EAS Journal of Humanities and Cultural Studies
Abbreviated Key Title: EAS J Humanit Cult Stud
ISSN: 2663-0958 (Print) & ISSN: 2663-6743 (Online)
Published By East African Scholars Publisher, Kenya

OPEN ACCESS

Volume-4 | Issue-5 | Sept-Oct-2022 | DOI: 10.36349/easjhcs.2022.v04i05.004

Review Article

A Review on the Weakness of UAM Corpus Tool

Wang Leyang[1], Liu Zhaoxia[2*]

[1]Associate Professor in Department of Foreign Studies, North China Electric Power University, NO 689 Road, North District, Baoding, Hebei, China
[2]Graduate Student in Department of Foreign Studies, North China Electric Power University, NO 689 Road, North District, Baoding, Hebei, China

Article History
Received: 16.09.2022
Accepted: 24.10.2022
Published: 27.10.2022

Journal homepage:
https://www.easpublisher.com

Quick Response Code

Abstract: Due to the scarcity of evaluation on UAM in previous studies and lack of knowledge of UAM corpus tool for beginners, this paper carries out a comprehensive investigation and makes a review on its merits and demerits, especially demerits. A lot of studies related to UAM from various perspectives within linguistics, mainly describe it as a research instrument rather than a research object. In the meantime, an evaluation standard has been employed in the assessment of UAM corpus tool in this study. According to these principles and other scholars' analysis on UAM, its main weakness contains limited access, restricted applied domain, not sufficiently accurate automatic and manual annotation, among others. In the longer term, this research will help users to understand and utilize UAM corpus tool critically and provide some inspirations to avoid the influence of its weakness as far as possible, for the sake of reliability and validity of their future study.

Keywords: UAM corpus tool, evaluation, weakness, limited access, restricted applied domain, inaccurate annotation.

1. INTRODUCTION

Since the mid-1990s, corpus linguistics is gradually becoming an interdisciplinary branch of linguistics and corpus research has developed rapidly. This is not only thanks to the development of computer hardware such as storage media and character recognition devices, but also owing to the improvement, update and innovation of all kinds of related software, especially tagging and retrieval software.

Nevertheless, faced with numerous annotation tools, it is a challenging and necessary task for researchers to find one that best suits their particular research purposes. As one of the annotation tools, UAM corpus tool was devised in 2007 and updated several versions from then on by Mick O'Donnell. He defined UAM Corpus Tool, as a "software for human and semi-automatic annotation of text and images", "a state-of-the-art environment for annotation of text corpora", which can be used for "annotating a corpus as part of a linguistic study, or building a training set for use in statistical language processing". (O'Donnell, 2008; website: http://corpustool.com/) Many scholars have accepted this tool and applied it to their researches. Recent data makes it clear that this tool has been downloaded 75362 times from its official website (retrieved October 14, 2022, from http://corpustool.com/)

Academic publication relating to UAM corpus tool is an important resource for disseminating the tool (Neves & Seva, 2021) and for assessing the novelty and popularity of the tool. A number of existing studies have documented UAM corpus tool as their research instrument. What has been most demonstrated about it is its usage procedures in details, often with pictures and tables attached in the studies. However, up to now, far too little attention has been solely paid to UAM itself as the research subject or the general assessment of UAM corpus tool, except the developer's related two publications in 2008. The papers titled "Demonstration of the UAM Corpus Tool for text and image annotation" (O'Donnell, 2008) and "The UAM Corpus Tool- software for corpus annotation and exploration" (O'Donnell, 2008) were published 14 years ago, as the most official publications and guidelines for users of this tool. Researches within the 14 years have been collected and systematically reviewed in this current study. Therefore, it is hoped that this study may provide an exciting opportunity to advance the present understanding and usage of UAM corpus tool.

*Corresponding Author: Liu Zhaoxia
Graduate Student in Department of Foreign Studies, North China Electric Power University, NO 689 Road, North District, Baoding, Hebei, China

The prerequisite before attempting to make an evaluation is to find a relatively objective and concrete standard. In order to uphold researchers and annotators in detecting the most appropriate tool for annotation purposes, and to identify defects for tool developers in their tools as well, Neves and Seva (2021) have surveyed 78 annotation tools in their hands-on experiments. Within their studies, there are a set of requirements and criteria to evaluate annotation tools. The full set of requirements involves five points: first, it should be available; second, it should be a web application, either online or downloadable; third, it should be able to be installed in up to 2 hours; forth, it should work for hands-on experiments; fifth, it should allow for the configuration of a schema. In addition, they also established 26 criteria in terms of both functional and technical aspects, and have split their criteria into four categories: (1) publication criteria, (2) technical criteria, (3) data criteria and (4) functional criteria. More accurately, they have elaborated the above criteria with a three-level scale and a score for the final evaluation of the tools.

On the basis of Neves and Seva's standard (2021) and other related researches, this paper reviews UAM corpus tool with the attempt to provide a brief and overall appraisal of UAM Corpus Tool as well as a detailed analysis of its weaknesses.

2. An Overall Appraisal of UAM Corpus Tool

What make UAM stands out are its characteristics, consisting of its authority, accessibility (convenience), feasible operability, and versatile functions. First, "authority" refers to the related researches, which is in agreement with publication criteria (Neves & Seva, 2021). The deviser has published the UAM Corpus Tool Manual in 2007 and his introductory literature in 2008 (O'Donnell, 2008); additionally, UAM possesses an increasing number of researches. Since 2007, numerous scholars at home and abroad who used this tool also published papers from different perspectives within linguistics. Second, accessibility (convenience) means that the tool is free of charge, readily available and easy to install on local disk from official website without too much time and energy consumption. The third characteristic is operability. UAM Corpus Tool manuals in different languages, such as English version by O'Donnell in 2007, Chinese version translated by Liu Xiaohan in 2008, instructions for using UAM Corpus Tool in Japanese by Motoki Sano and Spanish version translated by Mário Martins in 2010 have been provided for users. Manuals can be regarded as the official and systematic learning material to guide users to know about the usage, to operate following the specific procedures, and to pay attention to notions. Fourth, versatile functions include on-board search facilities, cross-layer searching, semi-automatic tagging, production of statistical reports from the corpus, visualisation of the tagged corpus, inter-coder reliability statistics, (O'Donnell, 2008), allowance of multi-label annotations, document-level annotations, saving documents and support for various languages (Neves & Seva, 2021). As Mick O'Donnell set a high value on it, UAM corpus tool is "perhaps the most user-friendly among all the annotation tools available, offering easy installation, an intuitive interface, yet powerful facilities for the management of multiple documents annotated at multiple levels." As a consequence, more and more scholars choose UAM as their research instrument.

Regardless of the above advantages, there are certain problems with the use of UAM corpus tool in terms of access to UAM, applied domain, annotation, etc. The following part analyses the deficiencies of UAM corpus tool in more details.

3. Weaknesses of UAM Corpus Tool

3.1 Limitation in access

UAM corpus tool was classified as non-selected tools in Neves and Seva's research results (2021). This is because it doesn't comply with one of the five requirements mentioned previously, web-based demand. UAM corpus tool is a stand-alone tool, which is required to be downloaded from its official website. Compared to web applications, UAM corpus tool is less convenient and readily available.

On the one hand, a main disadvantage of UAM as a stand-alone tool is that inconvenient access of UAM will probably bring more additional tasks and operational problems. As we all know, manual annotation is a demanding and challenging task. The less additional tasks annotators have to deal with, and the fewer distraction during annotation process they encounter, the better their annotation job will be accomplished. But downloading, updating a single tool, will cost more of users' time and disturb their main task; what's worse, if Macintosh and Windows users run into troubles when installing the tool, they have to find out the corresponding solutions in the "Downloading" section in its official website (http://corpustool.com/). All the above facts indicate the inconvenience of UAM. On the contrary, if a tool is a web application, it will guarantee that annotators can centre solely on the annotation task and save their time and energy in installing target tools, and the annotators can be absorbed in the whole annotation process.

On the other hand, another disadvantage of UAM as stand-alone software is that it struggles to meet users' demand in heavy workload and multi-person verification. Because the capacity of corpus construction is getting larger and larger, and the analysis is becoming more and more complex, stand-alone software may not meet users' demands and be overworked in view of the limited hard disk space and processing capacity of local computers. A large number of corpus need to be directly obtained from the network or uploaded to the network, so website applications tend to provide more choices than stand-alone software for the development of corpus tools (Lu & Hu, 2018). Besides, within stand- alone software, the realization of necessary verification by different people requires delivering different documents annotated to one another. Within annotation software, modification is a task that everyone does alone; and as a consequence, the revisions by different annotators cannot be presented at the same document and at the same time. Conversely, a web-based tool can make it possible that multi-person and real-time collaboration are implementable and modification results also can be traced from the page. So web applications have more advantages over stand-alone softwares. It becomes the reason why web application has become more and more popular with users than stand-alone software.

3.2 Limitation in Applied Domain

There is another potential concern: whether the theoretical foundation of UAM design will confine its application fields. The development of UAM corpus tool is dependent on the theoretical framework of systemic functional grammar. (Liu, 2011; website: http://www.isfla.org/Systemics/Software/Coders.html). Coincidently, most previous researches have failed to use UAM in other domains other than linguistics.

To be honest, this tool is "designed from the ground up to support typical user work flow and everything the user needs to perform a notation task is included within the software" (O'Donnell, 2008). That is to say, UAM corpus tool can be employed to any domains as there are layers provided in this software or layers established by a researcher himself, in accordance with his research objective.

However, the current study has discovered that most papers adopt systemic functional grammar as a theoretical basis and apply UAM corpus tool to analyze the text in the system environment, simplify and facilitate the statistical process, and help researchers find some imperceptible rules. The text analysis mainly focuses on different types of discourse and linguistic features within systemic functional grammar. For one thing, discourse differs from food blogs (Cesiri, 2020), history theses (Sawaki, 2020), invitation letters (Munalim & Gonong, 2019) to news reports (Bao, 2016) and so on. For another thing, different linguistic features have been investigated, including certain specific features like active-passive voice (Munalim & Gonong, 2019), transitivity (Munalim, 2017), engagement (O'Donnell, 2008; Bao, 2016), as well as rhetorical devices like anaphora (Lozano, 2016), ideational metaphor (He & Yang, 2018). And generally, characteristic words embodying these features are annotated in UAM.

Thus, it can be seen that previous studies which use UAM as an annotation corpus tool, have been limited to internal system of language and linguistic subjects, neglecting non-linguistics domains. First, in terms of internal system of language, multiple linguistic levels, like word, phrase, syntax, inter-sentence level and inter-paragraph level, haven't been explored evenly. Most scholars only apply UAM to annotate language at the word level. Second, within language subjects, a large number of published studies (as is mentioned in the last paragraph) does not engage with translation and literature domains. For example, the translation techniques and literary discourse can also be investigated by using UAM. Because researchers can define any layers related to translation or literature. Third, non-linguistics application of UAM is beyond the scope of current studies, like biomedical field (Neves & Seva, 2021) which can also be explored by means of UAM.

3.3 Limitation in Annotation

3.3.1 Introduction of the Annotation System in UAM

UAM Corpus Tool is a corpus tagging program, which adopts the tagging and database-building mode of XML, designed and developed by Mick O'Donnell, which integrates the functions of database construction, retrieval and statistics. The XML-based annotating system possesses diversified features. It allows users to annotate a corpus of text files at a number of linguistic layers; it provides users rights to define different layers and create a hierarchy of tags appropriate for that layer independently based on the research content; users can annotate the text at each layer by swiping the text to indicate a segment, and then assigning features from the tag hierarchy at that layer; the original corpus and the tagged corpus can be stored separately; the same corpus can be tagged in multiple dimensions; it integrates a variety of functions and linguistic concepts, which can be used for syntactic, register, semantic and other multi-dimensional tagging. (Wang, 2013; O'Donnell, 2008)

In fact, in spite of strong technical support for annotation, automatic annotation is not ready. This finding coincides with Heryono's (2020) viewpoint that UAM corpus tool concerns on manual as well as semi-automatic annotation due to unreliable automatic annotation. So most investigations depend on manual annotation. But on the other hand, there are still some problems to be solved in manual annotation.

3.3.2 The Shortcomings of Automatic Annotation

Automatic annotation suffers from restricted linguistic features and lack of accuracy.

Firstly, automatic annotation is unsatisfactory because it can just handle limited linguistic characteristics. Although automatic segmentation into sentences is provided from the previous version of UAM, some linguistic patterns, like semantic or pragmatic features, cannot be easily identified by computers. And the search facility in UAM can be used for semi-automatic tagging of text for limited linguistic features, for example, auto-code segments with the "passive-clause" feature. (O'Donnell, 2008) In order to solve these problems, O'Donnell has been devoting himself to automatic NP segmentation and structural tagging, including co-reference linking, rhetorical structuring and syntactic structuring, etc.

Secondly, automatic annotation has been of poor accuracy. Within the UAM software, the Stanford Parser syntactic tagging program is embedded. This program initially roots in Stanford Parser, a stand-alone syntactic annotating software developed by Stanford University. STNFD Parser in UAM can carry out automatic syntactic tree tagging for the corpus. (Wang, 2013) However, the results of automatic grammar analysis by means of Stanford Parse are not perfect and complete. There are many sentences in the text that are not correctly labeled after manual verification. Grammatical phenomena such as sentence length, punctuation, nonverbal symbols, juxtaposed structures, or omissions are high-frequency areas that cause syntactic annotation errors. Hence, the function of syntactic auto-annotation often has difficulties in tagging certain sentences.

3.3.3 The Shortcomings of Manual Annotation

With regard to manual annotation, it fails to conform to efficiency, objectivity, consistency and accuracy to a certain extent. There exists on obvious causal relationship between objectivity, accuracy and consistency. The deficiency in objectivity has an effect on consistency and accuracy of annotation.

Firstly, UAM corpus tool lacks in efficiency. Throughout the whole process, annotating process is dependent on human beings, from systematic learning and training, skillfully mastering UAM usages before annotation, defining consistent layers, tagging text during annotation, to checking and revising errors after annotation. Furthermore, erroneous operation may bring about great loss and increase annotators' workload, for example, the wrong execution of deleting of a layer. In the Project Management interface, there is a button "Delete" when adding layers. Without withdrawing directive function, executing "Delete" button means that this layer and all analysis at that layer in all texts will be deleted. It is generally used before hierarchical annotating; otherwise, the sub-layer will actually be deleted. Namely, once it happens, annotators have to add a new layer and tag the text again. Therefore, manual collection is time-consuming and laborious.

Secondly, subjectivity of UAM corpus tool is extremely disturbing as well. Individual definition of linguistic layers in UAM leads to the subjectivity of manual annotation. In UAM corpus tool, annotation of multiple texts uses the same annotation scheme, which demonstrates feature description of the text annotation. There are two means to create an annotation scheme: one is to copy existing schemes developed by others, such as pre-installed ones, including Peter White's evaluation network; the other is to design a unique scheme by users themselves according to different research purposes. In such a case, it is hard to ensure the objectivity in defining linguistic layers.

Take Bao's research (2016) as an example, she analyzed Engagement resources in the English news reports of Chinese newspapers and American newspapers on AIIB, by using UAM. She created her annotation scheme on a basis of scheme of engagement system by James R. Martin and Peter White. The list of Bao's classification of engagement resources is not exhaustive enough, which may probably result in divergences of identifying which sublayer an expression belongs to, or results in ignorance of some linguistic elements or resources. Thus, the process of manual annotation tends to be filled with subjectivity of freedom.

Thirdly, the difficulty of guaranteeing the consistency of manual annotating standards is quite a handicap for UAM users. During annotation process, any deviation from annotating standards may affect the study results. If the criteria of layers within and among groups cannot be unified and objective, there will be some disagreement. Peng, Yang and He (2012) have realized the significance of consistency of manual labeling and they proposed a cyclic process from "group labeling", "cross-check", "resolving differences and unifying standard" to "whole discussion group", "resolving differences and unifying standard once more" and "group labeling". It can greatly avoid the significant variation among researchers in manual labeling, and the consistency of manual labeling is basically guaranteed. Researchers can learn from this method and use it for reference on the issue of inconsistent manual annotation criteria.

Fourthly, a lack of accuracy is a real hassle as well. Inaccuracy results from various reasons, such as subjectivity of annotating, inconsistency of annotating principles and carelessness in the annotation process. Subjectivity of annotation can bring about different annotation schemes, layers and segments in the texts; inconsistent annotation criteria tends to cause disagreements of annotation results within and among different teams of annotators; and individual carelessness may lead to annotating mistakes, which can be solved by checking out several times. The accuracy of annotation plays an essential role in a study. It is the foundation of the subsequent statistical analysis, uncovering oblivious rules and ensuring reliable results of researches. Only with accurate annotation, can validity of the investigation be ensured to some degree.

To sum up, there are unfavorable qualities associated with the annotation of UAM corpus tool: for one, automatic annotation is only able to deal with limited linguistic feature and its results are not so complete and correct; for another, manual annotation may fail to ensure the efficiency of annotating task, the objectivity of annotating procedure, consistency of annotating standards and accuracy of annotating results.

4. CONCLUSION

This paper presents a brief review of the UAM corpus tool, an annotation tool for documents of text and images. A number of studies related to UAM from different perspectives at home and abroad have been reviewed. Most scholars have more or less made their assessment on advantages and disadvantages of UAM in their studies.

UAM corpus tool is attractive because its authority, accessibility (convenience), operability and versatile functions. But on the other hand, the main deficiencies of UAM includes: the access to UAM by downloading a stand- alone tool instead of web-based tool bringing more troubles; the potentially negative effect of theoretical framework of tool creating on applied domain of UAM; immature automatic annotation technology; inefficiency, subjectivity, inconsistency, inaccuracy, of the manual annotation. In addition to above mentioned limitations, there are certain disadvantages, such as imperfect corpus retrieval and statistical analysis.

It is hoped that this review may provide users with an overview in utilizing UAM corpus tool effectively and in finding solutions to overcome its weakness. To overcome the tool's shortcomings, users can combine UAM corpus tool with other tools to ensure the reliability and validity of their research results.

ACKNOWLEDGEMENTS

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

REFERENCES

- Bao, T. (2016) A Comparative Analysis of Engagement Resources in the English News Reports of Chinese Newspapers and American Newspapers-Take Reports about AIIB as Examples. (Master's dissertation). Anhui University. Available from Dissertations & Theses database in CNKI.
- Cesiri, D. (2020). The Discourse of Food Blogs: Multidisciplinary Perspectives. New York: *Routledge*.
- He, Q., & Yang, B. (2018). A corpus-based study of the correlation between text technicality and ideational metaphor in English. *Lingua*, 203, 51-65.
- Heryono, H. (2020). Optimizing UAM corpus for transitivity process regarding to covid-19 news in Indonesia from February to March 2020. *English Journal Literacy Utama*, *5*(1), 325-333.
- Liu, J., & Yang B. (2011). A Corpus based Study of Lexical Tendency in Systemic Functional Grammar. *Modern Foreign Languages*, 34(04), 364-371.
- Lozano, C. (2016). Pragmatic principles in anaphora resolution at the syntax-discourse interface: advanced English learners of Spanish in the CEDEL2 corpus. Margarita Alonso-Ramos, Spanish Learner Corpus Research: Current trends and future perspectives. Amsterdam: *John Benjamins.*
- Lu, Z., & Hu, J. (2018). An overview of the development of modern corpora in the context of Big Data Era. *Foreign Languages and Translation*, 25(04), 39-44, 98.
- Munalim, L. O. (2017). Mental processes in teachers reflection papers: A transitivity analysis in Systemic Functional Linguistics. *3L-Language, Linguistics, and Literature: The Southeast Asian Journal of English Language Studies*, 23(2), 154-166.
- Munalim, L. O., & Gonong, G. O. (2019). Filipino Students' Active-Passive Voice Preference in Invitation Letters. *The Normal Lights*, 13(1), 151-178.
- Neves, M., & Seva, J. (2021). An extensive review of tools for manual annotation of Documents. *Briefings in Bioinformatics*, 22(1), 146-163.
- O'Donnell, M. (2008, April). The UAM Corpus Tool: Software for corpus annotation and exploration. *Proceedings of the XXVI Congreso de AESLA*, Almeria, 3-5. Spain.
- O'Donnell, M. (2008, June). Demonstration of the UAM Corpus Tool for text and image annotation. *Proceedings of the ACL-08: HLT Demo Session*, 13-16. Columbus.
- Peng, X., Yang, X., & He, Z. (2012). Chinese-English Parallel Corpus of Appraisal Meanings. *Technology Enhanced Foreign Language Education*, (05), 3-10.
- Sawaki, T. (2020). Interacting voices structure a text: A quantitative investigation of dialogic elements across structural units in the introductory chapters of history theses. *Functions of Language*, 27(2), 174-206.
- Wang, D. (2013). An Evaluation on Multilayer Parsing TEGC Corpus by Using UAM-based Stanford Parser. *Electronic Test*, (09), 201-202.

Cite This Article: Wang Leyang & Liu Zhaoxia (2022). A Review on the Weakness of UAM Corpus Tool. *EAS J Humanit Cult Stud, 4*(5), 214-218.