The Use of UAM-Corpus Tool for a More Comprehensive Text Analysis

Idha Nurhamidah
Linguistic Department, Faculty of Cultural Sciences
Universitas Sebelas Maret
Surakarta, Indonesia
idhanurhamidah@student.uns.ac.id

Riyadi Santosa
English Department, Faculty of Cultural Sciences
Universitas Sebelas Maret
Surakarta, Indonesia
riyadisantosa@staff.uns.ac.id

Djatmika Djatmika
English Department, Faculty of Cultural Sciences
Universitas Sebelas Maret
Surakarta, Indonesia
djatmika@staff.uns.ac.id

Henry Yustanto
Indonesian Department, Faculty of Cultural Sciences
Universitas Sebelas Maret
Surakarta, Indonesia
henryyustanto@staff.uns.ac.id

Abstract— **The so called UAM-Corpus Tool has been claimed as a state-of-the-art environment for annotation of Text Corpora. This article aims at exemplifying how it can be used to construe the ideational meaning of a text leading to the description of the field of the discourse. Described step by step are the procedures for using the tool in order for linguistic analysts to perform text analysis more comprehensively and effectively than employing the traditional ways of text parsing. This tool can both present the statistics and linguistic evidence for the researchers as the basis of the analytical description and discussion of the research findings as a whole. Schema building for the componential analysis of transitivity is exemplified to analyze one short text on which to construe the ideational meaning. The article concludes that UAM-Corpus Tool may also be used to deal with any other componential analysis of text corpora.**

Keywords—UAM-corpus tool; annotation; scheme; text analysis

I. INTRODUCTION

It has been a tradition, over years of practice by undergraduate, master or even doctoral students, to parse sentences or clauses into phrases or words to investigate how the smallest linguistic units are syntactically manipulated to form larger constructions [1]. Not only is such a componential analysis [2] applied in Traditional Grammar at syntactic level but also in Systemic Functional Grammar (Linguistics) to deal with Transitivity Analysis [3] at the ideational meanings and meta-functions, Mood Analysis [4] at the interpersonal meanings and meta-functions, and Theme-Rheme Analysis [5] at the textual meanings and meta-functions. Text parsing [6] helps describe the linguistic phenomena by means of frequency counts or any other interpretation that may apply according the purpose of the study.

Normally parsing can be undertaken by bracketing or tabulating, such as done in the so-called immediate constituent analysis. This type of parsing is really boring and takes long time to complete, especially when dealing with longer texts such as speeches, moreover when some components may stay uncategorized (mostly in spoken language), which even causes "headaches" on the part of the students and teachers alike. Long and multiple slots have to be provided to accommodate each componential unit of a text. This may occupy a large piece of paper. However, the emergence of computer technology has been of some help. Students may create tables in Words or Excell even though such works may still cause some boredom and frustration when faced with complex structures of the clauses or sentences.

In response to the complex and tedious works of sentence or clause parsing in text analysis, applied and computational linguists have been working hard to come up with a product called UAM-Corpus Tool [7], which can be downloaded for free [8]. Alternatively, you can work 'Online'[8]. This product is compatible with Windows and Macintosh, and made user-friendly with moderate level of digital literacy.

II.   LITERATURE REVIEW

A. Why UAM-Corpus Tool

UAM-Corpus Tool hereinafter referred to as UAMCT is likely a great revolutionary breakthrough in text annotation for the 21st Century. It can accomplish both annotating a corpus as part of a linguistic study and establishing a training program in statistical language processing. Featured as distinctive points of this product are (1) annotation of multiple texts employing the same schemes designed by the user; (2) annotation of each text at multiple levels, such as NP, Clause, Sentence, including whole document; (3) ability to search for instances across levels as linguistic evidence, such as *finite clause containing company-np, particular verb processes throughout the text* etc. as assigned by the user; (4) comparative statistics across subsets, for example *contrasting conversational patterns* as used by *male and female speakers;* (5) all annotation is stored in XML files for the user to possibly share with other applications; (6) the use of "Stand-off" XML enabling multiple overlapping analyses of text.

Once downloaded (for those who prefer to work offline), this product provides a prospective user with simple and easy installation processes into a computer set. With a moderate level of digital literacy, users can easily get access to the user's guide, such as creating a project, data importing, making componential scheming etc. The operational processes as outlined in the manual are all easy to follow. A study on optimizing UAM Corpus tool [9] has also been conducted to support transitivity analyses for texts related to Covid-19 from February to March 2020. It is also argued that UAM-Corpus Tool is a convenient and comprehensive software [7]. This software's weaknesses have also been criticized, along with possible solutions offered therein [10]

B. Publications by Users of UAM-Corpus Tool

UAMCT, despite its recent existence in Indonesian linguistic research contexts, has actually been very much utilized all over the world as a powerful tool for text analysis. A study on 'Move'—steps of how units were linked to form lexical bundles in conclusion sections in research articles. It was methodologically stated that cconclusions containing lexical bundles were then imported into the UAM-Corpus tool for manual tagging of functions, structures, moves, and steps associated with lexical bundles [11]. This software was also particular used as exemplification in researches [12]. Political keywords have been abstracted, using this corpus tool [13]. Other publications in 2022 and traced 5 years back, even more, can be checked where the UAM-Corpus tool was used to perform componential analyses—prevalent all over the world, except Indonesia. This may be caused by lack of information how to report research as evidence of using UAM-Corpus tool. Therefore, it is high time, in Indonesian research context, that this kind of Corpus Tool be used in text analyses in place of traditional componential parsing and to report the research activities to the UAM-Corpus Website.

C. Componential Analyses in Social Research

A componential analysis [14] is very commonly performed in social research in which categorization becomes one of the research procedures. During the process of analysis, data may be parsed or broken down into smaller units or components to investigate how they are functionally, semantically and or syntactically related to form one unified whole. From this point, the frequency counts and distribution can be identified as the basis of further analysis leading to research findings and discussion. As previously mentioned, the UAM-Corpus tool [15] can accurately perform scheme building for the componential analysis and statistical operation, such as frequency distribution (percentage) and comparative components across single or multiple texts. The data are automatically presented in tables for ease of analysis, including displays of linguistic evidence in just several clicks on the corpus tool.

As a social science, linguistics which, in this article, is referred to as Halliday's Systemic Functional Linguistics (SFL) [16] is of no exception—that is no way of avoiding componential analysis to arrive at the research findings. Theoretically speaking, Halliday plays around with language as meaning making in order for the language users to socially and linguistically function in any assigned discursive practices. It is further argued that a language is used in three meta-functions, namely ideational, interpersonal and textual.

The ideational meta-function is systemically analyzed through its transitivity system [17] in which the componential analysis includes investigating the participants, verb processes and circumstances to arrive at the field of the discourse. Meanwhile, Mood analysis [18] is applied to the interpersonal meta-function through investigating the Mood system consisting of the subject and the finite verb in a particular clause to arrive at the tenor of the discourse. Finally, the textual meta-function [19] is analyzed through its thematic structure to arrive at the mode of the discourse. In short, SFL yields rich findings in lexico-grammatical features socially and functionally appropriate. Thus, not only does it deal with how language structurally work but also how it socially functions in three different but simultaneous meta-functions of language—ideational, interpersonal and textual.

III.   METHOD

This article is to introduce the so-called UAM-Corpus Tool for componential analysis. The materials with respect to features of the tool were extracted from its website. A step-by-step procedure to create one exemplary research project would be presented, among others including (1) naming a project, (2) creating componential schemes, (3) exampling componential analysis, and (4) showing statistical results.

In this article, we give you an example of text analysis of a short text dealing with the ideational meta-function of Halliday's Systemic Functional Linguistics (SFL) [20] through the analysis of the transitivity system [21] leading to the description of the field of the discourse.

IV.   RESULT AND DISCUSSION

Presented below is a suggested procedure of operating the UAM-Corpus Tool. It is, however, recommended that prospective users be equipped with a moderate or intermediate level of computer or digital literacy to avoid unnecessary technical problems during the operational processes.

A. To Start a Project

It should be again noted that operating the UAMCT may be performed both offline and online. Both start by getting access to the UAMCT Website to get the tool bar displaying (1) Home, (2) Feature, (3) Documentation, (4) Contact, (5) Research (6) Downloading (7) Online. If you want to undertake the research project offline, you can simply click on 'Downloading'. Conversely, if you prefer to work online, you can then click on 'Online'. You have to make your choice immediately and follow the instructions accordingly. In this paper, we try to use the online system. For this, you are supposed to create an account by entering your email address, followed by a preferred password in order for you to later login to the system.

When you manage to login, you will see the following screen and have immediately click on "Start New Project".

Fig. 1 Starting New Project

What you have to do next is to follow all the instructions, such as (1) naming the project, and (2) adding a text. In case you get confused in this point, simply click on the folder with the name of the project, then you will see the screen as follows:

Fig.2 Adding Files to Corpus

What you have to do next is to "Add Files to Corpus" and you may not know how to add a new file to the system. If that is the case, you can simply click on "Help". The system will display clear instructions; then click on Point 3 "Adding Files" under "Tutorial". By 'File' is meant the text you are going to analyze.

B. To Create Layers (Schemes)

In order to create "layers" or scheme, you can simply click on "Layers" on the tool bar as shown in Fig.2, then "Add New Layer". At this stage, you will be asked "How do you want to annotate?" as shown in Fig.3 below:

Fig.3 Selecting Type of Annotation

When you select "Automatic Annotation", the program will automatically annotate your texts, using a parser or tagger. Conversely, when you choose "Manual Annotation", you will have to identify segments in each text, and assign features to them. You may want to explore all features in both "Automatic Annotation" and "Manual Annotation to see if the program has the type of scheme (annotator) you want to use in your analysis. When you do not find a suitable scheme for your analysis, you have to select "Manual Annotation", then click on "Design Your Own" in which you will have to manually code segments within a document, e.g., clauses, phrases, etc.

We are going to give you an example of automatic annotation. We can click on "Automatic Annotation", then choosing "Grammar" in which we are offered several grammatical features, finally selecting "SFL Transitivity" as shown below
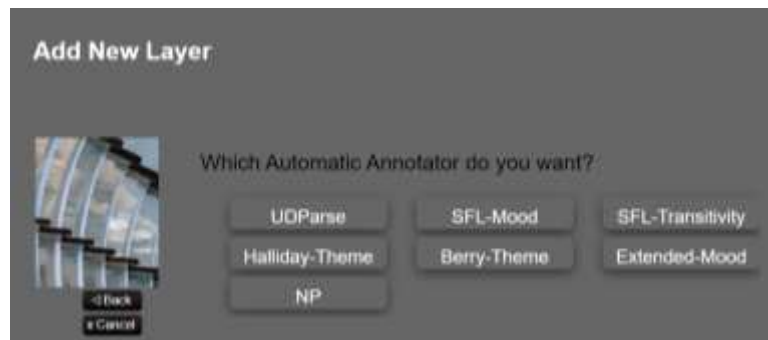
Fig.4 Creating Layers

By clicking on "SFL-Transitivity, the scheme (annotator) is automatically created for you. Conversely, when you the manual operation, you have to follow the instructions. Finally, you have then to click on "Edit Scheme" to start creating a scheme of your own design. When you find difficulties, you can always consult the problems by clicking on "Help", and follow the instructions. Both choices have each own advantages. Automatic annotation is already well-set to analyze the text. It knows what features are under analysis, and what features may be excluded from the analysis. Meanwhile, manual annotation requires you to segment the text to get matched with the scheme you designed. Unless you are equipped with thorough knowledge with the theory, it is recommended that you employ the automatic annotation.

It is important to note that in the use of automatic annotation, the text under analysis must be well-written— free from grammatical errors, including the writing mechanics. Grammarly software application can be used to clear up the text from errors prior to copying and pasting the assigned text on to the UAM-Corpus Tool System.

Displayed below is an automatic scheme (annotator) to analyze the ideational meaning (meta-function) of a text adopting Halliday's SFL- Transitivity.
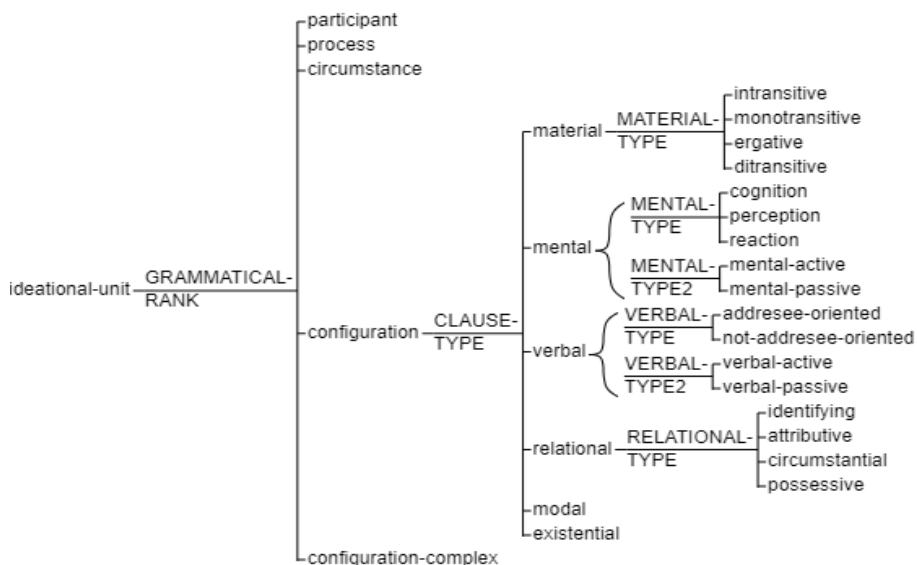
Fig.5 Transitivity Scheme (Annotator)

Theoretically, the ideational meta-function of a text is analyzed through its transitivity system which, in terms of components, consists of Participants, Process and Circumstance [22]. As the unit of analysis is "Clause", a text is analyzed in terms of Clause Type Configuration, including complex configuration. The central operation in transitivity system is the verb process, which is, consequently, further sub-systematized. Depending on the purpose of your research, you can still add more subsystems, by clicking on 'Edit Scheme'. Therefore, it is logical that researchers prefer to use manual annotation where they are free to include whatever items necessary.

C. To Perform Componential Analysis

In automatic annotation, things are all made easy. We can know the results of clause parsing in one click. From the tool bar, we can simply click on "Annotation", the result will display as follows (randomly) sampled out:

Fig.6 Clause Parsing

Fig.6 above shows that Root, as Sentence Representation, consists of two clauses, ideationally focusing on the word "data". It undergoes a passive verbal process "are taken" circumstantially "from the Model United Nations conference speech". The circumstance is further qualified by a sub-clause "which was held on 26th-27th March 2022". The sub-clause is further transitively analyzed as "Goal" referring to which (conference speech), undergoing a passive verbal process "was held" and further circumstantially qualified by "on 26th-27th March 2022".

From this point, the sub-field of the discourse can be formulated as having a short-term goal of informing the reader about where the data were taken (source of data) and when (time of data collection), and a long-term goal of partially informing the research report as a whole—that is the data collection method.

D. To Show the Statistical Results

Very often, researchers want to describe particular characteristics of the text by means of statistics. This can be done simply by clicking on "Statistics". There are a number of choices, depending on the researchers' interests, such as lexical patterns, feature patterns, inclusion patterns, or wordings. We randomly give you an example of lexical patterns, as follows:

Fig.7 Lexical Patterns

In terms of lexical patterns, the text, according to Fig.7, belongs to a written type of text, evidenced by the lexemes per segment of 3.8 and lexemes % of text of 60%, meaning that the text is of high lexical density. In terms of the author's subjective positivity, it is quite low (0.339), indicating a very good scientific text, and so is the subjective strength of only 0.292. Furthermore, it is strengthened the academic word use of 12.24% and academic rareness (uniqueness) of 2.255, all confirming that it is an academic text.

E. Limitation of this Article

Due to the constraints of time and the writer's limited knowledge on UAM-Corpus Tool, this article cannot be claimed as a thorough training manual. Rather, it can only be thought of as an introduction to the UAM-Corpus Tool. There are more features that have not been covered, especially the step-by-step instructions to create "Manual Annotator".

V.   CONCLUSION

The UAM-Corpus Tool has been introduced as a tool in linguistic analysis to replace the traditional parsing of sentences or clauses into their smaller componental parts for a better and more comprehensive analytical performance. This tool has proved efficient and accurate to perform componental analysis, especially dealing with text analysis. It is therefore recommended that such a tool be used in any linguistic research where componental analysis is required—either applying the automatic annotation or the manual one. Interested individual researchers should get access to UAM-Corpus Tool Tutorial for a better understanding of how it operates. There is only one way to master the use of UAM-Corpus Tool—direct practice with a selected research project. At one time or another, later on, when offered, we can join a complete training for the optimized use of the UAM-Corpus Tool, which up to now is still developing a new series, adding up more features in anticipation of more complex phenomena of social research, especially dealing with advances in computational linguistics in the digital era. It can also be argued that UAM-Corpus Tool may still be more optimally used to analyze data in other types of social researches.

VI. ACKNOWLEDGMENTS

VII. REFERENCES

[1]   G. Kazeminejad, M. Palmer, S. W. Brown, and J. Pustejovsky, "Componental analysis of english verbs," *Front. Artif. Intell.*, vol. 5, p. 780385, 2022, doi: https://dx.doi.org/10.3389/frai.2022.780385.

[2]   F. H. Aziz, "Componental analysis to John Steinbeck's Of Mice and Men's Characterization: A Semantico-Stylistic Study," *Hawlyat Al-Montada*, vol. 1, no. 53, 2023, [Online]. Available: https://www.iasj.net/iasj/article/261663

[3]   F. Liu, M.-Y. Qiu, and D.-H. Zhou, "Transitivity analysis of preference relations under the impact of indifferent voters," *Granul. Comput.*, pp. 1–16, 2023, doi: 10.1007/s41066-023-00380-3.

[4]   D. Luthfiyati and O. Purwati, "Mood System of Joe Biden's Speech on The Covid-19 Vaccination Program," *Elit. J.*, vol. 5, no. 1, pp. 117–124, 2023, [Online]. Available: http://www.elitejournal.org/index.php/ELITE/article/view/124

[5]   S. T. Siahaan, "An Analysis of Theme Types of Luke Chapter 17 in the English Standard Version Bible," *J. Teach. Educ.*, vol. 4, no. 3, pp. 271–284, 2023, doi: 10.31004/jote.v4i3.12313.

[6]   M. Yin *et al.*, "Two-stage Text-to-BIMQL semantic parsing for building information model extraction using graph neural networks," *Autom. Constr.*, vol. 152, p. 104902, 2023, doi: 10.1016/j.autcon.2023.104902 Get rights and content.

[7]   L. Wang, "Di Yu (2022). UAM Corpus Tool: A Convenient and Comprehensive Software," *Sch Int J Linguist Lit*, vol. 5, no. 10, pp. 352–356, 2022, doi: 10.36348/sijll.2022.v05i10.006.

[8]   M. O'Donnell, "UAM CorpusTool." p. 6, 2023. [Online]. Available: http://www.corpustool.com

[9]   H. Heriyono, "Optimizing UAM Corpus for Transitivity Process Regarding Covid 19 News from February to March 2020," *English J. Lit. Utama*, vol. 5, no. 1, pp. 325–333, 2020, doi: https://doi.org/10.33197/ejlutama.vol5.iss1.2020.484.

[10]   W. Leyang and L. Zhaoxia, "A Review on the Weakness of UAM Corpus Tool," 2022, doi: https://dx.doi.org/10.36349/easjhcs.2022.v04i05.004.

[11]   L. Deng and J. Liu, "Move--Bundle Connection in Conclusion Sections of Research Articles across Disciplines," *Appl. Linguist. Oxford Acad.*, vol. 14, no. 3, pp. 527–554, 2022, doi: https://doi.org/10.1093/applin/amac040.

[12]   N. Triki, "Exemplification in research articles: Structural, semantic and metadiscursive properties across disciplines," *J. English Acad. Purp.*, vol. 54, p. 101039, 2021, doi: https://dx.doi.org/10.1016/j.jeap.2021.101039.

[13]   I. Dilai and M. Dilai, "Automatic Extraction of Keywords in Political Speeches," in *2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT)*, 2020, vol. 1, pp. 291–294. doi: https://dx.doi.org/10.1109/CSIT49958.2020.9322011.

[14]   A. Bakenova, S. Mazhitayeva, S. Kenzhegaliyev, A. Zhartybayev, Z. Balmagambetova, and M. Yesmatova, "Componental Analysis of Ana/Mat'/Mother Words: Mother Prototype Extension," *Theory Pract. Lang. Stud.*, vol. 13, no. 6, pp. 1413–1420, 2023, doi: 10.17507/tpls.1306.09.

[15]   A. Skichko, "UAM corpus tool: an effective software program for the analysis of motivational speeches," *Sci. Collect. «InterConf»*, no. 158, pp. 135–137, 2023, [Online]. Available: https://archive.interconf.center/index.php/conference-proceeding/article/view/3848

[16]   A. Llinares and A. McCabe, "Systemic functional linguistics: the perfect match for content and language integrated learning," *International Journal of Bilingual Education and Bilingualism*, vol. 26, no. 3. Taylor & Francis, pp. 245–250, 2023. doi: 10.1080/13670050.2019.1635985.

[17]   D. E. N. Ikram and Z. Hussain, "The Transitivity Analysis of Sunja's Character in Pachinko," *PalArch's J. Archaeol. Egypt/Egyptology*, vol. 20, no. 2, pp. 493–506, 2023.

[18]   U. A. Azzahra and D. Rukmini, "The Mood Analysis as Interpersonal Meaning in the Donald J. Trump Presedential Letter Addressed to Recep T. Erdogan on October 9th 2019," *HUMANIS J. Ilmu-Ilmu Sos. dan Hum.*, vol. 15, no. 2, pp. 21–28, 2023, doi: 10.52166/humanis.v15i2.4247.

[19]   S. N. Hidayah, M. Wihadi, and W. Rahmatunisa, "Textual Metafunction Analysis about Education News Text on Selected National Online Newspapers," *Journey J. English Lang. Pedagog.*, vol. 6, no. 2, pp. 498–504, 2023, doi: 10.33503/journey.v6i2.3185.

[20]   L. Law and C. M. I. M. Matthiessen, "Revisiting Halliday (1990)'New Ways of Meaning: The Challenge to Applied Linguistics': What has Changed and What Still Needs to be Done?," *Linguist. Hum. Sci.*, vol. 15, no. 3, pp. 337–368, 2019, doi: 10.1558/lhs.23599.

[21]   H. Heryono, "Optimizing UAM corpus for transitivity process regarding to covid-19 news in Indonesia from February to March 2020," *English J. Lit. Utama*, vol. 5, no. 1, pp. 325–333, 2020, doi: 10.33197/ejlutama.vol5.iss1.2020.484.

[22]   M. A. K. Halliday, *Introduction to Functional Grammar-2nd Edition*, Second. Lodond, Merlbourne, Auckland: Edward Arnold, 1994.