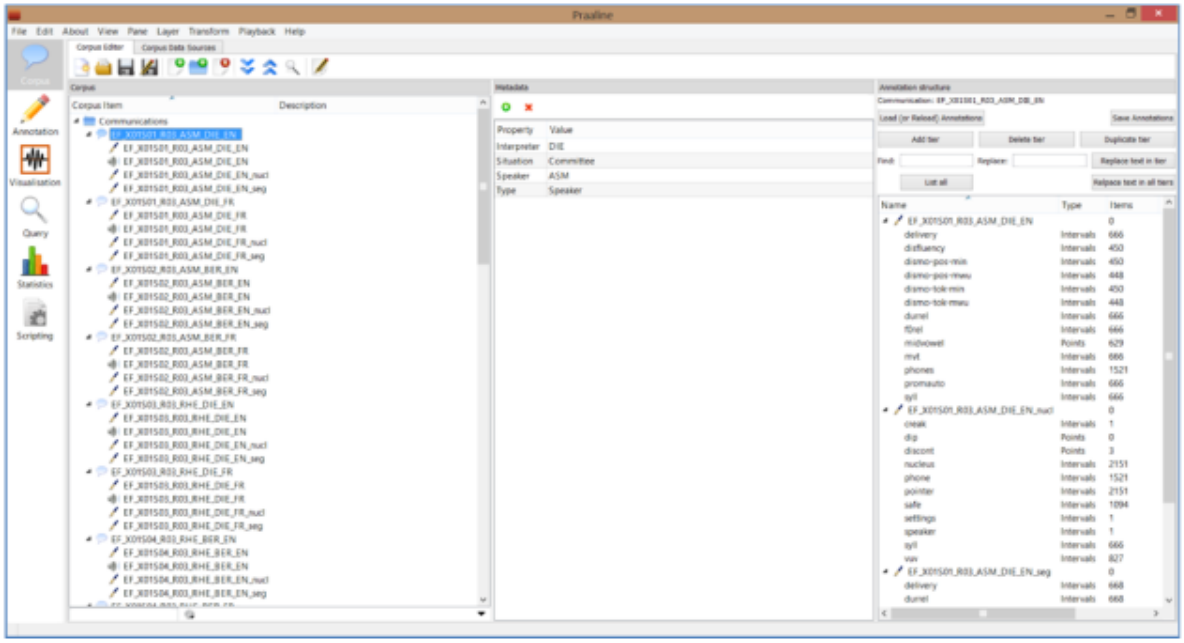

Figure 2. *Corpus management interface of Praaline*

In terms of linguistic prosody and phonetic analysis, Praaline far outreaches other programs. Using the program, comparing prosodic elements in spoken corpora and their simultaneous interpretation can be carried out. The results then can be then visualized into various graph types. Praaline's visualization component is built upon the Sonic Visualiser framework (Cannam et al., 2010) and supports the display of various types of visual data, including waveforms, spectrograms, pitch-related spectrograms, multiple annotation levels and tiers, as well as numerical representations such as data points, curves, histograms, and color-coded regions. Additional visualizations can be integrated through plug-ins. For instance, the Prosogram tool (Mertens, 2004) has been adapted within Praaline to present prosodic analysis results.

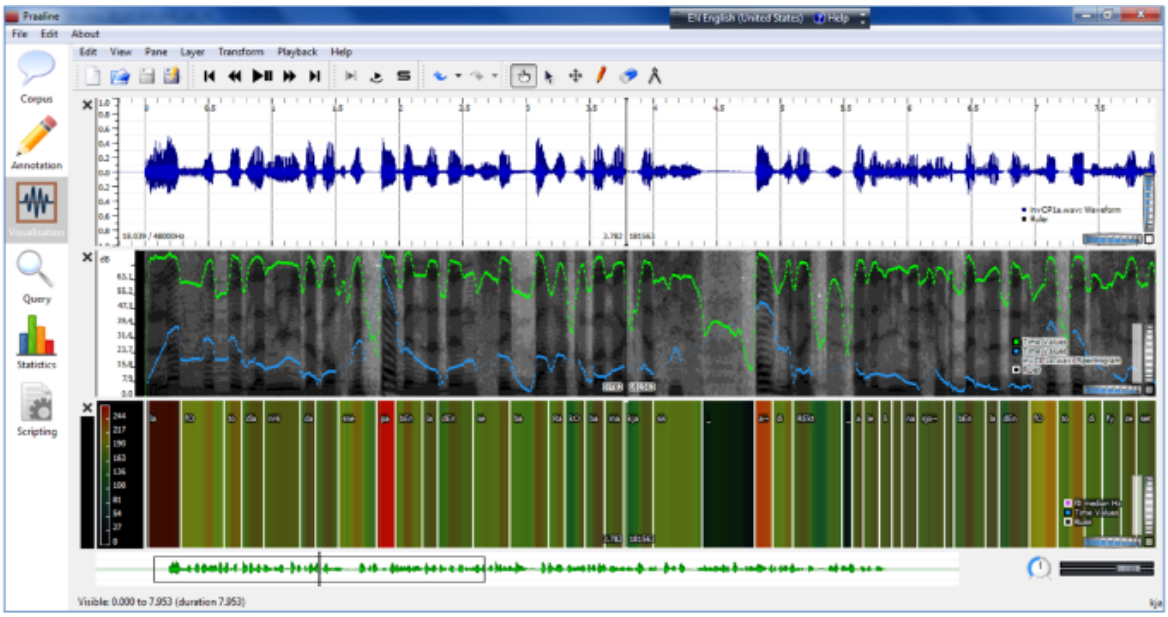

Figure 3. *Color-coded visualization of syllables in Praaline*

Research Methodology

This study employs a qualitative comparative design to evaluate two widely used corpus-based discourse analysis (CBDA) tools: AntConc and Praaline. The analysis focuses on their core functionalities, usability, and effectiveness in processing spoken discourse data. A representative spoken corpus of 15.000 words, complied from L2 users' interactions with AI chatbots, were transcribed into text format, was compiled and systematically analyzed using both tools. AntConc was examined for its text-based capabilities, including concordancing, keyword extraction, collocation analysis, and N-gram generation. In contrast, Praaline was evaluated for its integration with Praat and R, multimodal annotation features, and visualization capabilities for phonetic and prosodic elements. Each tool was assessed based on established criteria derived from prior research (e.g., Kennedy, 1998; Anthony, 2005; Partington et al., 2013), with attention to features such as user interface, extensibility through plug-ins or scripts, support for annotation layers, and overall workflow efficiency. The evaluation included hands-on use of both tools, with outputs compared in terms of analytical depth, flexibility, and relevance to spoken discourse research. The findings were thematically organized to highlight comparative strengths and limitations. This methodology allows for an informed assessment of how each tool supports CBDA practices and meets the needs of researchers working with complex, multimodal speech corpora.

Results and Discussion

When comparing AntConc and Praaline, it becomes clear that each tool has its unique features. AntConc is a user-friendly and accessible tool which is particularly effective when working with written or transcribed text. It allows researchers to explore word frequencies, collocations, and concordances, making it ideal for identifying patterns in vocabulary and structure across large corpora. Its simplicity and speed make it a great starting point for linguistic analysis, especially for those new to corpus tools. However, AntConc is limited to textual data – it doesn't support audio or time-aligned analysis – so it is less suitable for studying more dynamic features of spoken language like intonation, pauses or overlapping speech.

Praaline, by contrast, is designed with spoken discourse in mind. It builds on the acoustic strengths of Praat but adds powerful annotation and analysis features that allow users to work with both transcripts and audio together. With Praaline, researchers can investigate how speech sounds unfold over time, track features like pitch and rhythm, and analyze how speakers take turns or emphasize certain words. While it takes more time to learn and set up, it offers a flexible, customizable environment for working with complex, multimodal data. For researchers studying the finer details of spoken interaction – like timing, intonation, or speaker alignment – is an invaluable tool. Together, AntConc and Praaline show how combining text-based and audio-based approaches can lead to a deeper, more complete understanding of discourse.

Feature	AntConc	Praaline
Data Type	Text-based	Audio + Text (spoken data)
Main Functionality	Concordancing, keyword, collocation	Prosodic, phonetic, and timing
Multimodality Support	No	Yes
Ease of Use	High	Medium to Low
Ideal for	Thematic, lexical, frequency analysis	Turn-taking, prosody, intonation
Annotation Capabilities	Limited	Extensive (via TextGrids)
Visualization Tools	Concordance plots	Spectrograms, pitch tracks

Table 1. *Comparative evaluation of AntConc and Praaline features*

The table illustrates that AntConc and Praaline are two powerful tools that serve different purposes in corpus-based discourse analysis. AntConc is best suited for analyzing written or transcribed texts, offering user-friendly features like word frequency, collocation, and concordance analysis. Its simplicity makes it especially useful for beginners, though it lacks support for audio or time-aligned data. In contrast, Praaline is tailored for spoken discourse, allowing detailed multimodal analysis through synchronized audio and transcript annotations. Used together, these tools offer a complementary approach that enables a more comprehensive understanding of both the textual and prosodic features of discourse.

Conclusion

This comparative analysis highlights the distinct yet complementary roles of AntConc and Praaline in corpus-based discourse analysis. The results show that AntConc excels in exploring lexical and structural features within transcribed spoken data. Its user-friendly design and efficient processing make it particularly useful for identifying frequency patterns, keywords, and discourse markers. However, its limitations become evident when dealing with prosodic and interactional features, as it does not support audio integration or time-aligned analysis. Praaline addresses these gaps by enabling detailed multimodal analysis of spoken interactions. With its ability to align transcripts to audio and annotate features like pitch, intonation, and turn-taking, it provides researchers with deeper insights into how spoken language functions in real-time. While it demands more technical expertise, Praaline makes rich discourse-level phonetic and prosodic analysis possible. Incorporating these tools, a more holistic approach to studying spoken discourse can be formed by a thorough analysis of both corpus and discourse features of spoken interactions.

References:

1. Anthony, L. (2005). AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. *Proceedings of the IEEE International Professional Communication Conference*, 729–737.
2. Baker, P. (2006). Using corpora in discourse analysis. *Continuum*.
3. Boersma, P., & Weenink, D. (1992–2024). Praat: Doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org/

4. Barras, C.; Geoffrois, E.; Wu, Z. and Liberman, M. (1998). Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. *In Proceedings of LREC 1998*, pp. 1373-1376.
5. Cannam, C.; Landone, C.; Sandler M. (2010). Sonic Visualiser: An open-source application for viewing, analyzing, and annotating music audio files, *Proceedings of the ACM Multimedia 2010 International Conference*, pp. 1467-1468.
6. Christodoulides, G. (2018): Praaline: Integrating Tools for Speech Corpus Research, pp. 2-6
7. Kennedy, G. (1998). An introduction to corpus linguistics. London: Longman.
8. Mertens, P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In B. Bel & I. Marlien (Eds.) *Proceedings of Speech Prosody 2004*, Nara, Japan.
9. Partington, A., Duguid, A., & Taylor, C. (2013). Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS). Amsterdam: John Benjamins Publishing.