

CORPUS PROCESSING TOOLS IN BUILDING MULTILINGUAL CORPORA

Amirqulov Ma'rufjon Aliqul o'g'li

amirkulov.edu01@gmail.com

ToshDO'TAU magistranti

Abstract. Corpora annotation constitutes a pivotal phase in the development of multilingual corpora, encompassing various processes like lemmatization, tagging, parsing, and more. These procedures involve the addition of meta-information into the corpus, enabling the organization of textual elements such as words into specific categories like part-of-speech, thereby facilitating comprehensive analysis of their relationships and structures. Without such meticulous sorting and understanding of interrelations, the corpus data cannot be fully leveraged to its potential. Thus, this article discusses aforementioned matters briefly.

Annotatsiya. Korpus annotatsiyasi lemmatizatsiya, teglash, tahlil qilish va boshqalar kabi turli jarayonlarni o'z ichiga olgan ko'p tilli korpusni rivojlantirishning asosiy bosqichini tashkil qiladi. Ushbu protseduralar korpusga metama'lumotni qo'shishni o'z ichiga oladi, bu so'zlar kabi matn elementlarini nutqning bir qismi kabi muayyan toifalarga ajratish imkonini beradi va shu bilan ularning munosabatlari va tuzilmalarini har tomonlama tahlil qilishni osonlashtiradi. Bunday sinchkovlik bilan tartiblash va o'zaro munosabatlarni to'liq tushunib yetmaslik korpus ma'lumotlarining to'liq imkoniyatlaridan foydalanishga to'siq bo'ladi. Ushbu maqola yuqorida ta'kidlab o'tilgan masalalarni qisqacha muhokama qiladi.

Аннотация. Аннотации Корпуса представляют собой ключевой этап в разработке многоязычных корпусов, охватывающий различные процессы, такие как лемматизация, тегирование, синтаксический анализ и многое другое. Эти процедуры включают добавление метаинформации в корпус, что позволяет организовать текстовые элементы, такие как слова, в определенные категории, такие как части речи, тем самым облегчая всесторонний анализ их отношений и структур. Без такой тщательной сортировки и понимания взаимосвязей корпусные данные не могут быть полностью использованы в полной мере. Таким образом, в данной статье кратко обсуждаются вышеупомянутые вопросы.

Keywords: *corpora, concordance, aligner, metadata, lemmatizator.*

While manual annotation of metadata is theoretically accurate, it is exceedingly laborious and time-intensive. Hence, there is a prevalent preference for employing software tools to automate these tasks. Despite the potential for lower accuracy levels compared to manual methods, the practicality and efficiency offered

by automated annotation software make it the predominant choice in contemporary corpus development practices. This shift is primarily driven by the sheer volume of textual information present in corpora, especially those of considerable size, rendering manual processing unfeasible. Therefore, specialized software tools play a critical role in expediting the annotation process, thereby saving both time and resources.

In scenarios where mid-sized corpora or high precision outputs are required, semi-automatic annotation tools serve as a viable alternative. These tools automate a portion of the annotation process, which is then refined and enhanced through manual editing and further software training. Conversely, for smaller and specialized corpora with limited sample sizes, manual metadata addition proves advantageous, allowing for meticulous attention to detail and precision.

The thesis at hand exclusively focuses on automatic annotation tools, showcasing a variety of software options available for corpus processing. Examples include Wordfisher, Corpus Builder, BNC, Gestrolex, ANTConc, Sketch Engine, and Shoebox, each offering unique functionalities tailored to specific research needs. While some tools provide specialized features, others are comprehensive and intricate, catering to diverse requirements within the field of corpus linguistics.

Historically, early iterations of corpus processing software were notorious for their lack of user-friendliness and steep learning curves, often necessitating substantial computer literacy. However, modern software trends prioritize accessibility and ease of use, making these tools increasingly approachable for both linguists and non-experts alike. As a result, the current landscape emphasizes user-friendly interfaces and intuitive design, ensuring broader accessibility and usability across the corpus linguistics community.

The subsequent section of the thesis delves into fundamental tools designed to enhance the overall reliability of corpus data. These tools are geared towards differentiating between grammatical or syntactic categories, resolving language ambiguities, and refining structural analysis, rather than directly searching the corpus or providing finite data outputs.

LEMMAZIZER

A lemmatizer plays a crucial role in linguistic analysis by assigning a lemma—a collection of lexical forms with a shared stem and belonging to the same major word class—to each word entry in a corpus [Francis & Kučera, 1982:1]. This process simplifies the corpus by reducing words to their respective lexemes, enabling researchers to explore all variations of a specific lexeme without manually inputting every potential variant. Additionally, it facilitates the extraction of frequency and distribution data for each lexeme, as highlighted by Baker, Hardie, and McEnery [Baker, Hardie & McEnery, 2006:104-105].

Given the unique characteristics of each language, lemmatization tools must be developed separately for each linguistic system. However, linguists may

sometimes prioritize broader language features, such as grammatical categories and syntactic relationships, over specific lexical compositions. In such cases, it becomes essential to distinguish these features within the corpus texts. This is achieved through the use of linguistic markers known as tags. These tags help delineate the grammatical and syntactic structures present in the texts, allowing researchers to analyze language features beyond individual word forms.

TAGGER

Tagging, as described by Baker, Hardie, and McEnery [Baker, Hardie & McEnery, 2006:154] involves applying additional layers of annotation to corpus data, specifically adding codes to each word to indicate its part of speech [Hounston, 2002]. Initially, during the early stages of corpus development, tagging was a manual process. However, advancements in technology have led to the creation of software tools for automatic tagging.

Despite these advancements, fully automated tagging remains elusive. Instead, a manual tagging process is typically employed initially, using a sufficiently large text as a "tutorial" for the tagging software. Through a process of learning based on verisimilitude or predetermined linguistic rules, the tagger then assigns grammatical category markers (tags) to selected texts. According to Baker, Hardie, and McEnery [2006:154], a tag can be attached to various linguistic units, including phonemes, morphemes, words, phrases, or longer stretches of text.

Although human correction and post-editing are still necessary, the accuracy of automated tagging has surpassed 90%. This high level of accuracy allows tagging software tools to be utilized in much larger projects, streamlining the tagging process and enhancing efficiency in corpus annotation endeavors.

PARSER

Parsing is another essential procedure in corpus annotation, as it involves marking texts not only grammatically but also syntactically. Unlike tagging, which focuses on grammatical categories, parsing delves deeper into the syntactic structure of texts.

To enable parsing, a small number of manually processed texts are typically required to instruct parsers on recognizing and marking syntactical categories within a selected text in a corpus. Parsing tags may include delineating the start and end points of units such as noun phrases, verb phrases, and clauses. Additionally, parsing may provide information about how these syntactic units relate to one another [Baker, Hardie & McEnery, 2006:127]. However, parsing is less commonly used compared to tagging, primarily due to its complexity. Parsing a corpus is far more intricate and resource-intensive than tagging it. Moreover, there are limited parsing tools available to individual researchers or students, further contributing to its lower adoption rate in corpus annotation endeavors.

ALIGNER

In parallel corpus projects, a vital tool is employed to facilitate the alignment of text segments in one language with their counterparts in another language. This tool serves as a cornerstone for linguists, enabling them to synchronize specific parts of a text across different languages seamlessly. Often, this alignment process relies on the principle of verisimilitude, ensuring accurate correspondence between the texts. Once texts undergo modification through this software, researchers can proceed to delve into the corpus for their analyses. However, unlike the annotation tools previously discussed, the software tools described below offer tangible data outputs. These outputs provide researchers with concrete information that can be utilized in various ways. They can be treated as "finite," allowing for the formulation of generalizations and insights based on the extracted data. Alternatively, researchers may subject the data to further scrutiny and analysis, exploring nuances and patterns hidden within the corpus. Thus, these software tools serve as invaluable assets in the extraction and interpretation of linguistic data from parallel corpora.

WORD LISTER

One of the fundamental tools utilized in corpus processing is the word lister, which serves as a cornerstone for basic statistical analyses within a corpus. Often referred to simply as a word lister, this tool enables researchers to perform straightforward calculations and analyses on textual data. As described by Lynne Bowker, a word lister can calculate the total number of "tokens" and tally the frequency of each individual word form within the corpus [Bowker & Pearson, 2002:13]. These word lists can be sorted in various ways, such as alphabetically or by frequency, facilitating easy access to information. Furthermore, certain word listers possess the capability to generate lists of keywords, offering additional utility beyond basic word frequency analysis. These keyword lists can serve as valuable resources for activities like terminology mining and further web searches. For example, word lists derived from specialized corpora can be compared with keyword lists from other corpora, allowing users to discern differences and similarities in lexical usage across different fields of study or domains. This comparative analysis provides valuable insights into the language specific to various contexts, enriching linguistic research and facilitating cross-disciplinary investigations.

CONCORDANCER

A concordancer is a fundamental tool shared by all corpus-processing software, providing users with invaluable insights into the occurrences of specific search terms within a corpus. As defined by Baker, Hardie, and McEnery [Baker, Hardie & McEnery, 2006:42-43], a concordance is essentially a comprehensive list of all instances where a particular search term appears in the corpus, presented alongside the surrounding context in which each occurrence occurs. According to Lynne Bowker and Jennifer Pearson, a concordancer offers users the ability to

observe every instance of a particular word within its immediate contexts. This information is typically presented using a format known as keyword in context (KWIC). In a KWIC display, all occurrences of the search term are aligned at the center of the screen, with a specified amount of context displayed on either side. Similar to word lists, concordances can be sorted to facilitate the identification of patterns, enabling users to glean deeper insights into the usage and distribution of specific terms within the corpus.

CONCLUSION

The realm of corpus processing tools is vast, with a plethora of options available to researchers. In fact, the sheer variety of tools is such that an entire thesis could be dedicated solely to their description. Many software packages encompass a range of tools, some of which may be exclusive to a particular software (for instance, Word Sketches are unique to the Sketch Engine). These tools offer users a diverse array of functionalities, including the ability to access original files, search for collocations, save search results, visualize the clustering of search terms (such as in the AntConc Concordance Plot), utilize the corpus as a thesaurus, and consult additional corpora (as seen in the Sketch Engine). Despite the extensive range of tools available, certain basic tools are indispensable for anyone involved in corpus compilation. These tools, discussed earlier, form the foundation of corpus processing and are typically included in most comprehensive software packages. They serve as essential components of the corpus compilation process, facilitating various analyses and inquiries into linguistic data.

References:

1. Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press. (Baker, Hardie & McEnery, 2006)
2. Bowker, L., & Pearson, L. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. New York: Routledge. (Bowker & Pearson, 2002)
3. Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press. (Hunston, 2002)
- Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin. (Francis & Kučera, 1982)