International Journal of Innovation Scientific Research and Review
Vol. 02, Issue, 12, pp.608-612, December, 2020
Available online at http://www.journalijisr.com

IJISRR

Research Article

ISSN: 2582-6131

AUTHORSHIP ATTRIBUTIONS OF DAN BROWN, A CORPUS-BASED STUDY

*** Muhammad Imran Shah[1], Rabiya Nawaz[2], Eram Jamil[3]**

[1] PhD scholar (Applied Linguistics) UUM Malaysia
[2] Department of Applied Linguistics Government College University Faisalabad
[3] Assistant Professor University of Sargodha Sub-campus Bukkar

Received 25th September 2020; Accepted 18th November 2020; Published online 30th December 2020

ABSTRACT

Authorship attribution is a science of identifying the characteristics of the author from the text of the document he composed. The question of authorship is as old as the biblical books. Due to the advancement in technology, there are various ways to identify the writer's attributes in the text he/she produces. In this study, the authorship attribution of an American writer; Dan Brown has been investigated through the use of a specific discourse marker i.e. conjunctions. Corpus tools were used to tag and identify the conjunctions. The study not only focuses on the analysis of conjunctions but also the way the writer has used them to create coherence in his novels. Halliday and Hasan's (1976) model of cohesion was used while analyzing the data of the study. The results of the analysis show that the writer uses more additive conjunctions to maintain coherence in his writing. The study has strong research implications for the young linguists to find the authorship characteristics in the relevant texts of various authors.

Keywords: Authorship attribution, Conjunction, Discourse markers, Coherence, Cohesive devices.

INTRODUCTION

Writing varies from person to person, Just like no two minds think alike, no two writers have the same writing style. It all depends on how good a person is in expressing their thoughts and feelings through the proper use of coherent texts. Authors can be identified through their writing style, use of words, sentences and stylistics features etc. basically through their linguistics features. Authorship attribution is a procedure of identifying or recognizing the author of a given text. In today's world of increasing anonymous information, author attribution plays a significant role in various fields such as in detecting plagiarism, identifying the writer of an anonymous text or a text written under a pseudonym and resolving the issues of disputed authorship. It can also be used to know the real author when two people claim that they have written the text or when no one confesses the authorship (Juola, 2006). The linguistic evidence to verify authorship suggests that each speaker or writer uses a special idiolect which is expressed in the text by idiosyncratic and distinct choices (Halliday, McIntosh & Strevens, 1964). Using linguistic techniques for identification of the author started by Mendenhall in 1887 who started this by counting word length whereas Morton and Yule used sentence length for authorship attribution (Olsson, 2008). Others used features such as lexical repetition, word frequencies, and linguistic features. For finding the authorship of disputed works, Mosteller and Wallace (1964) studied the frequency of function words and found out that Hamilton or Madison was the author of the papers in The Federalist Papers. Different scholars suggested different and yet the same criteria for identifying the authorship of a text. Holmes (1985) suggested word length, sentence length, word frequency, lexical items i.e. their ratio of type-token words. On the other hand, parts of speech, average word and sentence length, and vocabulary were the features suggested by Allen (1974) whereas Foster (1996) focused on finding the syntax, unique words, and spellings in the text for authorship attribution. Stylometric features were given importance in this context. Stylometry is the style of an author and no matter how much he tries to change his style in writing, he will not be able to do it. Style markers are another feature for identifying authorship (Corney, 2003) This study is focusing on authorship attribution through discourse markers specifically through conjunctions. Discourse markers are also called pragmatic markers, discourse connectives or particles, etc. These are small linguistic items that play a significant pragmatic function in a text or conversation. (Anderson, 2001). They build a relationship between two sentences and different conjunctions are used according to the properties of the sentences. It can be syntactic (despite, and, but, etc.), connotations (after all, alternatively, again, so, therefore, etc.) and anaphoric expressions (therefore, by contrast, besides, etc.) (Fraser, 1990). According to Dulger (2007), there are fifteen categories of discourse markers and some of them are conjunctions, substitutions, verification, conclusion, summarizing, persuasion, etc. Discourse markers provide cohesion in a text. The cohesive relationships that tie the components of the text together are references (referring forward and backward in a text), conjunctions, substitution (replacing one linguistic item with another), the ellipsis (the absence of linguistic items), and lexical relationships (hyponymy, synonymy, collocation). The main function of coherence is to maintain the distribution of information in the text by topicalization, continuity, extending, introduction, and adding (Dijk, 1977).

"Cohesion is the semantic relation between one element and another in a text (Halliday & Hassan, 1976). A text is cohesive when the elements are tied together and considered meaningful to the reader. Cohesion occurs when the interpretation of one item depends on the other, i.e. one item presupposes the other."

(Bahaziq, 2016: p. 112)

Conjunctions in a text can be identified by their properties of linking makers and justification etc. The logical conjunction shows the arrangement of sentences as well as the sequence of thoughts in inter text i.e. cause &effect, addition, conclusion, and distinction (Gunay, 2007). Types of conjunction according to Halliday and Hassan theory of cohesion (1976) are the following

1. Additive – and, also, moreover, besides, nor, etc.
2. Adversative – but, despite, yet, anyhow, rather, etc.
3. Causal – so, thus, because, hence, etc.
4. Temporal – finally, then, after, before, etc.

***Corresponding Author: Muhammad Imran Shah,**
PhD scholar (Applied Linguistics) UUM Malaysia.

The study explores the conjunctions as cohesive devices used, particularly, by 'Dan Brown' in his novels, and proves this salient feature as the attribution of this author.

Research problem

'Authorship attribution' is the newest problem in information retrieval. Disputes on the ownership of text have always been around. The interest was developed in this area after reading several novels of various writers. Certain differences were detected in their writing styles, choice of words, stylistics techniques, discourse markers which they used. It has been decided to research and find out how we can identify authorship attribution through the way they have used discourse markers (conjunctions in particular). The distinctions in their use of conjunctions have been identified, their frequency of using all four types of conjunctions, and how this information can help in authorship attribution. Hence, an anonymous author can also be detected by the way he/she uses conjunctions as discourse markers.

Research Questions:

Through this study, following questions have been tried to answer:

1. How can discourse markers help in authorship attribution?
2. What is the frequency of the usage of various conjunctions in the novels of 'Don Brown'?
3. What are the characteristics of Dan Brown's writing found through conjunction as discourse markers?

LITERATURE REVIEW

Holliday & Hassan (1976) have categorized the grammatical cohesion into four types:

- Reference
- Substitution
- Ellipsis
- Conjunctions

Among these types, conjunctions are linking devices between sentences or clauses in a text. Besides grammatical devices, conjunctions express the 'logical-semantic' relation between sentences rather than between words and structures (Holliday & Hasan, 1976). In other words, they structure the text in a certain logical order that is meaningful to the reader or listener. There are four types of conjunctions, namely additive, adversative, causal, and temporal. Additive conjunctions connect units that share a semantic similarity. Examples of additive conjunctions are, and, likewise, furthermore, in addition, etc. "Adversative conjunctions are used to express contrasting results or opinions. This type of conjunction is expressed by words such as, but, however, in contrast, whereas, etc. Causal conjunctions introduce results, reasons, or purposes. They are characterized by the use of items such as, so, thus, therefore, because, etc. Temporal conjunctions express the time order of events such as, finally, then, soon, at the same time, etc" (Bahaziq, 2016).

Conjunctions are also used as a discourse marker, a phrase that is used to organize discourse into segments. The use of conjunctions is, therefore, considered the salient feature of authorship. Several researchers have attempted to find the 'authorship attribution' while using structural approach In Soler- Company and Leo Warner's (2017), the authors selected two data sets (Literary dataset and PAN Literary dataset) and found out four things i.e. gender identification, author identification, sourcebook identification, PAN author identification, and feature analysis by doing three experiments on literary datasets and one on PAN Literary dataset. The research concluded that discourse and syntactic features play a major role in the identification of the author and gender as well as author verification. The use of discourse markers by Tiryaki (2016) in justification types in argumentative texts of Turkish language teacher candidates, was a qualitative research in which survey model was used. This research found out 62 discourse markers that were used in two different aspects i.e. support and refutation justification and concluded how these types of studies can be used in a phased manner in the Writing and Writing Skills Course. Similarly, In BASIM (2012) the researcher applied the techniques of applied linguistics and forensic linguistics in identifying the authorship attribution of a suicide note. The suicide note was compared to another text which was written by the deceased. The researcher identified micro and macro linguistic features. The former was done to see whether the suicide note was written by the deceased or not and the latter was done to see if the suicide note was written under pressure or by someone else. It was concluded that the suicide note was not written under duress or threat.

The research article written by Belisa and Zufferey (2010) shows how discourse markers such as like and Well can play other roles as well i.e. as verb or adverb and analyzed the application of automatic classifiers trained via machine learning for disambiguating discourse marker son these ambiguous items. It was concluded that discourse markers are not a homogenous class and the lexical items such as like and well should be processed separately. Another study Duran, Sidorov, and Batyrshin (2014) showed the authorship attribution through syntactic n-grams as style markers. The writings of a few authors were studied through their short texts and then the researcher tried to associate a text whose authorship was unknown to one of them. The machine learning technique and specifically the vector space model was used. For modeling the writing style of the authors, the complete syntactic n-grams were applied. It was concluded that the syntactic n-grams are more effective than character n-grams because through them accurate results can be achieved. Moreover, Rexha, Kro, Ziak, and Kern (2018) studied how the humans identify and judge different writing styles. The researchers did two experiments. One was quantitative in which stylometric and content features ere studied and analyzed while the second experiment was qualitative in which the researchers evaluate the process and features on which the humans judge the writing styles. This study could help in plagiarism detection, automatic authorship attribution, and help forensic linguists in treating writer's profiles. Very scarce work has been done on discourse markers as the trait of the author in the text. Therefore the present research work is moving knowledge ahead in this area of linguistics.

METHODOLOGY

In this study, the pragmatic paradigm (Cresswell, 2014) will be used as it opens the door to multiple methods, different worldviews, different assumptions as well as different forms of data collection and analysis. Pragmatism acknowledges an adaptable way to deal with research problems. It says that there can't be one approach to tackle an issue however a blend of approach can more readily help solve the problem and discover the reality. Pragmatists accept that there can't be a single reality. This paradigm follows both positivism and interpretivism to look for the answers. Thus, this research worldview would propose a mixed-method approach to research in which the researcher has applied both qualitative and quantitative research methods. This investigation is going to be carried out by using a mixed approach i.e. both qualitative and quantitative. The novels were read first and then found out discourse marker conjunction by using corpus tools; making the analysis quantitative. Later, the

findings were further explored by using a theoretical framework thus making the analysis qualitatively.

Sampling:

Sampling can be defined as a particular rule used to choose individuals from a population to be included in the research (Din & Ghani, 2019). It has been noted that "because numerous populations of interest are too enormous to even consider working with legitimately, procedures of statistical sampling have been made to acquire samples from larger populations." (Proctor, 2003:). Therefore, because of the large size of the target population, researchers must choose the option to examine the sample taken from inside the population, generalize his findings and arrive at conclusions about the whole population. The sampling of the data not only makes the size of the population manageable and reduces the costs of the research but also, it helps in finding accurate results by analyzing the data in a more effective way (Dudovskiy, 2006). Dan Brown has written seven novels and the sample included three novels of the writer, selected randomly.

Data Collection

Data collection is the process of gathering and estimating data on variables of interest, in a set up a deliberate style that empowers one to respond to proposed research questions, test hypotheses, and evaluate results(Costa, Plonsky, & Starfield, 2018). The data collection part of the research is common to all fields of study including physical and sociologies, humanities, business, and so forth. While strategies shift by discipline, the emphasis on guaranteeing precise and legitimate collection remains the same throughout. There are two types of data collection methods i.e. secondary data collection and primary data collection method. The former data include the one that has been published in articles, books, journals, newspapers, online portals or magazines, etc. There is a large data available related to your research in these sources regardless of the research area or problem. To increase the level reliability and validity, it is important to select the appropriate set of criteria to select the secondary data. This criteria may include quality of discussions, reliability of the source, author's credentials, date of publication, depth of analysis or its contribution to your research etc. The primary data collection can be further divided into two categories i.e. qualitative and quantitative data collection methods. Qualitative method is related to words and other non- quantifiable elements and the data can be collected through interviews, observation, etc. whereas the Quantitative method is based on mathematical calculations and the data is collected through questionnaires (Dudovskiy, 2006). Three novels of an American writer Dan Brown have been selected. The novels are 'The lost Symbol, Origin and The inferno'. It has been compared them to see how Dan Brown has used conjunctions and how often he has used all four types of conjunctions to make his text cohesive. These results told us how his work was unique and attributed. It can also lead us to authorship identification of his work. These novels were compiled in pdf form from online sources; then processed on AntConc software for retrieval.

Tools

The tools which will be used in the analysis include pdf to txt converter and then corpus tools such as MAT tagger, Antconc (Lawrance A, 2005). In Antconc., we will be using word list and concordance tool.

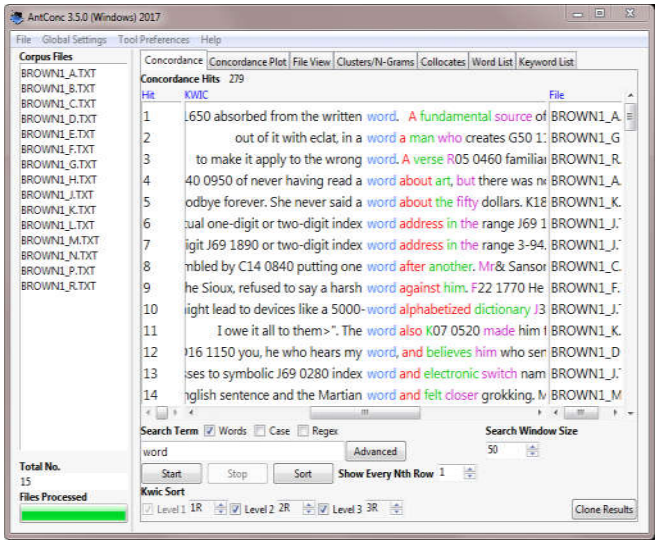

Figure 1. AntConc software

Method

First, the text of novels was downloaded then converted the pdf into txt format using pdf to txt converter. The text was tagged by using MAT tagger. The file generated by MAT tagger was put into Antconc. The specific conjunctions were looked for which needed to be analyzed, by using the concordance tool. Then after finding out the frequencies, the usage of conjunctions was described by using the cohesion model of Halliday and Hasan (1976) and then explained how it could help in authorship attribution of any writer.

ANALYSIS

The analysis was done by using both statistical and thematic operations. The frequency of the conjunctions was found in each novel and then calculate the average use of specific conjunctions and their percentage separately of all the three novels. The statistics of the data was compared and analyzed how he used conjunctions differently within his three novels. Finally, the reasons for the different use of conjunctions by the writer and how it could contribute to his authorship attribution was then explained.

Data analysis:

Cohesion comes in a text through these few things i.e. conjunctions, ellipsis, references, substitution, and lexical cohesion (Halliday and Hasan, 1976) Conjunctions are used as a cohesive device in these novels and most of the conjunctions have been within the text. The analysis of these novels provided us with an overview of how Dan Brown used conjunctions in his work. The additive conjunctions were the ones that occured the most. In total, 9988 additive conjunctions were used which made it more than half of the total number of conjunctions. Moreover the additive conjunction 'and' occured more time than any of the other conjunctions.

Table 1. Statistics of conjunctions of Don Brown novels

Conjunctions	Novel 1	Novel 2	Novel 3
Additive (besides, similarly, and)	3382 (68.53%)	3247 (73.41%)	3359 (69.86%)
Adversative (however, but, instead)	789 (15.98%)	551 (12.45%)	713 (14.82%)
Causal (because, therefore, so)	375 (7.59%)	341 (7.70%)	378 (7.86%)
Temporal (then, finally, at once)	389 (7.88%)	284 (6.42%)	358 (7.44%)
Total	4935	4423	4808

RESULTS AND DISCUSSION

The proper use of conjunctive markers is more important than their number of occurrences. One of the principal variables in English and in writing, according to Halliday (2004), is the presence and absence of conjunctive markers. So in the text, the cohesive items are not the ones that make a text textured, it is the conjunctive markers and their appropriate use. The writing should be highly structured and organized. There should be good relations between sentences and paragraphs through the use of conjunctive markers. The writer has used conjunctions appropriately to make his text cohesive. The analysis shows that he has used all four types of conjunctions even though their frequency varies considerably. The use of additive conjunctions are more than any of the other conjunctions and further in additive conjunction, the use of and has been used frequently and takes up the space of more than half of the total conjunctions used in the novels. The result also shows that 'and' occurs 68.53% in the first novel, 73.41% in the second novel whereas in the third novel it occurs 69.86%. All in all, it covers more than 70% of the additive conjunctions whereas the additive conjunctions like besides and similarly only constitute less than 25% of the text. The rest of the devices' frequencies are much less than the additive ones. For instance, the temporal conjunction's total frequency is 7.27 in all three novels and the same is the case with causal conjunctions. Their frequency is 7.72 i.e. there is no much difference between the occurrences of temporal and causal conjunctions. Although, the adversative conjunction occurs 10.68%. Conjunctions and conjunctive markers are important in novel writing because it provides coherence in the text and they also connect the text through logical connections making it easier for readers to relate to the content and rely on the writer. Good and proper use of conjunctions can make any piece of writing high in texture. The result of the analysis shows that the writer has used conjunctions appropriately to express his ideas and to deliver his message effectively. The conjunctions are used to express agreement and disagreement, maintenance, persuasion, causation, prominence, etc. They have helped the writer to deliver his ideas smoothly and to go from an idea to another without making the reader feel as if they are missing out on something. Such coherence makes the writing real and relatable for the reader. There is no repetition of the words or sentences in the novels because of conjunctive markers. Dan Brown's writing is highly interpretative, reader-friendly, and descriptive.

CONCLUSION

This study aimed to find the authorship attribution by the use of all types of conjunctions in the novels of Dan Brown. It also explains how conjunctive markers can be used to make the writing coherent, textured, and readable. It can be said that the writer uses all types of conjunctions but the frequency of additive conjunctions and the adversative ones are more than any other type of conjunction. This gives us an overview of how Dan brown uses conjunctions in his work and we can find the authorship of his texts through the analysis of the use of conjunctions. Nodoubt, the only finding out the conjunctions and how they are used in the novels are not enough for the authorship identification but it has strong implications as it can help in finding authorship attribution of the writing of any other writer. The study shows that the teachers do not have enough knowledge about authorship attribution (Zaphiris & Ioannou, 2018); teachers can use the information and analysis of the study to think, reconsider and change their course structure that will draw students to interact with other students and their teachers and share and work with pieces of writings. Teachers should see writing as consistently shared and social interaction i.e. authors as constantly negotiating and constructing meaning within and among others. Other than this, teachers can use this phenomenon to identify student's writings and detect plagiarism i.e. student's copying other student's work or copy material from the internet, more easily. The work which is summed up here does not even scratch the surface in terms of understanding the full scope of authorship attribution, of investigating its risks just as its possibilities. It can be concluded that the conjunctions can play an essential role in the authorship attribution of any writer because all writer uses conjunctions in their way to create connections and coherence in their writings.

REFERENCES

Anthony Lawrance. (2005). A Guide to using AntConc. 1–9.http://www.laurenceanthony.net/software/antconc/resources/help_AntConc321_english.pdf

Bahaziq, A. (2016). Cohesive Devices in Written Discourse: A Discourse Analysis of a Student's Essay Writing. English Language Teaching, 9(7), 112. https://doi.org/10.5539/elt.v9n7p112

Corney, M. W. (2003). Analysing E-mail Text Authorship for Forensic Purposes. Queensland University of Technology, (March), 181 p. https://doi.org/10.1.1.14.7427

Costa, P. De, Plonsky, L., & Starfield, S. (2018). The Palgrave Handbook of Applied Linguistics Research Methodology. In The Palgrave Handbook of Applied Linguistics Research Methodology. https://doi.org/10.1057/978-1-137-59900-1

Cresswell, J. (2014). Research Methodology-Qualitative, Quantitative, and Mixed Method. SAGE Publication India PVT Ltd.

Dijk, V. (2008). Text and context : explorations in the semantics and pragmatics of discourse . (Longman linguistics library , 21 .) London : (November), 113–119. https://doi.org/10.1017/S002222670000640X

Din, M., & Ghani, M. (2019). Analyzing Problem-Causing Factors for Pakistani EFL Learners in Translating Present Indefinite and Past Indefinite Tenses From Urdu Into English. English Language Teaching, 12(5), 194. https://doi.org/10.5539/elt.v12n5p194

Dudovskiy, J. (2006). Sampling in Primary Data Collection.

Dulger, O. (2007). DISCOURSE MARKERS IN WRITING Osman DÜLGER *. Selçuk University, Journal of the Institute of Social Sciences, 18.

Fraser, B. (1990). An approach to discourse markers. Journal of Pragmatics, 14(3), 383–398. https://doi.org/10.1016/0378-2166(90)90096-V

Fraser, B. (2017). An approach to discourse markers An Account of Discourse Markers. 2166(October). https://doi.org/10.1016/0378-2166(90)90096-V

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English (English Language Series). Longman, London.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). An Introduction to Functional Grammar. Rutledge, London.

Halliday, M., Mcintosh, A., & Strevens, P. (2008). The Linguistic Sciences and Language Teaching .Longmans ' Linguistic Library . London : (November), 187–190. Juola, P. (2006). Authorship attribution. Foundations and Trends in Information Retrieval, 1(3), 233–334. https://doi.org/10.1561/1500000005https://doi.org/10.1017/S0022226700001171

Juola, P. (2006). Authorship attribution. Foundations and Trends in Information Retrieval, 1(3), 233–334. https://doi.org/10.1561/1500000005

Li, J., Chen, H., & Huang, Z. (2006). A Framework for Authorship Identification of Online Messages : Writing-Style Features and. 57(3), 378–393. https://doi.org/10.1002/asi

Popescu-Belis, A., & Zufferey, S. (2011). Automatic identification of discourse markers in dialogues: An in-depth study of like and

well. Computer Speech and Language, 25(3), 499–518. https://doi.org/10.1016/j.csl.2010.12.001

Proctor, T. (2003). Essentials of marketing research. FT Prentice Hall.

Proctor, T. (2003). Essentials of marketing research. FT Prentice Hall.

Sidorov, G., & Batyrshin, I. (2014). Syntactic N-grams as machine learning features for natural language processing. https://www.researchgate.net/publication/262295081.

Soler-Company, J., & Wanner, L. (2017). On the relevance of syntactic and discourse features for author profiling and identification. 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference, 2, 681–687. https://doi.org/10.18653/v1/e17-2108

Tiryaki, E. N. (2017). Identification of Justification Types and Discourse Markers in Turkish Language Teacher Candidates' Argumentative Texts. Journal of Education and Training Studies, 5(2), 63. https://doi.org/10.11114/jets.v5i2.2033

Zaphiris, P., & Ioannou, A. (2018). Learning and Collaboration Technologies. Springe

****