International Journal of Applied Linguistics & English Literature
ISSN 2200-3592 (Print), ISSN 2200-3452 (Online)
Vol. 4 No. 5; September 2015
Australian International Academic Centre, Australia

Flourishing Creativity & Literacy

Text Variants and First Person Domain in Author Identification: Hermeneutic versus Computerized Methods

Omar A. S. Al-Shabab (Correspondence author)
King Abdullah Institute for Translation and Arabization, Imam University
Al-Malqa 225, Anas Ibn Malik Road, Riyadh 13524, Saudi Arabia
E-mail: osalshabab@gmail.com

Farida H. Baka
King Abdullah Institute for Translation and Arabization, Imam University
Al-Malqa 225, Anas Ibn Malik Road, Riyadh 13524, Saudi Arabia
E-mail: fhbaka@gmail.com

Received: 08-02- 2015 Accepted: 18-04- 2015 Advance Access Published: May 2015
Published: 01-09- 2015 doi:10.7575/aiac.ijalel.v.4n.5p.170 URL: http://dx.doi.org/10.7575/aiac.ijalel.v.4n.5p.170

Abstract

Since a language variety contains shared variants, and since a complete correlation between author and linguistic features is rarely acquired, it is suggested that linguistic features which fall outside the correlational agreement in a variety belong to the author's First Person Domain (FPD). Advances in computerized vocabulary profiling and readability provide useful characterization of features found in Academic English (AE), but they cannot capture the full range of linguistic features in a text. A corpus of about 38 extracts and texts (111.000 words) from local and international authors is analyzed to determine interpersonal and intrapersonal variations. The results show that language variation determines the features of FPD which are crucial for author identification and that computational methods are not adequately sensitive to insure a hundred percent author identification. Therefore, epistemological author identity profile (AIP) is suggested to plot alleged texts against the socio-physical and epistemological parameters of alleging authors.

Keywords: Vocabulary profile, Readability, Syntactic depth, Language variety, Author identification

1. Background

Since a text is normally assumed to be produced by an author, and since texts can be grouped in varieties by using situational parameters and linguistic features they share, it is safe to assume that the study of language variety has theoretical and practical implications for communal and individual use of language, implications which determine what an author observes due to conventions and what he/she can say even within the limits of one variety. The study of lexis, readability, grammar and textuality can highlight the boundary between conventional aspects of a variety, and thus a text in a variety, and individualistic formulation of that text. Aspects of what is found in actual texts and the overlap among texts (intra and inter textual properties) can modify our perception of the notion of text and language variety, especially when studying samples from authors who are not native speakers of the language and who claim to have written texts when textual and circumstantial evidence do not uphold the claim.

Linguistic features of language variety and variation commonly correlated with situational and geographical factors in terms of use, user and settings, all actively interact with (non-individual) parameters, overlooking any traces of a writer voice or author identity. It is rather surprising that although Author Identity (AI)[i] and author attribution of a text, is the focal topic for establishing a "Science of Text" for Dressler (1978) and De Beaugrande and Dressler (1981, 2002), still there is no theoretical recognition for distinctive features of the individual writer or author in terms of voice or identity.

2. Variety Features: the need for Text Variants

Approaches to characterize language variety are as old as Aristotle's *Poetics*; but research in linguistics has resorted to levels of linguistic analysis and situational or rhetorical parameters to identify shared features of a given variety. J. R. Firth expounded the "context of situation", a notion which was later developed by his followers in Britain to determine the demarcation of language variety. Hill (1958) was the first linguist to use the term "register" (Ellis, 1965) by illustrating that "sociological and institutional linguistics are concerned with what TONGUES PEOPLE use under what CONDITIONS" rather than "what PEOPLE (communities and individuals) use as TONGUE under what CONDITIONS" (Ellis, 1965, p. 5). The focus on people takes the discussion to "dialects", while the focus on "conditions" takes the discussion to langue *use*, i.e. variety. However, it was in (1964) that *register* received elaborate treatment in Halliday et al., who suggested that register identification and classification can be conducted in the form of: 1) field of discourse, 2) tenor of discourse and 3) mode of discourse (Halliday, McIntosh and Strevens, 1964, 90-92), notions which received further elaboration in Ure and Ellis (1977), Sinclair (1972) and Sinclair and Coulthard (1975) among others.

Ure and Ellis (1977, pp. 198-201) stipulated "complete" correlation between contextual features, of field, formality, mode and tenor, and linguistic features, for the attainment of a language register (ibid. 201). One may find an approximation of this complete correlation in certain cases like application forms, questionnaires and multiple choice test items, i.e. varieties which exemplify limited choices making a "sublanguage" found in ritualistic language and strict conventional formulae (marriage ceremony for instance). The room for variation is pre-determined, allowing the applicant to be male or female, married, single or divorced, white, black or colored, and the like. Most texts, however, are not so restrictive, and hence allow for choices at different levels from wide range of possibilities, especially at the level of the mental lexicon and grammatical structure (Author, 1986 and 2012).

Despite the power of conventional and regularity forces associated with linguistic use, there remains a considerable room on the continuum of linguistic infinity for personal style, idiosyncrasies and identity, features which need to be described and explained by a text theory. The need for developing an "inner voice" in a writer or in a language learner has been recognized and encouraged (Russell, 1999), and the need for seeking empirical method and evidence for author attribution has valid, and even "ethical" grounds, since questions of identity of author and "text" have direct bearing on questions of academic argumentation, forensic evidence, copyright disputes and literary criticism. But linguistically, the vital and central ground, the object language, the linguistic unit under the spotlight in this search is the "text". In a corpus-based approach, including variety and variation studies, the text is a real physical entity, a processing mechanism and a theoretical frame is at stake. The two approaches of discourse and text linguistics merge in the shift from focusing on discourse features, variety features, or stylistic features to relocating the "text" in the central paradigm for knowledge and linguistic study. Whether one takes the realization of linguistic formulation (paraphrase, abstracting, plagiarizing, cut and paste), translating or investigating language variety, style, or discourse, the unit under scrutiny is the *text* as realization or abstraction.

Text variant is a feature of language user or linguistic use. The term is an all-comprising notion which refers to features of the speaker, the situation, the geographical space, or the speaker-specific human or non-human speaker including divine speakers, imaginary entities and creatures assumed actual or mythological. Therefore, any of the purposes described in the above section (text and variety description, writer identity, author attribution, or studying one specific text) will be best served by focusing on text variants, and consequently questions of text integrity can be captured and served.[ii]

3. Writer Identity from Pedagogical and Ethnographic Perspectives

In pedagogical settings encouraging learners of English as a foreign/second language to express themselves, takes various forms one of which is keeping diaries or writing an autobiography or "autoethnography". Chamcharatsri uses composition classes to examine how second language learners "construct their identities" and how autoethnographical aspects of their personality play a role in identity building by Asian students in American universities (Chamcharatsri, 2009).

For Ivanić learning and teaching about academic writing can evolve around the self and identity, a notion which receives extensive treatment in her (1997) book. Identity involves complex and powerful constructs such as “social identity”, “the self” and “discoursal identity”, helps in developing discourse and academic community (Ivanić, 1997) and it also helps in exploring the “multiple possibilities of self-hood” as an academic writer (Ivanić, 1997). One can say that according to Ivanić’s model, the journey of an academic apprentice writer is tenuous and problematic as having to convey content while giving a representation of the self, in multilayered structures of socio-communal groupings in the wider environment and academia. Ivanić’s model positions the “self” and “identity” in the face of socio-cultural settings which can be penetrated only through discoursal compromises between the self and the other.

Catherine Russell (1999) explores the “autobiographical” and “autoethnographical” as two basic foundations of the construction and representation of identity in filmmaking. The deconstruction of the other through the “transformation of “personal expression” in the avant-garde to a more culturally based theory of identity” (Russell, 1999, p. 25) is the way to vitalize and create the dynamics of identity. This is achieved by harnessing the authobiograpical elements and the autoethnic expression in self-expressing discourse. The “inner voice” expressed by the “I” as opposed to “you” finds its echo in the “mental voice” found in expressions of identity in women intimate relations (Moonwomon-Baird, 2000). In all three perspectives reviewed above, linguistic (or discourse) identity hinges on extra-linguistic factors, biography, community, ethnicity, or libido motivation.

According to Klein and Kirkpatrick (2010), writing can be a tool for “communicating and learning", distinguishing two types of variables, moderator variables “gender, previous writing experience” and mediator variables “genre knowledge and approach to writing”. They found that gender predicted previous writing experience but was not affected by instruction, while instruction affects genre knowledge” (Klein and Kirkpatrick, 2010).

Patchan et al. (2009) studied the writer's identity by comparing comments by students (peer review), a writing instructor and a content instructor, to test the hypothesis which states that students are capable of rating their peers. Writing instructors’ comments were largely evaluative (72%) rather than coaching (20%) or common reading (8%) (Smith 2003 quoted by Patchan et al.). Patchan et al. were mainly concerned with “directive comments” in “feedback … important for improving writing” (Patchan et al. p. 127). For Bruke (2010), Korean students in American universities face difficulties as they engage in a "power struggle" in their attempt to construct "their authoritative identities in the U.S. academy – which requires authoritative writer identity" (Bruke, 2010, p. 13).

4. Author Attribution

Universities in USA prepare students guides to ensure knowledge about violations of copyright and plagiarism offences (see: http://w.w.w.judicial affairs,sa.ucsb.edu/ Academic Integrity, Academic Integrity: A Student's Guide). The implicit assumptions behind such documents reveal that texts and authors have their entity, identity and integrity, terms which provide notional frames that need demarcation, but that also pose questions about the linguistic ground and implications of these terms.

Studies of Author Attribution (AA) have a long history (see Grieve 2005 for a review), but they have recently experienced a surge with the availability of experimental methods enhanced recently by computational linguistics. Grieve puts the earliest date for studying AA as (1787) referring to Edmond Malone work on the three parts of Henry VI in which Malone used meter and rhyme as features of author attribution (Grieve 2005, p. 4). In his review of AA, Grieve discusses the main issues in the area including: meter and rhyme, word length, sentence length, punctuation, contractions, vocabulary richness, graphemes, etymology, errors, words, word position, N-Grams[iii] and syntax (as in Startvik's analysis of forensic evidence: the discrepancies in two written statements, Grieve, 2005, p. 53).

With advances in computer and information technology, one can predict the rush towards using and developing available technologies. This trend is felt in the Internet link for Appen Speech Language Technology Inc. (http://www.appen.com.au). Three short quotations will clarify the point:

Appen Text Attribution Tool (TAT) was developed under US government funding and sponsorship to meet and identify needs of intelligence and law-enforcement organizations.

[*underlining by researchers] (Appen,, 2008a, p. 1)

TAT determines author's age by passing a document's features to a machine classifier (an SVM; SMO as implemented in WEKA [6]). By using features other than surface level, the TAT is able to identify constructs that reveal an author's true age.

[*underlining by researchers] (Appen, 2008a, p. 2)

The TAT is intended to support human analysis by identifying candidate material for more assessment. It is not intended to provide definitive analysis. In its law-enforcement and intelligence configurations, the user brief was to provide an investigative profiling tool than evidentiary tool.

[*underlining by researchers] (Appen, 2008b, p. 2)

The first quotation declares that TAT is customized to serve intelligence and law-enforcement organizations, and that TAT makes the task too specific for a linguist and operates as an indicator motivated by problem-finding.

In the second quotation, the customer is promised that TAT will definitely identify the "construct that reveals an author's true age", a claim which is made using indirect language "identifying constructs" and which is simply not correct, since "true age" can be illusive and the constructs used including "slang/jargon, specialized vocabulary, or context specific language varieties as text spk (SMS text speak)" are all open to abuse and misuse.

The third quotation provides a disclaimer and a retreat from the position announced earlier, since it states now that the results do not provide "definitive analysis" and that "the developers are told of no legal incrimination against individuals who may be wrongly classified and consequently accused" (Appen, 2008, Internet site).

In another online paper (Appen, 200b), Appen reveals another tool, Data Stream Profiling Tool, which "uses biometrics modeling, specifically mathematical abstractions of a user's typing behavior, in order to identify them". One of the three main components, the Keylogger, is a "small software component that is installed (covertly if necessary) on any computer to be monitored" (Appen, 2008b, p. 1). The "covert" option used here is not for protecting the persons being monitored but to put them under surveillance, a matter which raises ethical and legal questions. Technically, the significant factors DSP works include "typing cadence; duration for which keys are held down; timing transitions between key sequences" (Appen, 2008b, p. 1). It is clear that although the linguistic product of the individual being monitored is in the background (i.e. is being processed), the factors being monitored are extra-linguistic factor relevant to non-verbal behavior. The DSP shows the wide range of issues which may be evoked in what is termed "profiling". Author linguistic identity in this paper, is limited to language, and may be most productive when it is limited to one text-type.

Although not all works on text/author identification or attribution are set for specific or narrow band of customers with hidden agenda, still by its nature, unlike pedagogical applications, text identification and author attribution can easily slide to a forensic type of investigation.

5. Author identification

The works reviewed below make a random selection in which the emphasis is on academic questions concerning the possibility of achieving Author Identification (AI) and the type of linguistic features and techniques (tools and methods) employed to achieve AI.

Stamatatos et al. (2001) carried out experiments in genre detection, author identification and author verification tasks to test their method which they developed. Their technique utilizes one-word and two-word frequency. They maintain their method and technique is promising and that the “distributional lexical measures, i.e. functions of vocabulary richness and frequency of occurrence of the most frequent words” is better than most available methods for author identification (Stamatatos et.al, 2001, p. 471).

Hoover (2003) questions the “usefulness of vocabulary richness for authorship attribution” rejecting the assumption that “vocabulary richness can capture an author’s distinctive style or identity” (Hoover 2003, p. 152). But in Hoover (2006), a large corpus (200.000 words) of American poetry and a large corpus of 46 Victorian novels, are used to test the usefulness of “the less than 100 most frequent word units, only to conclude that word lists were unable to identify author or style (Hoover, 2006). Hoover is hopeful, however, that refinement in search measures and the large corpora that can be treated today are promising in enabling us to explain “why and how word frequency analysis is able to capture authorship and style” (Hoover, 2006, p. 1).

The critical issues of the works of suspected authors and number of words needed for suspected texts were discussed by Luyckx (2011). He adopted the “traditional number 10.000 per author as a minimum for an authorial set” (Luyckx 2011, p. 35). He clearly illustrates text categorization models in figure 1:

	Corpus	
Training data	Feature selection	
	Training instances	Machine learning
Test data	Test instances	
		Labeled test instances

Figure 1. Luyckx and Daelemans, 2011, p. 38.

Luycks and Daelmans write:

> It is possible that different types of features (e.g. character n-grams or function word distributions) are reliable for small as well as large sets of authors, the specific features may be very different in both conditions.
>
> (Luycks and Daelmans, 2011, 42)

Considering forensic investigations in general, one can say that they reveal suspicion of text manipulations which compromise text “integrity” and which hide negative malicious intentions. Hence, the degree of offensive text manipulation and the degree of compromising text integrity is matched by human involvement in the manipulation, resulting in at least three types of offence:

1) Criminal persona corresponds to a criminal offense in which a text is forged or adapted as part of a crime.

2) Shadow/Ghost author corresponds to academic offense in which a text is wholly or partially claimed by a person other than the original writer (including cases in which the text is totally or partially commissioned to appear with a name other than that of the original writer).

3) Twilight assistant corresponds to soft offenses in which students or trainees receive assistance and/or lift material from outside sources (including commissioning) to unjustly earn grades, prizes or recognition (but not to the extent of procuring a complete project or thesis).[iv]

Zhao and Zobel's investigation of a literary English corpus of (634) texts by famous authors "to further explore the properties of AA methods" (Zhao and Zobel, 2006), focusing on three linguistic features taken to represent style in the authors under investigation: 1) function words; 2) part-of-speech (pos) tags and pos pairs; 3), and combinations of these (ibid). Their main results show 85% accuracy in positive examples, 95% accuracy in negative examples, 10% accuracy in parts of texts, and 53% accuracy in 10.000 words extracts. Interestingly, the main error (misattribution) originated from translated texts "suggesting that style" – as measured by Zhao and Zobel – "does not survive the translation process." (Zhao and Zobel, 2006, p. 2). Conclusions show that token-level is the most reliable discriminating factor, and that the analysis level measures are more reliable than the phrase-level. Secondly, texts less than 1.000 words are less likely to be correctly classified. Thirdly, according to Stamatatos et.al the method has achieved a higher accuracy than Burrow’s lexical method, which used fifty most frequent words (see Stamatatos et al., 2001. P. 212).

Stylometry is most often used for detection of plagiarism, finding authors of anonymously published texts, for disputed authorship of literature or in criminal investigations within forensic linguistic domain (Stańczyk and Cyran, 2007, p. 151)

Stańczyk, and Cyran (2007) investigate two Polish writers using nine function words, eight punctuation marks, and combinations of function words and punctuation marks. They reported that the “textual descriptors” they used showed a preliminary advantage for using “syntactic attributes” in author attribution (see Stańczyk, and Cyran, 2007, p. 157).

The brief review above outlines three main concerns: 1. Identity development in pedagogical and academic settings represented by Ivanić (1997), 2. Autoethnographical self, represented by Russell (1999), 3. Author attribution and author identification recently represented by numerous researchers (e.g. Grieve, 2005, Grieve, 2005 and Zhao and Zobel, 2006). It is clear from the comments made in relation to each of the reviewed works that the main emphasis is pedagogical, ethnographical or computational, which leaves the role of identity in text interpreting and making unexplained. Hence, placing the notion of identity in the linguistic network in the form of the IF, would hopefully reveal some aspects of interpretation, identity and author identification, by focusing on vocabulary, and readability. At the same time, investigating the author-specific features will address the features which fall outside the scope of use in variety studies and user in sociolinguistic and dialect studies, in addition to testing the potential of available computer programs. Author and text specific features are investigated through a simple experiment reported in the following paragraphs.

6. The Present Experiment

6.1 Author Linguistic Identity: Rationale

In order to construct an author identity (AI) outside the boundaries and concerns of traditional variety analysis of register, text-types and genre, and to test some of the stylometric techniques reviewed above, a well-researched sample is needed. The sample needs to be controlled in various ways including size, variety field and author to guarantee a better approximation in the results obtained and to allow comparisons of texts, or parts of texts, by overtly stated alleging authors and anonymous authors whose works are included for the purpose of shedding light on author identity and text integrity. Putting diverse authors in a list will not help in the search for a "possible author" (Grieve, 2005, p. 87).

6.2 Method and Sample

The primary method utilized here is observation of details of linguistic behaviour at the level of lexis (Nation's vocabprofile), text (readability scales) and syntax (sentence length and clause type). Observation of details and rigorous testing will enable researchers to obtain viable conclusions which can take the discussion beyond mere description and classification.

6.2.1 Sample Size and Diversity

The size of the sample is limited by considerations of availability of text in electronic forms and capacity of computer programs being used such as Paul Nation's Vocabprofiler and Flesch ease score, which added "human interest" to ease of reading in an attempt to supersede earlier formulas (Flesch, 2006).[v] Syntactic analysis was manually conducted, a practice which poses its own constraints on the amount that can be handled, but which is preferred to Xiaofei Lu's computerized syntactic complexity analyzer due to the limitations of Lu's classification of clause types.[vi] The sample is limited to one variety of English, academic English (AE), and within AE, only works from the field of linguistics are included, with one exception in the form of a poem by W. H. Auden for comparison and contrast. The diversity and size of the current sample are specified in Table 1 below.

The sample includes two selections from M.A. dissertations and Ph.D. theses, a sample from the introduction and method, and a sample from the survey of previous works, which allows examining this crucially intertextually mediated part of academic works. There are three types of summaries in the sample: M.A. summaries, Ph. D. summaries and academic papers summaries (column 5-7, Table 1). Also included in the sample are complete texts of academic papers and a selection from 3 books on Linguistics by two authors.

It is hoped that the sample will give results about various aspects of academic English in the field of language and Linguistics.

The works from which the samples are taken are by Arabic speaking academic staff specialized in English language, kept anonymous for privacy and ethical reasons, though the third out of the three (coded THREE) is the current author. Since the other two are anonymous authors (ONE and TWO) known by the author for long period of time, the author has first-hand circumstantial knowledge of authors ONE and TWO, knowledge which can be crucial for the purpose of author identity and author attribution. Works from four well-known linguists, John Sinclair, John Swales, M. A. K. Halliday and Noam Chomsky, are included to act as a yardstick against which other works are measured and compared. The poem from Auden acts as a reminder of the semantic possibilities and potential of the language, and it helps in evaluating various techniques and parameters used, such as lexical density, word frequency, readability and syntactic depth.

Table 1. Number of words and sources of current corpus.

		ONE	TWO	THREE	Swales	Sinclai	Hallida	Chomsk	Auden	Total
MA Introduction		1.000	1.000	955	-	-	-	-	-	2.955
MA Literature		1.000	1.000	1.000	-	-	-	-	-	3.000
PHD Introduction		1.000	1.000	1.000	-	-	-	-	-	3.000
PhD Literature		1.000	1.011	1.000	-	-	-	-	-	3.011
MA Summaries		245	292	146	-	-	-	-	-	683
PhD Summaries		316	764	326	-	-	-	-	-	1.406
Papers/Summarie		755	695	294	368	694	134	282	-	3.222
Papers	P1	5.199	5.808	10.716	2.003	2.001	1.999	10.101	-	37.827
	P2	3.098	9.621	5.406	2.000	2.049	-	10.215	-	32.389
	P3	-	4.427	-	-	-	-	-	-	4.427
Books	B1	-	-	14.509	-	-	-	2.000	-	16.509
	B2	-	-	-	-	-	-	2.001	-	2.001
Poem		-	-	-	-	-	-	-	463	463
Total		13.613	25.618	35.352	4.371	4.74	2.133	24.599	463	110.893

6.2.2 Basic Issues

There is a vast string of topics, issues and concerns which can be addressed by studying the current samples and in light of the results obtained from it. But the main issues of immediate interest here can be stated in terms of priority in the questions below:

1. Comparing one author with another: Are there any significant differences among and/or within the works of various authors in terms of vocabulary, readability, and syntactic depth?
2. Comparing variety with another variety: Are there any significant differences among the varieties (theses, summaries, papers or books) in terms of vocabulary, readability, or syntactic depth?
3. Comparing text with another text by the same author: Are there differences between two or more texts by the same author?
4. Comparing one part of a text with another part in the same text: Are there any differences among the parts of academic works of MAs and PhDs in terms of vocabulary, readability or syntactic depth?

The scope covered by investigating text and author identity is both dynamic and open-ended; whereas the features and the parameters utilized in computer programs are necessarily fixed, and currently limited. Three such programs have been utilized in the current analysis:

1. Paul Nation's Vocabprofiler, which handles various aspects of vocabulary statistics including the features of: a. total tokens, b. total types, c. K1, d. K2, e. word frequency.

2. Readability scales from which the following are obtained: a. number of words, b. number of sentences, c. ease score, d. readability level in terms of (school) grade.

6.3 Vocabulary, Readability and Syntactic Depth in MA and PhD Theses

Are there specific lexical tendencies or lexical behavior reflective of, or bearing the stamp of, a specific author? Can the impressions or pre-theoretical hunches and assumptions about a specific way of talk, a unique print or a "linguistic DNA" be supported by systematic observation and empirical investigation? The answer cannot be given lightly if one remembers the seriousness of the ethical, practical and material implications of cases of false authorship and identity theft, cases which range from text integrity to plagiarism. But the complexity and sensitivity of the questions are not reasons for delaying tackling them, nor should the present shortage in research tools and lack of effective software stop a preliminary attempt at evaluating currently available methods and techniques and suggesting future direction.

To narrow down possible differences in the results, one language variety is examined at a time, starting with academic theses where results from works of three unrevealed authors are reported including the percentage of K2 words (2^{nd} most frequent thousand in English) which shows reasonable similarity except in TWO Ph.D. Literature survey which uses 2.66% almost 50% less than TWO Ph.D. methodology 4.90%, where the two cases are less than 5% reported in ONE and THREE. Using 5.60% K2 words in Ph.D. literature is also attractive, since circumstantial knowledge of ONE and TWO puts ONE as low in writing ability. The notable results of Ph.D. K2 words point to a clear case of disciplinary deficiency compared with the two other authors as well as samples from MA and Ph.D. from the same author, TWO.

Another indicator, type/token ratio, shows similar distribution across the three authors except for TWO Ph.D. literature which is 0.41 compared with 0.37 in ONE and THREE for the same section of the thesis; otherwise, this parameter yields similar results across author comparison.

In readability indicators, the average words in a sentence shows a big difference in TWO MA (27.27 words) and in THREE MA (29.62 words); while ONE MA shows the lowest number of words in a sentence and the highest results in Flesch ease scale, which is not surprising taking into consideration the low writing skill found in ONE. In Readability indicators, Average Words in sentences, Flesch ease score, Flesch-Kincaid Grade level and Readability consensus confirm the weakness of ONE and surprisingly high scores in TWO MA, putting it most difficult and least easy, with highest number of words in sentences followed by THREE MA, Method.

Readability scales use three grammatical indicators: number of words, number of sentences and average number of words in a sentence, which means that they leave important significant syntactic features unrevealed. To carry on with the analysis of sentence length, a number of syntactic parameters have been investigated manually, including:

1. Number of words; 2. Number of sentences in text; 3. Number of clauses; 4. Number of clauses per sentence; 5. Number of main (independent) clauses (α); 6. Number of coordinated clauses (coα); 7. Number of subordinate clauses (β); 8. Number of coordinated to subordinate clauses (coβ); 9. Numbers indicating syntactic depth, calculated from total of β and coβs (cf. Lu). The final parameter of syntactic depth is calculated by the number of successive subordinated clauses, taking a coordinated inside a subordinated clause to be the same level of depth. Depth, together with sentence type, in terms of coordination/subordination, may prove to be indicators that distinguish author and/or text. The mechanism of coordination and recursive subordination are fundamentally different from the size of the Mental Lexicon, type/token ratio and lexical density. Ideally, a thorough description of syntactic complexity would take into consideration the degree of nominal modification and the degree of verb-phrase complexity, to account for depth at the level of the phrase as well as the level of the clause (Author, 1989). Essentially, syntactic complexity at the level of the clause in the present analysis may prove to be informative, and hence may point to distinctive features in an author's written texts; it carries identity features.

The results of examining various aspects of clause types and depth show that "clauses per sentence" is quite promising, since it sets THREE higher than ONE and TWO except in TWO. The higher number of clauses per sentences, has direct bearing on syntactic complexity, which confirms the difference of THREE from ONE & TWO; by using more clauses, especially subordinate clauses, in a sentence. ONE is lowest in depth and THREE is highest. But logical connectors, meta-textual deictics, and organizational elements are shown to be more salient as textuality indicators.

One can conclude that Vocabulary Profiler, readability scales, syntactic complexity and meta-textual connectors, have a measured degree of success in distinguishing one author from another when writing in the same academic variety, e.g. MA and Ph.D. theses. Numerous parameters show different results in the same work when two samples from methodology and literature reviews are examined. These results require further testing and the more sensitive (indicative) parameters closely monitored in varieties other than MA/Ph.D. Theses and works from more authors, is addressed in the following section when academic abstracts are examined.

6.4 Vocabulary, Readability and Syntactic Complexity in Academic Abstracts

The abstracts samples are necessarily small in words number, but they include abstracts of papers by the three anonymous authors and by four linguists, one American and three British. The highest percentage of K2 is found in ONE's papers, while the lowest K2 is in THREE MA and Chomsky's papers. As such, these results do not reflect much about author's vocabulary profile. Type/token ratio does not show any significant tendency; the lowest is ONE's papers (0.39), THREE (0.36), Two's Ph.D. (0.37) and Sinclair's (0.46); the highest Type/token ratio is Halliday's paper (0.65), Chomsky's papers (0.60), TWO's MA (0.55) and THREE's papers (0.55). Type per token reflects this conflicting picture, while lexical density is quite similar across the board, ranging from 0.65 in ONE to 0.53 in Chomsky, leaving only Sinclair's papers lowest 0.48. In all, the vocabulary profile is not telling in this case, which naturally takes the discussion to readability and syntactic depth.

In terms of readability, all four British/American linguists use relativity more words per sentence, worthy of noting is the exception of TWO MA and TWO Papers, where sentence length is higher than that of Swales and Halliday (19.60 and 26.11 words in TWO and 24.67 and 22.67 words in Swales and Halliday respectively); whereas in THREE the number of words per sentence is in line with the rest, 20.86, 21.73 and 23.45 for MA, Ph.D. and papers respectively. Flesch ease score does not correspond to the number of words per sentence in all cases: the easiest 45.9 (ONE Ph.D.) agrees with the lowest number of words per sentence (17.72 in ONE Ph.D.), and the least easy (Chomsky's papers 18.2) corresponds to the highest number of words per sentence (Chomsky's papers 35.12). Still, the full picture gives mixed signals, since in TWO the ease score is 48 M.A. and 20.5 Ph.D. for 19.60 and 21.83 words per sentence (more than double in ease scores compared with only 3 words difference in sentence length in Chomsky). A case seen in THREE, when 44.3 and 20.8 in ease score correspond to less than 2 words difference in sentence length; *less words per sentence is not indicative of ease of readability*. The easiest in terms of Flesch ease score are: TWO MA (48), followed by ONE Ph.D. (45.9) and THREE MA (44.3). The least easy are Chomsky's papers (18.2) followed by TWO Ph.D. (20.5) and THREE PH.D. (20.8), a result which needs further analysis in light of the fact that TWO is being observed for experimental purposes.

Moving to Flesch-Kincaid Grade level, the highest Grades are found in the following abstracts: Chomsky Papers (19.4), TWO Ph.D. (15.9), THREE Ph.D. (15.8), Sinclair Papers (15) and Swales Papers (14.6). The lowest grade levels are: TWO MA (11.4), ONE Ph.D. (11.5) and THREE MA (12.3), showing the greatest discrepancy in TWO (from Grade 11.4 to Grade 15.9); *whereas the four British/American linguists have abstracts assigned for Grade 14.6 and above. Flesch Grade Level is supported by Readability consensus showing Grade 15 and higher for the four British/American linguists,* and Grade 16 for TWO Ph.D. and THREE PH.D., i.e. one grade higher than Swales, Sinclair and Halliday, a result which deserves further treatment.

The results in the Readability consensus offer more moderate, and maybe more credible, grades; Grade level and Gunning Fog text scale tend to assign much higher grades and Colmn-Liau Index tends to assign lowest grades; *while leaving Flesch-Kincaid nearest to the consensus in line 11, being less than half point different from the consensus (19.4 by Flesch-Kincaid and 19 by Consensus). The best indicators of Grade Readability are Readability Consensus and Flesch-Kincai, both of which put TWO Ph.D. quite high in terms of grade adding suspicion to AI of TWO.*

Syntactic analysis of the same sample of abstracts examined above show that readability, which is measured and influenced by sentence length (i.e. number of words per sentence), is not related to depth. In other words, *what is easy to read or suited for higher grade readers in Readability scales is not necessarily characterized by syntactic depth*.

The authors whose abstracts are most complex, in that they show the highest number of clauses per sentence (Halliday papers (7.75), Sinclair papers (4.30) and THREE MA (4.50)), does not correspond to the works (texts) assigned highest grades in the Readability Consensus (Chomsky Papers (19), TWO Ph.D. (16), and THREE PH.D. (16), followed by Swales, Sinclair and Halliday (Grade 15 each). It is surprising that ONE's lowest scores in clause per sentence (2.29, and 2.42 in ONE Papers and ONE PH.D. respectively) should be assigned to reasonably high readability grade (14 and 12 respectively). In spite of this relatively high readability score, ONE PH.D. and ONE Papers, have also the lowest score in syntactic depth (47.05% and 51.27% respectively). The highest scores in syntactic depth, i.e. the most grammatically complex (Halliday Paper (83.87%) followed by THREE MA (77.77%), Swales Papers (68.87%) and Sinclair Papers (67.67%) and TWO MA (61.15%)) are markedly low in Readability Grade: Halliday (15), THREE MA (12), Swales and Sinclair Papers (15 each) and TWO MA (12) respectively. *Thus, there is no relationship between syntactic depth and Readability Consensus*.

In the author Textuality profile, it is noticeable that the same text by the same author exhibits consistent use of textual markers of different types or it consistently lacks textual markers, which means that when textual markers are preferred by an author, they appear in different types even in a small sample, Sinclair Abstracts (See Sinclair Papers, THREE Papers and THREE Ph.D., and TWO MA). *Hence, textual markers like those currently used seem to be promising in author profiling*.

6.5 Vocabulary, Readability Scales, Syntactic Depth and Textual Markers in Academic Papers and Books

6.5.1 Author Profile: Vocabulary

The vocabulary profile in the sample of academic research articles, academic books, and one poem, has the largest number of words in addition to being a more advanced stage on scholarship than in the sample from MAs and Ph.Ds. Therefore, it presents further evidence from published research claimed by the person whose name appears on the research article in journal, conferences and/or the Internet. One notable result is the high percentage of K2 words in one paper by ONE, all three papers in TWO, one paper in THREE and one paper by Sinclair (above 5.8% and 6.30% in TWO).

The poem is markedly different (9.72%) of K2 words. One strange result is found in THREE where in one research paper out of three the percentage of content words is (73.83%) compared with (26.79%) for function words; with the next highest percentages in ONE (40.54%) compared with (36.31%) for function words. This may be explained by the presence of foreign words in the translational data used in that paper by THREE (the current author). *But the overall picture remains rather mixed with no clear trend in the distribution of content-versus-function words.*

Type-token ratio does not show any consistency or special trend; the lowest ratio appears spread across the board: ONE (0.18), TWO (0.17), THREE (0.17 and 0.18), Chomsky (0.17). The highest type/token ratio, however, appears in British/American authors, up to (0.43) in Chomsky Book. But the highest of all type/token ratios is in the poem, recording (0.45) a result which might be influenced by the small number of words in the poem (663 words) compared with 2.000 to 1450 words in works in the research papers sample.

The type/token ratio is inversely related to the number of tokens per type, which is lowest in Auden's poem (1.84) tokens per type. Lexical density seems not to be susceptible to the size of the sample or to author, and hence it ranges in a narrow band between (0.51) in Auden’s poem and (0.64) in ONE Papers, which means that *lexical density is not significant for determining author identity or profile.*

6.5.2 Author Profile: Readability

Moving to Readability designates parameters and scales, one finds that “Average words per sentence” does not correspond to Flesch ease score, as ONE Paper1, Paper2 show, since sentence length (16.57) and (14.00) words per sentence correspond to almost the same ease score of (45.5) and (45.8) respectively. The most striking off-the-point ease score, is that of Auden’s poem where the average number of words per sentence is (27.35) and the ease score is (62.1), which is, supposedly, the easiest of all authors and texts. This result is both counter intuitive and not correct, since even experts on literature face difficulties in interpreting the poem as seen in the Author’s work on contextualization (forthcoming). Another case which deserves some comment in relation to the average number of words per sentence and ease score, is Swales Paper1 in which the average number of words per sentence is (25.81) and the Flesch ease score is as low as (27.7), which is related to sentence length, but contradicts Auden’s poem where the number of words per sentence is very near to that of Swales (27.35) but the ease score is very high (62.1) as observed above. Flesch-Kincaid Grade level shows a similar trend assigning Swales Paper1 to grade (17) and Auden’s poem to grade (11.4), a trend which is carried over to the Readability score of grade (17) for Swales Paper 1 and grade (11) for Auden’s poem, which is put at the same level seen in ONE Paper1 (Grade 11). *The problem is surely not with the poem being so easy, but with the readability indicator which seems to be made for texts types that do not belong to literary genres.*

6.5.3 Author Profile: Syntactic Depth

Syntactic features show discrepancies among the works of the same author in terms of clause per sentence; Sinclair’s three papers have (6.15, 4.58 and 9.09) clauses per sentence; whereas Auden’s poem has a very moderate and comparable number of clauses per sentence (4.61). The number of clauses coordinated to main clauses (co α) is markedly high in Auden’s poem (46.66%) compared with (08.59%) in Sinclair Paper2 and less than this in the rest of the sample. Naturally, the high percentage of coordinated clauses lead to a low score in syntactic depth (31.66%) in Auden’s poem compared with (64.16%) and more in the rest of the sample, which shows mixed scores and relatively narrow scope of difference (64.16%) lowest and (82.71%) highest. In brief, *there is no clear-cut trend or differences in syntactic depth among various works by the various authors, with the significant exception of Auden’s poem.*

Textual indicators reflect a similar message of mixed usage with no clear predictable trend attached to an author or a text. This leaves us with more questions about the various parameters of vocabulary, readability, syntactic and textual parameters used in the present study (Author forthcoming, on intertextuality). One such question may be posed about the performance of the parameters in relation to the academic text-type of writing academic theses, abstracts of research articles and research articles, but a thorough author vocabulary profile needs to map up the full range of ML manipulated in the written works of an author, which in the present case should ideally include in the sample all texts by one author. The ultimate purpose is to specify the individual ML and the Communal ML, which means in the present context the ML shared by all authors in the sample (Author forthcoming). The aspects of linguistic author identity covered in the current paper leave much to be done at various levels of the language of one individual.

6.6 Vocabulary, Readability, Grammatical Depth and Textual Markers

6.6.1 Text-type Vocabulary Profile:

Examining the percentage of K2 in the three academic text-types of theses, abstracts and articles/books, one finds slight differences among the three text-types, with more differences with the works of the same author. To obtain a reasonably better point of comparison of K2 results, one needs to shift attention to a completely different variety of English, e.g. poetry. In a short poem by Auden, one finds the percentage of K2 words to be about double the average found in the three academic text-types: (9.72%) for Auden's poem compared with (2.80%) in Swales' Papers and a noticeably high percentage of (6.30%) in TWO Paper3.

Type/token ratio show minor differences among the three academic text-types in the works by ONE, TWO and THREE. With the exception of Halliday's abstract (0.65) and Halliday's paper (0.33) no big difference is observed, even in Auden's poem whose ratio (0.45) is slightly higher than the rest (but the number of words in the poem is only (463) words, which may influence the Type/token ratio. Even lexical density does not distinguish the poem or any of the three academic varieties: (0.51) for the poem and a range between (0.64) and (0.49) for the three academic text-types. The vocabulary profile has not shown the three academic text/types to be characterized differently; Auden's poem exemplifies a non-academic variety.

6.6.2 Text-type Readability Profile:

Flesch ease Scale reveals that ONE Paper Abstracts are markedly less easy (i.e. more difficult) than the texts and abstracts of other text-types and texts including abstracts, by the same author, a fact which calls for more investigation of this particular text. In fact, most abstracts are less easy (Flesch ease Score) in the abstract compared with, which seems reasonable, since the abstract is more condensed and focused than the body of the text.

Comparing the results from *Flesch-Kincaid Grade level* does not assign markedly higher grades to ONE Paper in the abstract compared with markedly high grades assigned to other authors. But again among the works of the same author, grades vary remarkably in the same text-type, e.g. TWO Abstracts are assigned to grades 11.4, 15.9 and 14.9, and TWO Papers are assigned to grades 13.7, 12.1 and 10.6. One noticeable result is Chomsky Abstract, assigned to grade (19.4). The Readability Consensus shows remarkable differences among the authors; but relatively higher grades are assigned to abstracts compared with papers with the exception of Swales and Sinclair when the papers are assigned to grades (16/17 and 16/14) and the abstracts are assigned to grades (15) in both cases. *The overall picture is mixed and far from showing a tendency of conformity.*

6.6.3 Text-type: Grammar and Syntactic Depth Profile

Clause per sentence in ONE theses and Abstracts are similar, but in ONE Papers it is slightly higher, whereas in TWO and THREE Theses have more clauses per sentence than abstracts and papers. *The number of clause per sentence in the three academic text-types is mixed and does not show a clear tendency*.

The number of subordinate clauses shows no significant differences among the three text-types except in Halliday Abstract (only 125 words) which has (83.84%) of subordinated clauses, the next linguist is well below that (62.96%) in THREE MA Abstract. The more indicative measurement of syntactic depth does not reveal marked differences among the three academic text-types, again with the exception of Halliday Abstract (83.87%) compared with (71.51%) for Halliday Paper. Sinclair Paper3 has (82.71%) in syntactic depth and only (67.67%) for his abstracts, which is the opposite trend of Halliday's short sample. As for textual markers, simple observation of the results shows that the abstracts use less textual connectors, whereas all three text-types scarcely use meta-textual organizational markers. *In conclusion, it can be said that the three profiles of vocabulary, readability and syntax, do not characterize any of the three academic text-types, stamping them as consistently and significantly different.*

To specify the size of the ML, Communal and Individual, a case-specific software needs to be developed, and that would lead to the creation of a comprehensive vocabulary profile which maps up all the known/available corpus of a given author.

7. Conclusion

The search for author-specific features has, on the whole, been approached primarily from the vantage point of surveying the total linguistic features found in a text, which calls for more scrutiny. The first observation comes from the fact that in variety studies in which it is acknowledged that each variety, regardless of the term used to designate a variety, is characterized by variety-specific features. The second basic notion relates to the fact that in variation studies, social and geographical factors, whether gender, dialect or jargon, show variation-specific features. If these two types of features found in a text are put together they may amount to a good percentage of the total linguistic features in a text, depending naturally on the degree of correlation and development of the text-type in language under analysis. This means that the FPD features will necessarily belong to the non-variety and non-variation features, a group which can be labeled non-communal factors, among which the author-specific, or FPD, features constitute a prominent component. Author identification should establish these features in the alleging author, and then should move to investigate them in the suspected or alleged text. In the absence of the techniques and data required for isolating the author-specific features in a text, circumstantial evidence can be considered. The best candidate for describing author identity profile is to establish a method for constructing and reconstructing aspects of author profile which will help map up the alleged text against the detailed profile of the alleging author. The following components can be suggested:

Table 2. Author Profile based on Author–specific features.

Elements	Social/physical parameters	Existential parameters	Epistemological mental parameters
Being	⁄	⁄	⁄
Environment	⁄	⁄	—
Understanding	—	⁄	⁄
Experience	⁄	⁄	⁄
Assertion	—	⁄	⁄
Identity	⁄	⁄	⁄

In a comprehensive treatment of author identity profile (AIP), Author states some criteria for applying the above configurations:

The reconstruction of the mental-epistemological identity should be congruent with the social/physical parameters of identity. The relationship between the two is judged against the criteria of: 1. Consistency (to eliminate mismatch), 2. Plausibility, 3. Ethical code, 4. Claimed pronouncements (texts). Each of these criteria may prove to be necessary in certain cases, can be more or less relevant depending on the case, i.e. author and text. (Author, forthcoming)

For a possible application, a simple profile is attempted for the British linguist, M. A. K. Halliday, and the American linguist, Noam Chomsky by Author (forthcoming).

However, if we leave elaborating on the AIP for a future work, the results reviewed in details above enable us to make a number of remarks by way of conclusion and observation concerning what is needed to reach a more decisive position.

1. The results, except for Auden's poem, do not give evidence of coherent consistent features which mark an author with distinctive systematic set of features that justify a "whole" profile typical of a specific author.
2. The parameters used in the study of vocabulary, readability, structure and textuality do not allow any of the three academic varieties of theses, abstracts, papers and books to be assigned variety-specific features.
3. Considering one and two above and taking the differences at various levels in the three academic text-types, it is logical to assume that a more viable approach to author identification and attribution, must take text as the departure point in matters related to author identification and author profiling, and the text, rather than the author or the text-type, as a linguistic unit can be the focus.
4. Auden's poem included with academic articles and books, manifests meaningful distinctiveness at certain, but not all, levels of analysis.
5. In case of anonymous texts, the open question of "Who is the author of X?" is far more difficult to address than the narrower question of "Can X be written by Y?", where X is a specific text and Y is a possible (claiming or ghost) author. In other words, with reference to a particular author, it will be easier to determine if "X is or is not written by Y" than to determine if "Z is the author of X" (where X is a given text, Y a claiming author and Z any possible author). Unlike most cases in forensic linguistics where the speaker/writer is unknown, certain cases of attributing academic text (not author attribution) can be narrowed down, and subsequently resolved, by applying the "X" to this author, which means that in certain cases of academic texts the question about authorship can be formulated in the following question: "Can this individual (claiming author) have written this text he/she attributes to him/herself?"

The above question can best be answered by comparing claimed text(s) with a sample of actual writing by the claiming author, comparing various claimed texts in the author profile for consistency and for present or absent fingerprints, comparing claimed texts with standard features of corpus representing the text-type to which the claimed text(s) belong. In the absence of circumstantial evidence, and in the absence of sample texts written by the "suspected" author, it is difficult to be certain about author identity when the available texts are all in doubt.

It is more difficult to determine authorship when the author is nonnative speaker and does not have a text known to have been actually produced by him/her. For the purpose of author identification (AI), text-type, text, and author should be viewed as living dynamic beings in the making, rather than a fixed definitive entity. AI should incorporate this fact about text and author, and *not* deny or fight them by taking these two notions to be static.

References

Author (1986). *Organizational and Textual Structuring of Radionews Discourse in English and Arabic.* (Unpublished doctoral thesis). Aston University, UK.

Author (2012). First Person Domain: Threshold Mental Lexicon and Arab Learners of English. *Proceedings of the Second Symposium on English Language Teaching in KSA: Realities and Challenges: Research Papers,* Riyadh, Saudi Arabia (9-11 April, 2012), 141-208.

Author (forthcoming). Text Integrity, Editorial Practices and Author Epistemological Profile. In Author. *New Horizons in Linguistic Interpretation: Poetry Translation.* Manuscript.

Appen Speech and Language Technology Inc. (2008). Internet Safety Technical Task Force: Technology Submission – Text Attribution Tool. Retrieved from http://www,appen.com.au

Aristotle (written 350 B.C., Translated by S. H. Butcher). *Poetics*. Retrieved from http://classics.mit.edu/Aristotle/ poetics.html

Academic Integrity. Retrieved from https://www.google.com.sa/search?biw=1307&bih= 342&q=asu+academic+integrity+resource+guide&oq=ASU+Academic+Integrity+Resource+Guide&gs_l=serp.1.0.35i39.15784.19768.0.24396.12.8.0.0.0.0.263.1111.0j2j3.5.0.msedr...0...1c.1.62.serp..8.4.887.0.nXOemal80fo

Baka, F. (1989). *The Discourse of Biology Lectures: Aspects of its Mode and Text Structure.* Ph. D. thesis, Aston University in Birmingham, UK.

Bruke, S. B. (2010). *The Construction of Writer Identity in the Academic Writing of Korean ESL Students: A Qualitative Study of Korean Students in the US.* (Unpublished doctoral dissertation) Indiana University Pennsylvania.

Chamcharatsri, P. B (2009). Negotiating Identity from Auto-ethnography: Second Language Writers' Perspective. *The Asian EFL Journal: Professional Teaching Articles*, 38, 3-19.

De Beaugrande, R. and Dressler, W. (1981). *An Introduction to Text Linguistics (*digital 2002). London, Longman.

Dressler, W. V. (1978): *Current Trends in Textlinguistics*, Berlin & New York, Walter de Gruyter.

Ellis, J. (1965). Linguistic Sociology and Institutional Linguistics. *Linguistics*. 3, 19, 5–20.

Flesch, R (online). *FLESCH READING EASE READABILITY FORMULA*. Retrieved from http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.phpFlesch, R. (2006). A New Readability Yardstick. *The Classic Readability Studies*, Costa Mesa CA, Impact Information, 96-111. Retrieved from http://www.ecy.wa.gov/quality/plaintalk/ resources/classics.pdf

Flesh, R (1948). A New Readability Yardstick. *Journal of Applied Psychology.* 32 (3), 221-223. Retrieved from http://psycnet.apa.org/journals/apl/32/3/

Grieve, J. W. (2005). *Quantitative Authorship Attribution: A History and an Evaluation of Techniques*. Unpublished M.A. thesis, Simon Fraser University, USA.

Halliday, M. A. K., Mcintosh, A. and Strevens, P. (1964). *The Linguistic Sciences and Language Teaching* (Longmans' Linguistic Library). London, Longman.

Hill. T. (1958). Institutional Linguistics. *Orbis,* 7, 441-455. Retrieved from http://www. degruyter.com/dg/viewarticle/j$002fling.1965.3.issue-19$002fling.1965.3.19.5$002fling.1965 .3.19.5.xml;jsessionid= 8E93A3F08 AA99E4F5B5E1BA367FE422B

Hoover, D. L. (2003). Another Perspective on Vocabulary Richness. *Computing in the Humanities,* 37, 151-178 Kluwer Academic Publishers, Netherland. Retrieved from http://link.springer. com/article/10.1023%2FA%3A1022673822140

Hoover, D. L. (2006). Word Frequency and Keyword Extraction. *HRC ICI Method Network*, Centre for Computing in the Humanities, Kay House, London, 1-8.

Ivanić, R. (1997) Writing and Identity. *Discoursal Construction of Identity in Academic Writing,* Amsterdam: John Benjamin Publishing, (Chapters, 6 & 10).

Klein, P. D., & Kirkpatrick, L. C. (2010). A framework for content area writing: Mediators and moderators. *Journal of Writing Research, 2* (1), 1-46. Retrieved from http://www. jowr.org/articles/vol2_1/jowr_2010_vol2_nr1_klein_kirkpatrick.pdf

Lu, X. (2010). Automatic Analysis of Syntactic Complexity in Second Language Writing. *International Journal of Corpus Linguistics*, 15(4), 474-496. Retrieved from http://www.personal.psu.edu/faculty/x/x/xxl13/papers/Lu_inpress_ijcl.pdf

Luyckx, K. and Daelemans, W. (2011). The Effect of Author Set Size a Data Size in Authorship Attribution. *Literary and linguistic Computing*, 26, (1), 35-55

Moonwomon-Baird, B. (2002). What do Lesbians do in the daytime? *Journal of Sociolinguistics*, 4, (3), 348-378. Retrieved from onlinelibrary.wiley.com/doi/10.1111/1467-9481.00120

Patchan, M. M., Charney, D. and Schunn, C. D. (2009). A Validation Study of Students' End Comments: Comparing Comments by Students, A Writing Instructor, and Content Instructor. *Journal of Writing Research*, 1 (2), 124-152. Retrieved from http://www.lrdc.pitt.edu/schunn/ research/papers/JoWR_2009_vol1_nr2_Patchan_et_al.pdf

Russell, C. (1999). *Experimental Ethnography: The Work of Film in the Age of Video.* Durham (NC), Duke University press.

Sinclair, J. McH. (1972). *A Course in Spoken English: Grammar*. Oxford, Oxford University Press.

Sinclair, J. M and Coulthard, M. (1975). *Towards an Analysis of Discourse: The English Used by Teachers and Pupils.* Oxford, Oxford University Press.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2001). Automatic Text Categorization in Term of Genre and Author. *Association for Computational Linguistics,* 26, (4), 471-495. Retrieved from http://delivery.acm.org/10.1145/980000/971883/p471-stamatatos.pdf?ip=62.120.162.148&id

=971883&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144 511F3437&CFID=502443111&CFTOKEN=45652212&__acm__=1429224979_d0e31cd59adbdabb650ad3f9bdef67f4

Stańczyk, U. and Cyran, K. A. (2007). Machine Learning Approach to Authorship Attribution of Literary Texts. *International Journal of Applied Mathematics and Informatics*, 1, (4), 151-158. Retrieved from http://www.naun.org/main/UPress/ami/ami-22.pdf

Ure, J. and Ellis, J. (1977). Register in Descriptive Linguistics and Linguistic Sociology. Ed. by Oscar Uribe-Villegas. *Issues in Sociolinguistics,* The Hague, Paris, New York: Mouton Publishers, 197-243.

Zhao, Y. and Zobel, J. (2006). Searching with Style: Authorship Attribution in Classic Literature. *Twenty-Ninth Australian Computer Science Conference (ACSC)*, *Conferences in Research and Practice in Information Technology (CRPIT)*, 48. V. Estivill-Castro and G. Dobbie, Eds. Retrieved from http://goanna.cs.rmit.edu.au/~jz/fulltext/acsc07yz.pdf

Notes:

[I] The term Author Identity is preferred to writer or speaker identity, because the question at hand has to do firstly with the fact that a writer is not a speaker (mode difference), and secondly, because the authorship claims implied in the word author, since a writer is not necessarily and author.

[ii] For the discussion of text integrity see Author forthcoming.

[iii] N-Grams is a method of measuring letters and space configurations in terms of their actual occurrences in a text, a method based on dissimilarity formula demonstrating relative success in text/author attribution.

[iv] But there is a forth case in which textual influences are paramount and are quite accepted, or at least left as a gray area, a case seen in examples of interference from editorial boards and editors in general, in changes implemented by language editors, and in various kinds of amendments and rewriting resulting from “advice and opinions” of academic supervisors and thesis examiners. In all these and similar cases, inter-textual manipulations are accepted, but the name of the editor, reviewer, secretary or supervisor, may well not appear on the work.

[v] Flesch’s paper, which originally appeared in (1948) in the *Journal of Applied Psychology*, 32 (3), is introduced and evaluated in its proper context of readability formulas by W. H. DuBay (Ed.) *The Classic Readability Studies*. Retrieved from

[vi] The Lu’s parameters included clause types but it does not indicate the level at which each type appears in a sentence, i.e. it does not sensitive to the level of embeddedness ordepth. See http://aihaiyang.com/synlex/ syntactic/http://www.personal.psu.edu/ xxl13/downloads/ l2sca.html