OPEN ACCESS
Scholars International Journal of Linguistics and Literature
Abbreviated Key Title: Sch Int J Linguist Lit
ISSN 2616-8677 (Print) |ISSN 2617-3468 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Review Article

UAM Corpus Tool: A Convenient and Comprehensive Software

Leyang Wang[1], Di Yu[2*]

[1]Associate Professor, Department of Foreign Studies, North China Electric Power University, Baoding, Hebei, China
[2]Department of Foreign Studies, North China Electric Power University, Baoding, Hebei, China

DOI: 10.36348/sijll.2022.v05i10.006 | **Received:** 17.09.2022 | **Accepted:** 22.10.2022 | **Published:** 25.10.2022

***Corresponding author:** Di Yu
Department of Foreign Studies, North China Electric Power University, Baoding, Hebei, China

Abstract

With the development of corpus linguistics, more and more corpus tools have come into being. As a relatively powerful annotation tool, UAM corpus tool is worth linguistic scholars' attention and promoting. This paper aims to review the application of UAM in linguistic field, and tries to summarize the advantages by comparing it with other corpus tools, like AntConc and SPSS. It shows that UAM is more effective and practical with comprehensive functions. Inevitably, due to the combination of both annotation and corpus statistics, users may need extra time to get familiar with the operation procedure and to know what functions a corpus provides before doing research. In a word, UAM corpus tool is a suitable software for corpus annotation and exploration.

Keywords: Corpus linguistics; UAM corpus tool; statistics; annotation.

1. INTRODUCTION

In recent years, there has been a growing use of software to assist researchers in the annotation of text corpora. Part of this growth has been due to the increasing number of linguists interested in exploring linguistic patterns in text which cannot be explored with simple concordancers. Linguistic features which cannot yet be automatically tagged, such as semantic and pragmatic features, need to be identified by a human, and good annotation software can facilitate this task.

Additionally, there is a growing interest in statistical-based language processing, for example, machine translation, parsing, etc. These systems typically require a training set, which is usually provided by human annotators. Human-annotated texts can also be used as 'golden standards' to facilitate the evaluation of such systems.

The UAM Corpus Tool has been developed to address these needs. In fact, some linguistic researchers are still trapped in the ocean of corpus and have to annotate corpus by themselves, or doubt about the feasibility and reliability of UAM. Though one may once read articles that use UAM as a corpus tool, few articles can be found to review its strength or practicability exclusively in previous researches. Thus, a review of application and advantages of UAM Corpus Tool is needed to help researchers get a critical understanding of the functions and usage of this tool.

This study mainly introduces and evaluates how UAM Corpus Tool performs its role and analyzes its strength in the field of linguistics.

2. Introduction to UAM Corpus Tool

UAM (Universidad Autonoma de Madrid) Corpus Tool, is a software for the annotation of text corpora and allows manual and automatic annotation of collections of text at a number of linguistic layers. It is developed by the computational linguist Mick O`Donnell, who has found that, though lots of annotation tools have been developed, they are not readily adaptable to different annotation problems. To a certain extent, they have been limited in that they allow only certain types of annotation to take place (O`Donnell, 2008). However, the UAM Corpus Tool can fill this gap. It is designed from the ground up to support typical user workflow, and everything the user needs to perform annotation tasks is included within the software.

3. Review of UAM Corpus Tool

3.1 The Convenience of UAM Corpus Tool

The UAM Corpus Tool is a text annotation tool primarily aimed at the linguist or computational linguist who does not program, and would rather spend their time annotating text than learning how to use the

system. Thus, it is convenient for the novice users become competent.

3.1.1 Files and Layers

Once UAM Corpus Tool is installed on computer, users can begin working with it. The first thing to do is to create a new "project", they just click on the "Start New Project" button and provide a name for this project. Then, they can add files needed to annotated, which in the form of either a single or folder text files, even paste from the clipboard is allowed. It also allows users to delete source files from the project, or to open a specific file for annotation at a specific layer.

As to the layers, it allows users to add new annotation layers to the project, and edit or extend the annotation scheme for each layer. To add a layer means to specify what analyses the users require in the project. Automatic and manual annotation are both provided, users can choose either of the two according to their needs, with grammatical structure and part of speech available in the automatic part. In the manual part, there are built-in schemes which include clause grammar, appraisal, rhetorical structure theory and error analysis, or to be more specific, they can design a scheme according to their research purposes, even they can select a scheme file from another UAM project they have established before. What's more, most of the current text annotation tools lack built-in facilities for creating and editing the coding scheme. UAM Corpus Tool uses a hierarchically organized tag scheme, allowing cross-classification and multiple inheritance. The scheme is edited graphically, adding, renaming, moving or deleting features, adding new sub-distinctions, etc.

3.1.2 Corpus Statistics

The Corpus Statistics pane allows various statistics to be derived from users' tagged corpus. It can perform two kinds of studies on the corpus, which includes general text statistics and feature usage. The former offers general statistics of the corpus, such as total number of segments, number of words per segment, lexical density in the corpus, pronominal usage, etc. While the latter one can specify a feature in a layer, most typically, the root feature of the layer, and the program describes the features usage in the corpus at that layer, such as counts, mean, and standard deviation. These studies can be done for a single dataset (descriptive statistics), two datasets (comparative statistics), or showing results for each document individually.

A contrastive feature study is often analyzed through UAM, which has some functions similar to SPSS. When a comparative study is done, it is possible to measure whether the differences between the two datasets are statistically significant. UAM Corpus Tool uses two measures of statistical significance, and presents them both in the results, they are t-statistic and chi squared. T-Stats are the numbers on which the level of significance of result can be derived. The bigger it is, the higher the level of significance, but this also depends on how much data the users have. In some academic papers, people might be requested to provide T-Stats, but it is quite rare in linguistics. Chi Squared, in recent years, particularly in linguistics, Chi Squared statistics are becoming the preferred means of testing significance. Corpus Tool provides the Chi Squared statistics for each comparison, and the level of significance that corresponds to this.

From the above, we can see that UAM Corpus Tool is a user-friendly and convenient annotation tool, offering easy installation, an intuitive interface, yet powerful facilities for management of multiple documents annotated at multiple levels, which is suitable for both novice user and scholars.

3.2 The Comprehensiveness of UAM Corpus Tool

Since UAM is usually employed as a corpus tool in empirical studies, it usually appears in the abstract of an article. When the author searched UAM Corpus Tool as keywords of abstract on the Internet, it found that UAM has been involved in many fields, like linguistics, translation, medicine and so on. The reason why using UAM Corpus Tool as the keyword instead of UAM is that there are lots of abbreviations named UAM in multiple subjects (e.g., medicine science, engineering and management), UAM Corpus Tool is more concrete and effective without misunderstanding. In fact, through searching, the review of the application of UAM Corpus Tool is relative rare, one reason is that few researches are devoted to studies of UAM, or people have not yet realized its importance and irreplaceability. Thus, the following tends to explore the application of UAM in several fields so far and comments on its main advantages.

Among the searching results, most of them are about systemic functional grammar, which shows a view of language in both structure (grammar) and words (lexis), and its latest development, appraisal theory, which are used to explore the discourses of news reports, diplomatic speeches, translation and so on.

Some studies deal with the systemic functional linguistics (SFL). Firstly, in the level of lexis, Zhang, Tan and Ling (2015) use the retrieval function of UAM to take an analysis of various adverbs in excerpts of father-and-son, mother-and-son and sibling relationships and combine this tool with thematic structure theory of Functional Linguistics, from the perspective of narrative discourse, to see how the application of adverbs and shifting of themes contribute to demonstrating the relationships between characters. Also, aided by the UAM corpus tool, Liu and Wang (2016) try to find the identity words to investigate the

means and order of logos in identity construction in Chinese and English doctoral dissertation acknowledgements. In the level of move, the corpus established by Akbas (2021) is analyzed in terms of the promotional and rhetorical moves based on a model developed by the researchers using the UAM Corpus Tool.

Besides, several annotation schemes in UAM are served to analyze the transitivity, a grammar structure, which is popular in SFL. Based on the transitivity system in SFL, Feng and Hong (2022) set out to analyze the headlines of the China-related news reports in The New York Times during the COVID19 pandemic. And another research (Ren, 2014), with the assistance of UAM Corpus Tool, investigates the distribution, collocation, progression and changes of factors in the system of Transitivity of *The Art of War* and its translation, in order to explore the norms in the translation of military strategies and tactics.

As a new outcome of SFL, the Appraisal Theory attracts much attention. Appraisal theory consists three systems of attitude, engagement and graduation. The subsystems of attitude involves the source of attitude, while graduation deals with the way in which attitude can be amplified and hedged, and engagement introduces a range of voices into a text. UAM provides a possibility for researchers to have a more careful and work-saving experience in the annotating process. What's more, UAM is usually applied in master's dissertations to analyze a relatively more abundant language corpus.

UAM has a built-in scheme for AT, which enables users to annotate attitude, engagement and graduation respectively, or use these schemes all in one network. Dai (2020) tends to find the distribution frequency of appraisal resources in the dialogue between Liu Xin and Trish Regan according to the framework of Appraisal Theory by annotating the attitudinal resources, engagement resources, gradable resources and their sub-categories with the help of UAM Corpus Tool 5. Besides analyzing the whole scheme of AT, thesis about engagement only is also involved. In order to analyze the distribution and rhetorical strategies of move resources in the discussion section of thesis, Deng and Zhang (2021) comprehensively examines 20 academic papers of applied linguistics using the UAM corpus tool. Through searching, it is clear that researches of graduation are fewer than the other two subsystems.

UAM Corpus Tool is also widely used in the field of translation. Bartley (2022) also, through a transitivity analysis of three issues from Dabiq (an online magazine), explores how the in-group (the believers) and the other (the non-believers) are represented in the magazine. Besides, in the field of translation, one of the papers aims to discuss the transitivity choices, comparing the short story "Love", published by Clarice Lispector in the book *Family Ties* in Brazilian Portuguese, and its translation into English. The analysis was based on the use of a "quantitative/qualitative software UAM Corpus Tool, which allows the mapping of linguistic systems in a functional perspective (O'Donnell, 2016) and an online alignment tool for corpora. The studies of Xu and Nesi (2019) were coded with the corpus software UAM Corpus Tool 3.0, and the typical features of and the differences between the domestic and overseas research were analyzed with respect to the identified four main areas, namely pedagogical practices research, training evaluation and proposed strategies, process/product research, and LSP-based translation technology surveys.

Even the field of medicine has put UAM into application. One corpus-based study (Williams, 2019) examines women's framing of health issues in online forums by collecting posts from December 2016 to April 2017 and annotating them with the assistance of UAM Corpus Tool to examine emergent categories and compare them to three time periods: pre-, during, and post-ACA. Data within posts were coded as to the linguistic moves being made.

All the investigations mentioned above use the UAM as the analyzing tool, and it is found that though at the beginning, UAM was designed for linguists, more and more scholars have realized the advantages of it and nowadays, it was applied in a relative comprehensive fields, due to its convenient and practical functions, which may not have in other tools.

3.3 Superiority to Other Tools

Compared with other tools, UAM combined annotation, statistic analyzing and contrastive study together to offer a more comprehensive and convenient service. SPSS is a tool that mainly used for statistical analysis of the data. It is widely used in various areas like health care, marketing, educational research, survey companies, education researchers and many others. It provides data analysis for descriptive statistics, numeral outcome predictions, and identifying groups. This software also gives data transformation, graphing and direct marketing features to manage data smoothly. Thus, SPSS is suitable for scholars to focus on the entire analytics process, while it fails to deal with such linguistic issues as semantic division, grammar or structural analyzing and so on.

On the contrast, UAM can not only analyze comparative statistics across subsets, e.g., contrasting conversational patterns used by male and female speakers, but also can do annotation of multiple texts by using the same annotation schemes of your design, or of each text at multiple levels e.g., NP, Clause, Sentence, whole document. Some researchers have not grasped the main function of UAM and they are apt to use tools

which have overlapping properties. For example, one master's dissertation's qualitative and quantitative methods are utilized to analyze and count the frequency of Themes and TP patterns, with the help of WPS Office, SPSS 26.0 and UAM Corpus Tool (Zhuang, 2022).

AntConc, another tool for working with language corpora, is a free software programme using a graphical user interface. Within AntConc are a number of 'tools' that support linguistic analysis by enabling the user to -- for example, search corpora, to generate lists of words in corpora, and to browse 'concordances' of word use in corpora. Nevertheless, one of the weakest areas of AntConc is in its handling of annotated data such as data encoded in HTML/XML format. Although AntConc offers a simple way to view or hide embedded tags used in HTML/XML and other annotation methods, much more sophisticated methods need to be implemented if the full power of annotated data is to be realized (Anthony, 2004).

However, UAM fills this gap. All annotations within it are stored in XML files, meaning that your annotations can more easily be shared with other applications. Meanwhile it uses "stand-off" XML, which means the annotation files do not contain the text, just point to the text. This allows for multiple overlapping analyses of the same text, not so easy in standard XML.

4. CONCLUSION

After a short review of the usage and application of UAM Corpus Tool, and the comparison with SPSS and AntConc, one can find that UAM is a practical annotation tool with function of statistic analyzing. This tool is aimed particularly at those wishing to perform linguistic studies, and thus provides on-board search facilities and statistical reporting. It provides functionalities for coding several documents at multiple annotation layers as well as an auto-coding function and evaluation and visualization functions. UAM focuses on manual as well as semi-automatic annotation. Since it has a comprehensive functions, users may need extra time to get familiar with the operation procedure and to know what functions a corpus provides before doing research, but the interface is easy to learn and easy to use. Different tools target different user needs, and no tools can address all needs. It nowadays is widely used in Systemic Functional Linguistics and its new development: Appraisal Theory. Also, with its convenient and comprehensive functions, researchers in various fields such as translation, journalism and medicine have realized its advantages and put it into application according to their research issues. In the longer term, more necessary features are planned to be achieved, such as supporting storage of project files online, allowing multiple users to work on project files at the same time, controlling access to avoid two people working on the same file at the same time. The software is undergoing continual development and will be improved based on user response.

ACKNOWLEDGEMENTS

This research was supported by the Fundamental Research Funds for the Central Universities (2019MS139).

REFERENCE

- Akbaş, E., & Farnia, M. (2021). Exploring rhetorical moves in a digital academic genre: A cross-disciplinary study of the highlights section. Ibérica, (42), 85-114.
- Ammara, U., & Anjum, R. Y. (2019). The Transitivity Analysis of Woolf's 'Kew Gardens': A Corpus Based Study. Corporum: Journal of Corpus Linguistics, 2(02), 16-37.
- Anthony, L. (2004). AntConc : A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. Proceedings of IWLEL, 2(12), 7-13.
- Bartley, L. (2022). "They fabricated lies against us and described us in the harshest of ways": An analysis of the transitivity patterns used in the online magazine DABIQ. Pragmatics and Society, 13(3), 431-452.
- Dai, L. F. (2020). A Comparative Analysis of Trish Regan and Liu Xin's Dialogue on Sino-US Trade Disputes from the Perspective of Appraisal Theory. Xi'an International Studies University.
- Deng, X. H., & Zhang, D. Q. (2021). The Rhetorical Research on the Discussion Section of Academic Dissertation of Applied Linguistics from the Perspective of Engagement. Foreign Language and Translation, 28(02), 53-60.
- Feng, E. H., & Hong, G. (2022). A Research on the Image of China in Pandemic Related News Headlines: A Transitivity Analysis. Technology Enhanced Foreign Languages. 01, 48-55+108.
- Lima-Lopes, R. E. D. (2018). Translation choices as representational systems: a study of the processes in the short story "Love" by Clarice Lispector. DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, 34, 17-39.
- Liu, X. F., & Wang, Y. (2016). A Contrastive Study of Logos in Chinese and English Doctoral Dissertation Acknowledgements. *Journal of PLA University of Foreign Languages*, 39(06), 27-35+158.
- O'Donnell, M. (2008). The UAM CorpusTool: Software for corpus annotation and exploration. In Proceedings of the XXVI Congreso de AESLA (Vol. 3, p. 5). Spain: Almeria.
- O'donnell, M. (2016). UAM CorpusTool-Versão 3.3. Recuperado de http://www. corpustool. com.
- Ren, X. F. (2014). Study on the Transitivity of The Art of War and Its Translation by UAM Corpus Tool. Proceedings of 3rd Northeast Asia

International Symposium on Language, Literature and Translation, 381-386.

- Williams, S. A., & Dhillon, S. (2019). Women's obstetric and reproductive health care discourse in online forums: perceived access and quality pre- and post-Affordable Care Act. Preventive Medicine, 124, 50-54.
- Xu, X. Y., & Nesi. H. (2019). Differences in engagement: A comparison of the strategies used by British and Chinese research article writers. *Journal of English for Academic Purposes*, 38, 121-134.
- Zhang, J., Tan, L., & Ling, Z. Q. (2015). Interpersonal Functions of the Adverbs and Thematic Structure Analysis of The Cement Garden. *Journal of Tianjin University (Social Sciences)*, 17(04), 361-365.
- Zhuang, B. Y. (2022). A Genre Analysis of English Writing by Junior High School Students Based on Thematic Structure and Thematic Progression. Guangdong Polytechnic Normal University.