NWU®

NORTH-WEST UNIVERSITY
NOORDWES-UNIVERSITEIT
YUNIBESITI YA BOKONE-BOPHIRIMA

Towards valid linguistic measurement: what digital humanities can bring to the forensic linguistic table and vice versa

SADiLaR Digital Humanities Colloquium

13 October 2021 at 10:00 AM (SAST)

Karien van den Berg

North-West University

South Africa

karien.hattingh@nwu.ac.za

Introduction

- The issue: one of authorship
- The question: Did the accused write the denied texts or not?
- The approach: a combination of stylistic and stylometric analysis, as implemented by Kotzé in *State v Kerr Hoho* (Kotzé & Bezuidenhout, 2007; Kotzé, 2010).
- The purpose: demonstrating how DH methodology can supports the validation of authorship verification methodology

Forensic Linguistics concerns:

Broadly:

- -the written language of the law; i.e. the meaning of legislation and legal jargon)
- -spoken interaction in legal contexts; in terms of legal processes, such as having your rights read to you upon being arrested; implementing fair language policies etc)
- -language as evidence; the linguist as an expert presents evidence on questions about language use, e.g. for the purpose of authorship comparrison

The narrow definition restricts the discipline to language as evidence alone (Coulthard & Johnson, 2010)

The nature of authorship verification

- "Even within a single genre, textual features that work well to differentiate author A from a set of peers, might fail to separate author B from the same set of peers. Due to the many idiosyncracies that occur in an individual’s writing style, this makes it challenging to develop systems that can be robustly scaled across many different individuals. Modeling authorial writing style requires bespoke models that are tailored to the characteristics of a single author or a specific set of authors" (Kestemont et al., 2021)

Forensic Linguist as expert witness

- The expert witness must be called to give evidence on matters calling for specialised skill or knowledge.
- The expertise of the witness should not be overstated to such an extent that the court's own capabilities and responsibilities are disregarded.
- The witness must be a qualified expert with sufficient skill or expertise.
- The facts upon which the expert opinion is based must be proved by admissible evidence and must not be based on hypothetical scenarios.
- The guidance offered by the expert must be sufficiently relevant to the matter in issue which is to be determined by the court.
- Opinion evidence must not usurp the function of the court. A witness should not be permitted to give an opinion on legal matters and must not be called to answer questions which the court has to decide – this is sometimes referred to as “the ultimate issue”- doctrine.

Research question

- How can the (applied) linguist present forensic evidence that is sound, fair and based on the valid measurement of idiolectal idiosyncracies?

▶

Points of departure

1. Forensic linguistics is heavily linguistically oriented but is closely related to Law and legal practice.
2. Like DH, Forensic linguistics is essentially interdisciplinary
3. Unlike DH, Forensic Linguistics is a field of enquiry. From my perspective DH is a methodological commons.

Figure 1: An intellectual and disciplinary map of humanities computing (Willard McCarty and Harold Short, via the HUMANIST discussion group. 2006. Alliance of Digital Humanities Organizations).

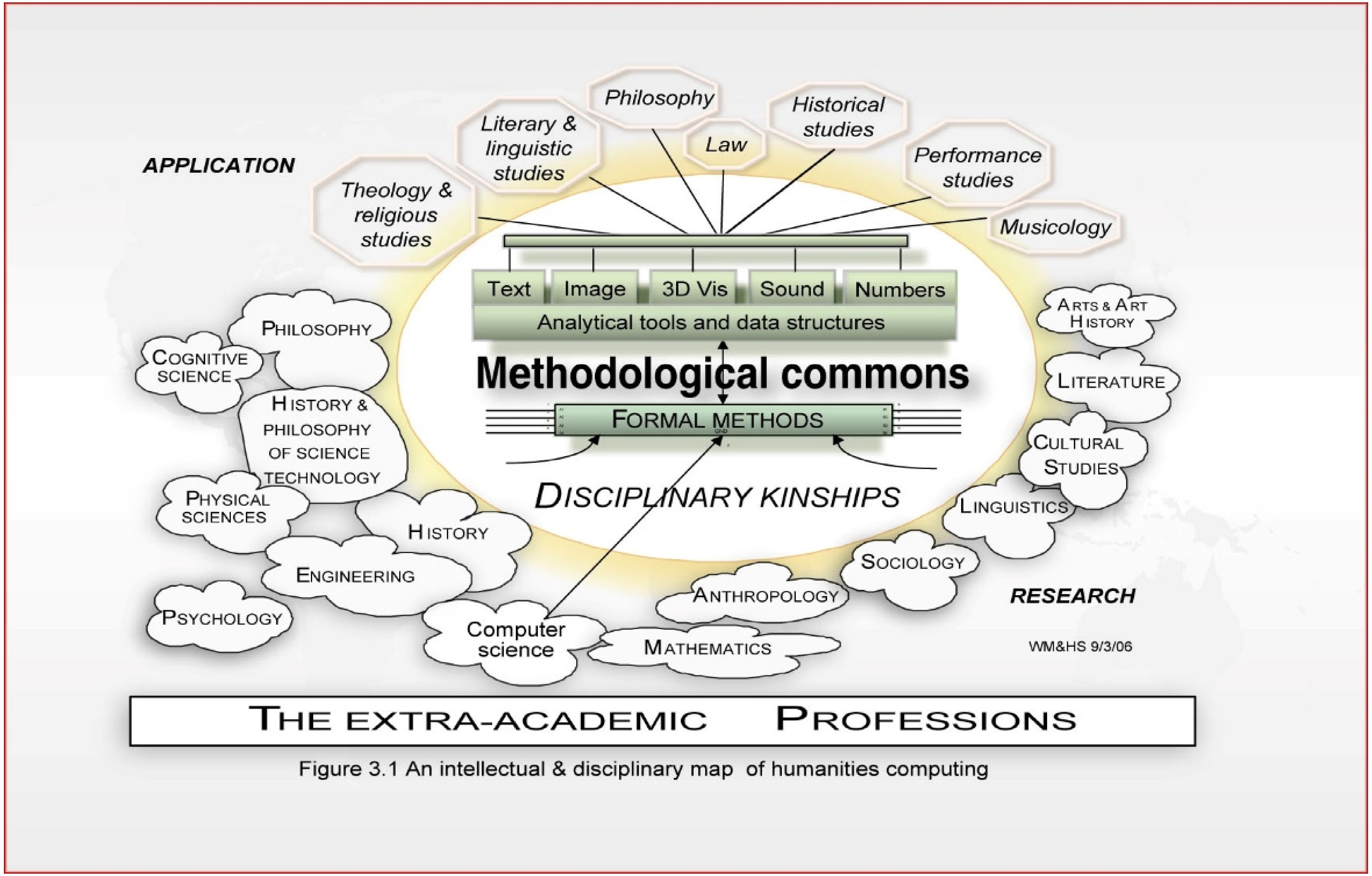


Figure 3.1 An intellectual & disciplinary map of humanities computing

A perspective on Digital Humanities

“Since approximately 1990, a wide range of computational technologies which previously went their separate ways started to cohere into a community of practices, with changing names: *Computing in the Humanities, Humanities Computing*, and most recently *Digital Humanities* or DH. As a consequence, many who worked in fields as diverse as computational linguistics, natural language technologies, computational lexicology, concordancing for theology or law, computational musicology and electronicmusic, digital media production and archiving, documentary linguistics and ethnology, found themselves under this new umbrella. The domains of investigation of DH are extremely heterogeneous,and the methodologies are many, but the different domains have a family resemblance in terms of the methods applied to different domains. In numerous lectures, Harold Short has termed this criterion for coherence ‘metamethodology’ within a ‘Methodological Commons’. In the present contribution, the shared goals of Digital Humanities are seen as the use of computational technologies with formal numerical and structural models to document, archive and describe fields which have traditionally used hermeneutic, interpretative methods.” (Gobbin, 2019)

Points of departure (continued)

4. Authorship attribution is a text-based task.

5. Authorship verification is never a simple task.

6. The notion of an idiolect informs authorship verification tasks.

"much of the research in present-day computational authorship identification is implicitly underpinned by a basic assumption that could be summarized as the “Stylome Hypothesis”. This hypothesis, seminally formulated by van Halteren et al. [1], states that all writing individuals would leave a unique stylistic and linguistic “fingerprint” in their work, i.e., a set of stable empirical characteristics that can be extracted from and identified in a large-enough writing sample. In the analogy of the human genome, the assumption is that this fingerprint is a sufficient means to identifying the author of any given writing sample, provided it is long enough. The Stylome Hypothesis is an attractive working hypothesis, but remains hard to demonstrate, let alone prove“ (Kestemont et al., 2021).

Lakoff vs Fitzgerald on idiolectal co-selection

Lakoff singled out 12 words and phrases for particular criticism, on the grounds that they were items that could be expected to occur in any text that, like these two, was arguing a case – *at any rate, clearly, gotten, in practice, moreover, more or less, on the other hand, presumably, propaganda, thereabouts*, and words derived from the roots *argu** and *propos**.

-3 million documents

-69 documents

(Coulthard & Johnson, 2010:162-163)

Point of departure (continued)

- 7. The duty of the forensic linguist in forensic investigation is to "see what might not be evident to the naked eye" (Correa, 2013:7)

[L]inguists know what to listen for in a conversation [or look for in texts]. They listen for topic initiations, topic recycling, response strategies, interruption patterns, intonation markers, pause lengths, speech event structure, speech acts, inferencing, ambiguity resolution, transcript accuracy, and many other things. Scientific training enables linguists to categorize structures that are alike and to compare or contrast structures that are not. Linguists understand the significance of context in the search for meaning in a conversation and are unwilling to agree with interpretations wrenched from context by either the prosecution or the defense (Shuy, 1993, p. xviii).

Points of departure (continued)

8. DH offers a methodological commons that has the potential of expanding the forensic linguistic toolbox, of amplifying linguistic exploration, description and analysis.

9. The use of various tools can increase reliability and validity of authorship verification methods

  - Language test developers collect evidence to support claims of validity in the form of:
    - -complex **quantitative** (reliability coefficients; factor analysis; multi-faceted Rasch measurement)
    - - and **qualitative** data (expert opinion, discourse and conversation analysis, observation and verbal protocol analysis – cf. Lazaraton & Taylor, 2007)

Case Study: Authorship Verification

- Kotzé in *State v Kerr Hoho* (Kotzé & Bezuidenhout, 2007; Kotzé, 2010)
    - Validation
    - Demonstration of DH

The facts

- "ADMIT": the baseline data set containing a set of emails exchanged between the accused and a legal practitioners. I was given to understand that these texts were written by the author, without any doubt and this set therefore formed the basis of comparison with the second set.
- "DENIED": the comparative data set, containing a series of blog inscriptions, suspected to have been produced by the accused. However, the suspected author denies having written these blog publications.
- 3 subsets of texts written in similar vein on the same topic, but in the form of either informal or more formal newspaper articles.

Data

Current Research

Case #1: Defamation

- *Admit* (3614 words) vs *Denied* (5674 words)
  - Subset 1) 1 8296 words;
  - Subset 2) 2178 words;
  - *Subset 3) 3 x Various Authors* individually comprise fewer than 1000 words per text.
- **Methodological components**
  - Stylometric analysis (quantitative)
  - Stylistic analysis (qualitative)

Comparative Research (cf. Kotzé, 2010)

Father Punch: **Defamation**

- Memoranda *(*5 482 words)
- 11 chronicles (25 431 words)

- **Methodological components**
  - Stylometric analysis (quantitative)
  - Stylistic analysis (qualitative)

The question Case #1:
Did the accused write *DENIED & Subsets 1-3?*

For the analyses, I departed from the following general hypothesis:

- H1: The contested documents (DENIED; Sub-sets 1-3) display a variety of linguistic similarities rather than differences with the uncontested or ADMIT documents, suggesting that these texts are likely to have been written by the same author;
- H0: The contested documents (DENIED; Sub-sets 1-3) do not display a variety of linguistic similarities rather than differences with the ADMIT documents, suggesting that these texts are less likely to have been written by the same person.

As point of departure, I aimed to explore whether the null-hypothesis could easily be accepted. If not, this would suggest that H1 to be more plausible.

Case #1 Stylometric Analysis

Instrument

- Wordsmith Tools (6.0) Word list, Keyword and Concordance function

Findings:

- Overall, the results suggest that it is highly unlikely that different authors composed the documents in question.
    - Interpreted in consideration of constraints such as text length and genre restrictions
    - Results for some *individual* documents contained in the Sub-sets 2 and 3 set offer some contestable evidence in this regard.
- Further investigation is required in the form of stylistic consideration.

TEXTS COMPARED	TOTAL NUMBER OF KEYWORDS ABOVE THRESHOLD OF SIGNIFICANCE	TOTAL KEYWORD AVERAGE	TOTAL KEYWORD MEAN	STATISTICALLY SIGNIFICANT DIFFERENCE: OVERALL?	NUMBER OF GRAMMATICAL ITEMS	GRAMMATICAL KEYWORD AVERAGE	GRAMMATICAL KEYWORD MEAN	10 GRAMMATICAL KEYWORDS AROUND THE MEAN	STATISTICALLY SIGNIFICANT DIFFERENCE: GRAMMATICAL ITEMS?	TEXT LENGTH IN WORDS
Suspected texts										
Denied	58	1,27	1,4	no	27	-1,14	-0,99	-0,05	NO	
Sub-set 1									NO	
	59	5,61	5,37	no	17	-0,42	1,68	5,06	NO	2178
	41	10,83	9,52	no	8	11,73	11,25	10,91	no	
	7	19,19	10,97	no	2	10,49	10,49	10,97	no	104
	20	13,39	10,93	no	6	8,43	7,18	8,43	no	479
	40	10,05	8,42	no	13	8,79	8,45	8,06	no	1567
Sub-set 2									NO	
	76	1,48	1,25	no	29	-0,16	-0,54	-1,78	NO	8296
	9	25,31	26,23	yes	1	24,11	24,11	26,23	yes	191
	53	11,58	10,66	no	19	11,08	9,54	10,91	no	1861
	28	9,82	9,75	no	9	9,1	8,18	9,04	no	542
	37	13,02	12,03	no	13	12,58	10,93	9,32	no	993
	42	11,77	11,19	no	17	10,25	9,23	8,79	no	1424
	33	15,44	12,41	no	9	10,04	10,91	10,07	no	1316
	18	18,63	16,14	yes	8	14,48	11,6	14,12	no	554
	30	20,94	17,86	yes	10	12,4	9,87	9,87	no	777
	20	28,53	25,56	yes	6	24,14	23,79	23,86	yes	233
	9	31,3	21,84	yes	2	22,63	22,63	21,84	yes	165
Sub-set 3									NO	
	33	18,96	15,27	yes	9	14,35	11,23	14,1	NO	758
	27	17,31	16,51	yes	10	20,06	18,69	18,68	yes	331
	38	15,57	12,74	no	11	8,46	7,28	7,28	no	495

Stylometric Results

Based on the results, there is:

- Strong evidence of shared authorship between document sets ***Admit* and *Denied*** set.
    - overall keyness average and mean of 1,27 and 1,4 respectively;
    - a grammatical keyness average and of -1,14 (based on 27 itmes) and -0,99 (based on 23 items).
    - The average of the 10 midrange grammatical keywords of -0,05 provides additional statistical support

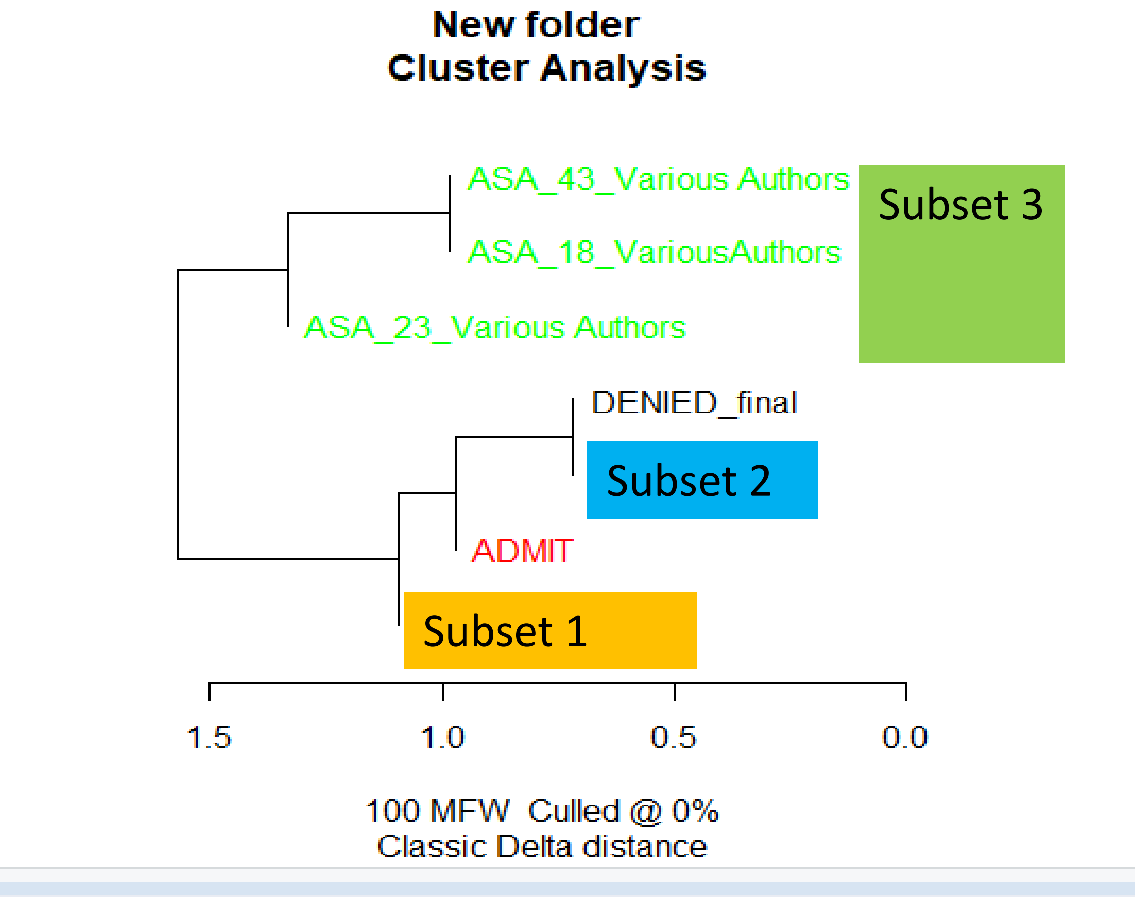
Therefore, stylometric results suggest refuting H0.

Stylometric Results

- strong evidence that the document set ***Sub-set 1*** was authored by the same person who compiled the *Admit*-set.
- Fair to strong evidence that the document set ***Sub-set 2*** was authored by the same person who compiled the *Admit*-set. **contains a number of shorter texts*
- Fair evidence that the document set ***Sub-set 3*** was authored by the same person who compiled the *Admit*-set. **subject to genre restrictions*

Therefore, stylometric results overall do not support H0.

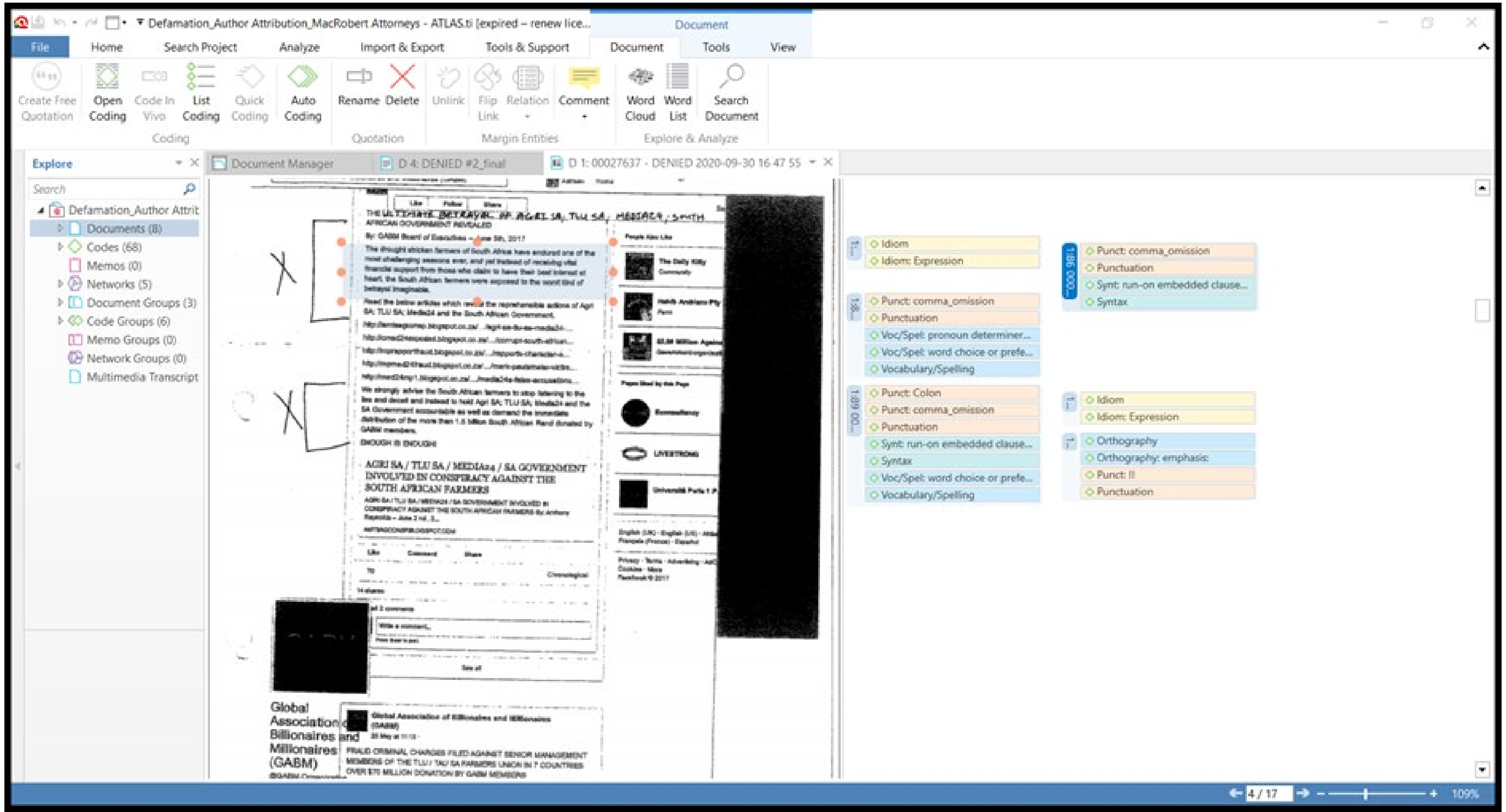
Digital tool: Stylo



Stylistic Analysis with digital tool Atlas.ti

Instruments

- Data were annotated using computer software Atlas.ti (8.0)



Like Follow Share

THE ULTIMATE BETRAYAL OF AGRI SA, TLU SA, MEDIA24, SOUTH AFRICAN GOVERNMENT REVEALED

By: GABM Board of Executives – June 5th, 2017

The drought stricken farmers of South Africa have endured one of the most challenging seasons ever, and yet instead of receiving vital financial support from those who claim to have their best interest at heart, the South African farmers were exposed to the worst kind of betrayal imaginable.

Read the below articles which reveal the reprehensible actions of Agri SA; TLU SA; Media24 and the South African Government.

We strongly advise the South African farmers to stop listening to the lies and deceit and instead to hold Agri SA; TLU SA; Media24 and the SA Government accountable as well as demand the immediate distribution of the more than 1.5 billion South African Rand donated by GABM members.

ENOUGH IS ENOUGH

AGRI SA / TLU SA / MEDIA24 / SA GOVERNMENT INVOLVED IN CONSPIRACY AGAINST THE SOUTH AFRICAN FARMERS

Like Comment Share

Chronological

People Also Like

The Daily Kitty

Community

LIVESTRONG

English (UK) · English (US) · Français (France) · Español

Privacy · Terms · Advertising · Cookies · More

Facebook © 2017

[redacted]

Global Association of Billionaires and Millionaires (GABM)

Global Association of Billionaires and Millionaires (GABM)

FRAUD CRIMINAL CHARGES FILED AGAINST SENIOR MANAGEMENT MEMBERS OF THE TLU / TAU SA FARMERS UNION IN 7 COUNTRIES OVER $70 MILLION DONATION BY GABM MEMBERS

1:...
- Idiom
- Idiom: Expression

1:8...
- Punct: comma_omission
- Punctuation
- Voc/Spel: pronoun determiner...
- Voc/Spel: word choice or prefe...
- Vocabulary/Spelling

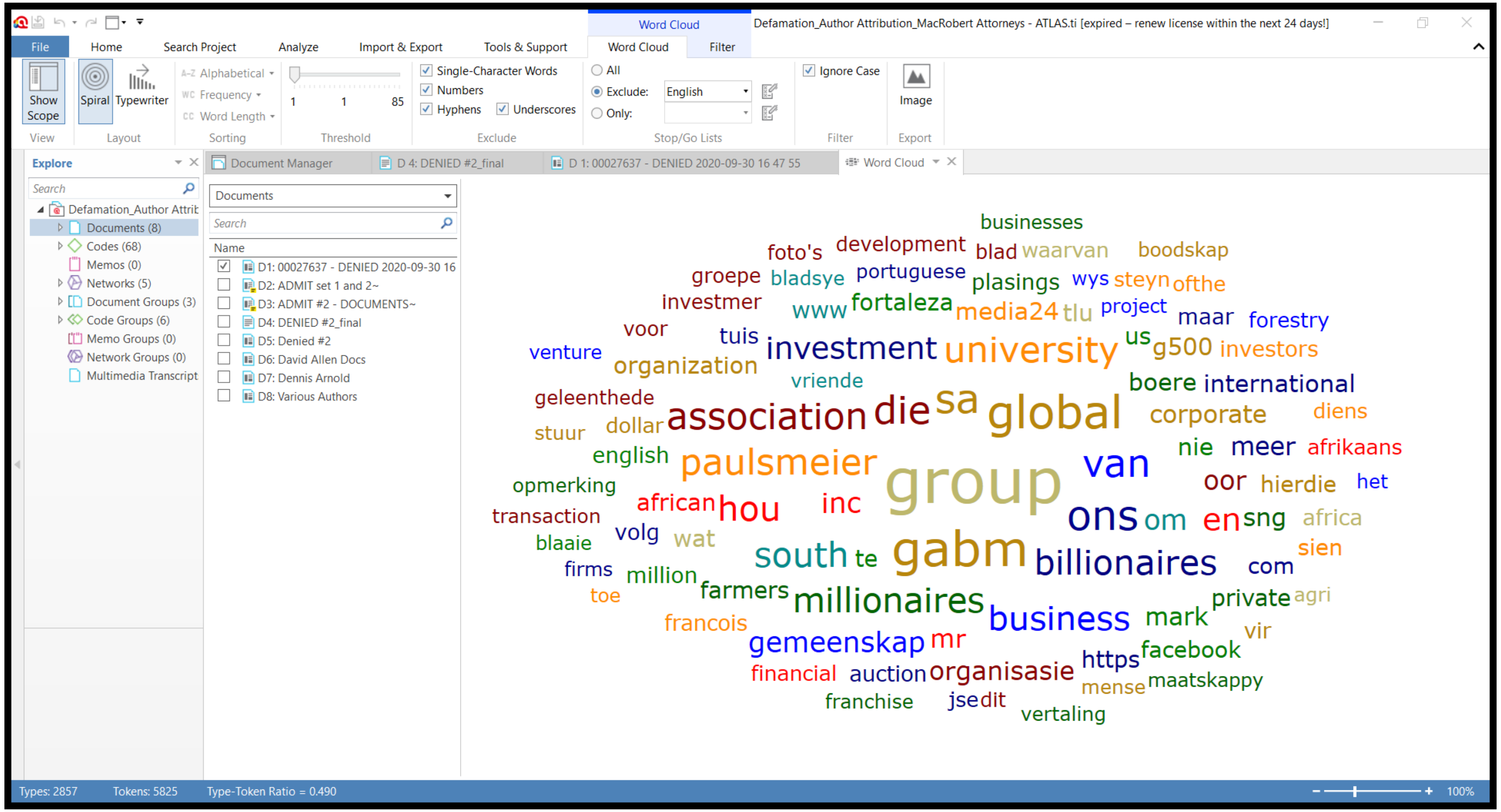
1:89 00...
- Punct: Colon
- Punct: comma_omission
- Punctuation
- Synt: run-on embedded clause...
- Syntax
- Voc/Spel: word choice or prefe...
- Vocabulary/Spelling

1:86 000...
- Punct: comma_omission
- Punctuation
- Synt: run-on embedded clause...
- Syntax

1:...
- Idiom
- Idiom: Expression

1:...
- Orthography
- Orthography: emphasis:
- Punct: !!
- Punctuation

Word cloud



Defamation_Author Attribution_MacRobert Attorneys - ATLAS.ti [expired – renew license within the next 24 days!]

☑ Single-Character Words
☑ Numbers
☑ Hyphens ☑ Underscores

○ All
● Exclude: English
○ Only:

☑ Ignore Case

Threshold: 1 1 85

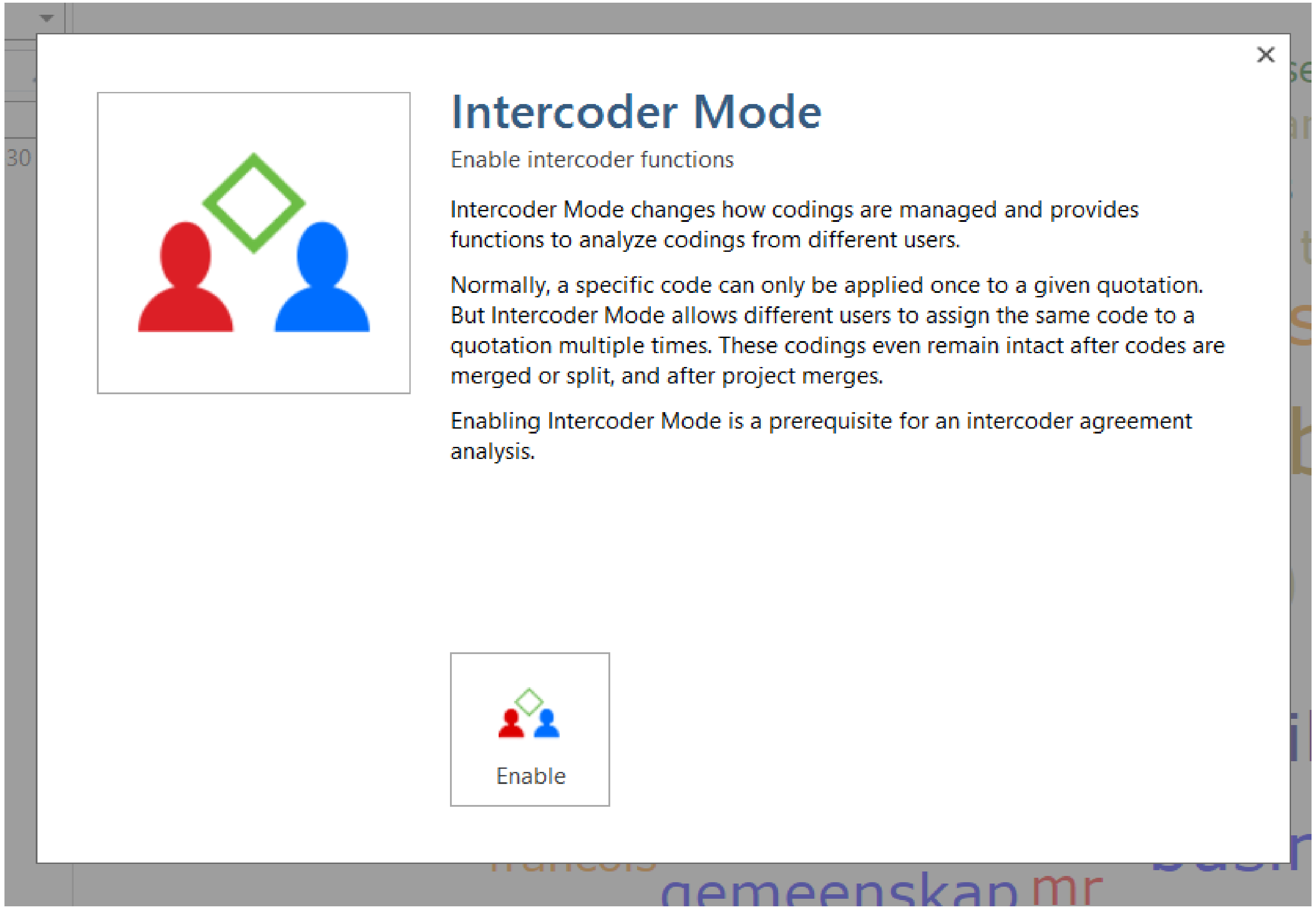
Documents

	Name
☑	D1: 00027637 - DENIED 2020-09-30 16
□	D2: ADMIT set 1 and 2~
□	D3: ADMIT #2 - DOCUMENTS~
□	D4: DENIED #2_final
□	D5: Denied #2
□	D6: David Allen Docs
□	D7: Dennis Arnold
□	D8: Various Authors

Types: 2857 Tokens: 5825 Type-Token Ratio = 0.490 100%

Word Cloud
File
Home
Search Project
Analyze
Import & Export
Tools & Support
Word Cloud
Filter
Show Scope
Spiral
Typewriter
Alphabetical
Frequency
Word Length
View
Layout
Sorting
Threshold
Exclude
Stop/Go Lists
Filter
Image
Export
Explore
Search
Defamation_Author Attrib
Documents (8)
Codes (68)
Memos (0)
Networks (5)
Document Groups (3)
Code Groups (6)
Memo Groups (0)
Network Groups (0)
Multimedia Transcript
Document Manager
D 4: DENIED #2_final
D 1: 00027637 - DENIED 2020-09-30 16 47 55
Word Cloud
Search
businesses
foto's development blad waarvan boodskap
groepe bladsye portuguese plasings wys steyn ofthe
investmer www fortaleza media24 tlu project maar forestry
voor tuis investment university us g500 investors
venture organization vriende boere international
geleenthede association die sa global corporate diens
stuur dollar english paulsmeier van nie meer afrikaans
opmerking group oor hierdie het
transaction african hou inc ons om en sng africa
blaaie volg wat south te gabm billionaires com sien
firms million farmers millionaires business mark private agri
toe francois gemeenskap mr vir
financial auction organisasie https facebook
franchise jse dit mense maatskappy vertaling

Intercoder mode



Intercoder Mode

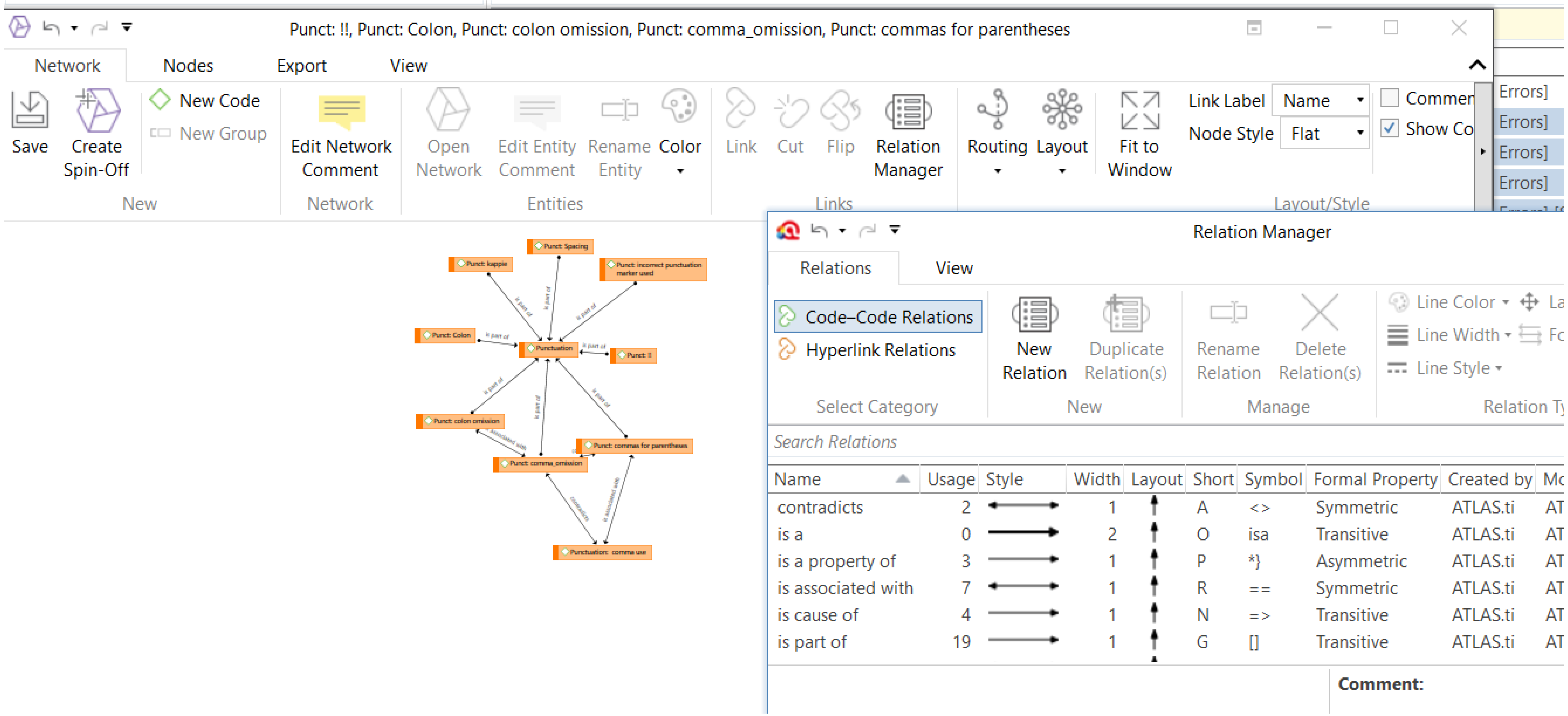
Enable intercoder functions

Intercoder Mode changes how codings are managed and provides functions to analyze codings from different users.

Normally, a specific code can only be applied once to a given quotation. But Intercoder Mode allows different users to assign the same code to a quotation multiple times. These codings even remain intact after codes are merged or split, and after project merges.

Enabling Intercoder Mode is a prerequisite for an intercoder agreement analysis.

Open Network



Punct: !!, Punct: Colon, Punct: colon omission, Punct: comma_omission, Punct: commas for parentheses

Relation Manager

Search Relations

Name	Usage	Style	Width	Layout	Short	Symbol	Formal Property	Created by	Mc
contradicts	2	↔	1	↑	A	<>	Symmetric	ATLAS.ti	AT
is a	0	→	2	↑	O	isa	Transitive	ATLAS.ti	AT
is a property of	3	→	1	↑	P	*}	Asymmetric	ATLAS.ti	AT
is associated with	7	↔	1	↑	R	==	Symmetric	ATLAS.ti	AT
is cause of	4	→	1	↑	N	=>	Transitive	ATLAS.ti	AT
is part of	19	→	1	↑	G	[]	Transitive	ATLAS.ti	AT

Comment:

Stylistic enquiry

- Systematic deductive analysis of the uncontested documents, followed by an inductive description of the contested samples.
- UCREL Log-Likelihood and Effect-Size Calculator to investigate whether statistically significant differences were evident between the most prominent features identified during the error analysis and stylistic analysis.

Findings

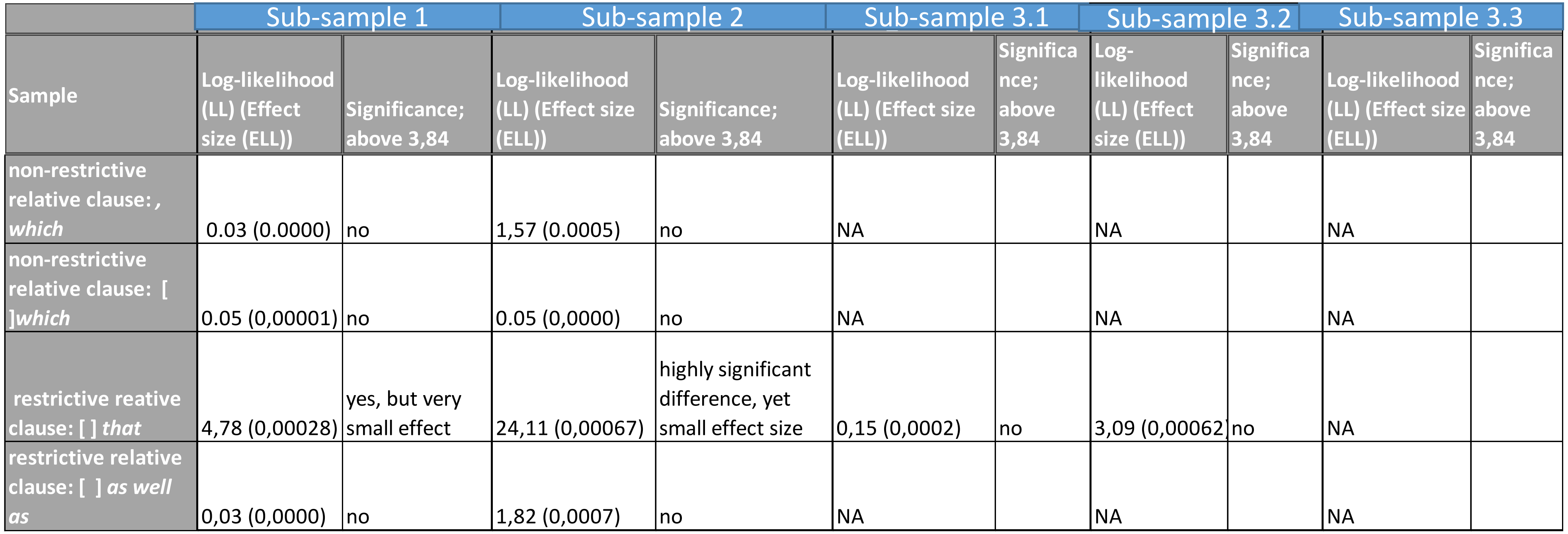
- The language used across samples is complex.
- Style is generally formal to very formal, regardless of the genre of the document.
- The author/s appear to be well-educated and highly proficient in English.
- Various instances across suspect sets display similar errors;
  - Specific combination of tense and aspect errors as well as and syntactic punctuation choices
  - errors are characteristic of second language speakers of English

Syntax

a. Tense and aspectual preferences and errors	(i) Please allow me to explain the purpose behind my earlier communication as well as the generous offer extended to you to settle the subject matter before it is going to court. (ii) as a result of the intentional criminal actions by [NAME] and [NAME] requires that the matter being taken to the High Courts of South Africa…	(i) Media24, the [NAME] open letter has been served at your corporate offices more than 3 months ago. (ii) The appointment of the [NAME] security division to recover more than $120 million on behalf of 233 members of [NAME] has been confirmed yesterday by [NAME]
b. Subject-verb agreement errors	…here is very good reasons to amend these dates	(i) You, [NAME], has *indicated* …
c. Sentence complexity: marked word order and semantic relation	(i) It will neither be expected from [NAME] to manage a fund	(i) The emergency funding commitment by the [NAME] members in terms of the "XXX Project" is strictly
d. Sentence complexity: complex subordination and omission of syntactic punctuation markers (i.e. run-on sentences)	(i) The serious nature of the defamation involving several international legal persons and I as well as the resulting financial losses incurred by myself as a result of the intentional criminal actions by [NAME] and [NAME] requires that the matter being taken to the High Courts of South Africa, which ultimately would have disastrous consequences for both [NAME] and [NAME].	(i) To my utmost shock and surprise the quest was one of the detectives I was working with and he came to sell a 20 carat diamond he stole from a raid and at the same time he warned them about me. (ii) You, [NAME], has indicated to the media and South African farmers that[NAME] can and will not support the $70million [NAME] Project because the [NAME] organization refuses to submit to [NAME] its private audited financial statements, banking details, names of auditors, names of the [NAME] Board of Executives and other detailed corporate information on the GABM organization.
e. Sentence structure and clausal subordination: relational clauses with which; as well as and that.	(i) This particular letter entails two aspects, which approval in our opinion is merely academic [*comma omission*] but we need that nonetheless. (ii) There is also no legal framework that governs virtual currencies in SA, [comma used here to signal parenthesis, but the parenthesis is not closed] and for your information the [NAME] has published "[NAME] Position Paper on Virtual Currencies", which indicate that transactions are not regulated by the SARB and thus do not need to obtain [NAME] approval. (iii) …and media releases which would inform non [NAME] members as well as members of other agricultural unions about the subject matter.	(i) [NAME]'s vision has always been to establish his own private investment group that is very exclusive and lucrative, which subscribes to the highest standards of international corporate governance practices and ethical business practices others can only aspire to. (ii) The incident was reported in the media, which resulted in a huge affair at the police headquarters. (iii) At the same time the world economy is still in dire straits which [non-restrictive; grammatically incorrect} make it even more difficult for developing countries like South Africa to respond to the plight from commercial farmers to access disaster funding; (iv) Commercial farmers that wish to apply for emergency funding are required:

Statistical confirmation

Table 1 Summary of Log-Likelihood and Effect Size Results for the use of "which", "that" and "as well as" in the contested samples



	Sub-sample 1		Sub-sample 2		Sub-sample 3.1		Sub-sample 3.2		Sub-sample 3.3	
Sample	**Log-likelihood (LL) (Effect size (ELL))**	**Significance; above 3,84**	**Log-likelihood (LL) (Effect size (ELL))**	**Significance; above 3,84**	**Log-likelihood (LL) (Effect size (ELL))**	**Significance; above 3,84**	**Log-likelihood (LL) (Effect size (ELL))**	**Significance; above 3,84**	**Log-likelihood (LL) (Effect size (ELL))**	**Significance; above 3,84**
non-restrictive relative clause: , *which*	0.03 (0.0000)	no	1,57 (0.0005)	no	NA		NA		NA	
non-restrictive relative clause: [ ]*which*	0.05 (0,00001)	no	0.05 (0,0000)	no	NA		NA		NA	
restrictive reative clause: [ ] *that*	4,78 (0,00028)	yes, but very small effect	24,11 (0,00067)	highly significant difference, yet small effect size	0,15 (0,0002)	no	3,09 (0,00062	no	NA	
restrictive relative clause: [ ] *as well as*	0,03 (0,0000)	no	1,82 (0,0007)	no	NA		NA		NA	

Punctuation and Orthography

a) Punctuation and orthographical preferences: **i**. the use of all capital letters, e.g. in headings, openings of letters and emails; or to draw attention; **ii.** the use of exclamation marks to express anger and/or urgency; iii**.** the use of inverted commas (i.e. quotation marks) to identify names, as opposed to italic print, for example; **iv.** the use of headings, which creates white space on the page; **v**. the use of numbering and bulleted lists to structure information and improve the readability of densely written texts (lists are often introduced by a colon in mid-sentence)	RE: DEFAMATION LAWSUIT- [NAME] (ii) Something is seriously wrong here! (iii) , and for your information the SARB has published "[NAME] Virtual Currencies", which indicate that transactions are not regulated (iv) INTRODUCTION SA Agri's actions to date: (v) 1. The first aspect entails the….. 2. The second aspect entails the confirmation from the Ministry of Agriculture	(i) The R1,5 billion [NAME] donation to the drought stricken farmers that was misappropriated MUST BE RETURNED! (ii) In response to [NAME] excellent open letter to [NAME]….they have apparently deliberately and intentionally published in terms of the $70 million [NAME] Project: [followed by three points, numbered i, ii and iii]. (iii) The R1,5 billion GABM donation to the drought stricken farmers that was misappropriated MUST BE RETURNED! (iv) "Secret Millionaire Project" HISTORY OF THE G500 PRIVATE INVESTMENT GROUP PART 1: PORTUGUESE INVESTORS GROUP -1981 TO MARCH 1994 … Preamble (v) The objective of the diamonds sales were to: - generate cash reserves - the acquisition of properties in Europe - the acquisition oi businesses in Europe - fund various business deals.
b) Syntactical punctuation omission	(i) This would be done by direct communication to [NAME] members, [*comma use for listing*] and media releases [*comma omitted to signal subordination before determining pronoun 'which'*] which would inform non-[NAME] members [*comma omission preceding 'as well as' to signal parenthesis*] as well as members of other agricultural unions [*omission of comma to close parenthesis*] about the subject matter.	(i) Why did you, [NAME], [*comma use for parenthesis*] made the deliberate false and incorrect statement to a Beeld journalist which [*comma omitted to open parenthesis and signal subordination*] alleges the [NAME] Bond financial instruments are different from other fiat currencies, [*parenthesis closed*] and that same cannot be exchanged to South African Rand [*comma omitted to signal conditiona* l] if it enters South Africa [*comma omitted to show conjunction*] as would be in the case with the US Dollar, British Pound [*comma use for listing*] or Euro currencies.

Vocabulary, semantic and spelling preferences

a) Spelling preferences	N.A.	(i) to the greater good as appose to personal interest above all (ii) is making the $70 million donation as appose to the actual 233 GABM members (iii) : Russian investors for an amount of $50 million as appose to the real market value which should be in the range of $120 million.
b. Lexical choices	(i) The first aspect entails the confirmation from the Ministry of Finance that [determiner] confirms that [relative pronoun] the donations will be tax free to SA commercial farmers, and that [relative pronoun] [NAME] indeed are not required to obtain [NAME] approval to sell the investment [NAME] virtual currency Bearer Bonds to SA citizens. (ii) …and media releases which would inform non [NAME] members as well as members of other agricultural unions about the subject matter.	(i) The Monarch Consortium partners however out-voted [NAME] in the decision that [relative pronoun] the [NAME] Inc should be based and operated from South Africa, (ii) and the SA Government accountable as well as demand the immediate distribution of the more than 1.5 billion South African Rand donated by [NAME] members.
c. Idiomatic Expressions	(i) …the said information and facts… (ii) …publishing of the malicious, false and misleading…	(i) …apparently deliberately and intentionally (ii) …false allegations, alternative facts, fake news, defamatory statements, misleading information…

Stylistic Conclusion

- **Based on the stylistic analysis, H0 is refuted and so, H1 is supported.**

- Ample evidence to support shared authorship across the sets.
- Little evidence to suggest that texts were produced by another author who is an L1 speaker of a specific variety of English.
- Rather, the combination of stylistic features evident across the samples support the notion of a shared author and most likely one who is not a native speaker of supposed variety of English.

The approach

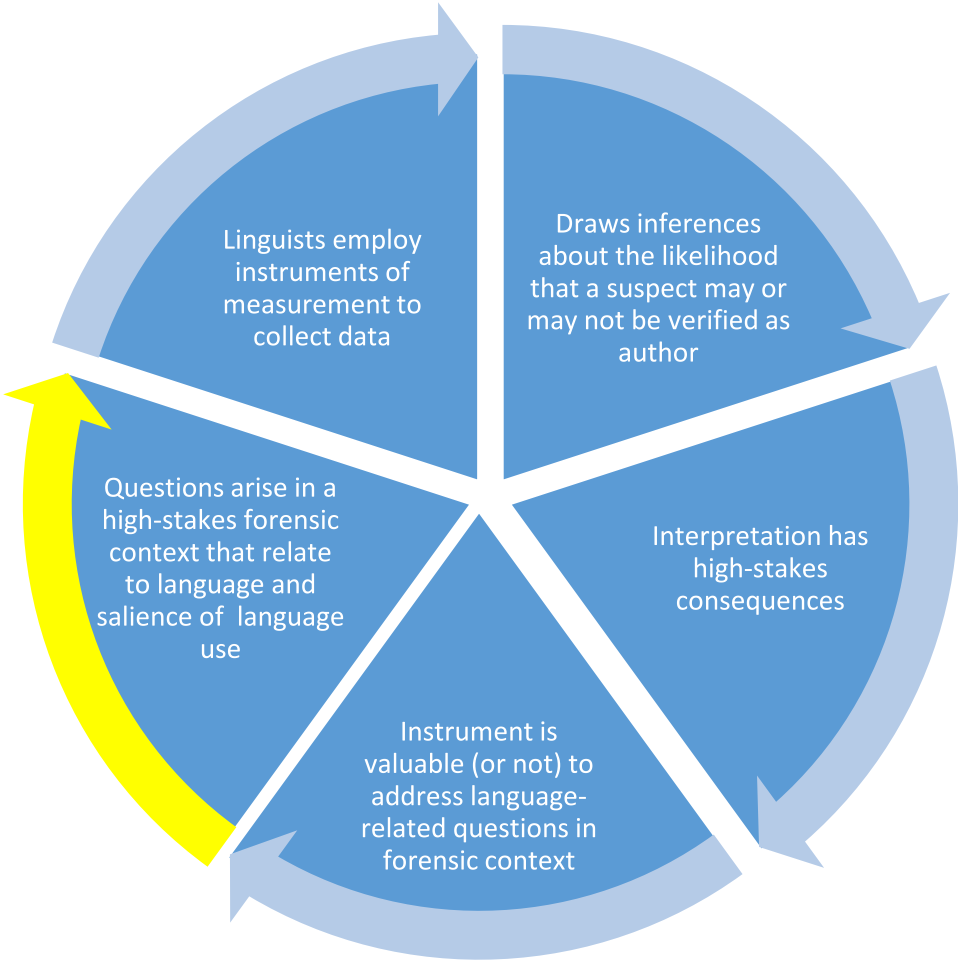
A combination of stylistic and stylometric analysis and viewed from opposing perspectives as first implemented by Kotzé in *State v Kerr Hoho* (Kotzé & Bezuidenhout, 2007; Kotzé, 2010).

“The combination of different approaches, and, in fact, opposite hypotheses, regarding the identification of authorship on the basis of an attested source of data, must contribute to the principle of scientific rigour underlying the credibility and acceptance of expert evidence submitted by linguists in court. The fact that the stylometric calculations are based on principles of statistical significance which can be demonstrated in a transparent way in court, and that clear-cut correspondences between the source text and those under investigation, linguistic and otherwise, can be identified and serve as complementary evidence to corroborate the findings of the analyst, has contributed to the recent acceptance of such evidence as valid and reliable before the High Court in South Africa and the Council of a South African university. I believe that the application of the combined approach to a wider range of text types should lead to an increasing refinement of this methodology in future” (Kotzé, 2010:194-195).

Validation of the approach

- Argument-based approach (Kane, 1989; 2012; Kane & Wools, 2019)
- Various methodologies are combined to collect data that motivates the use of these instruments to elicit relevant responses in authentic ways that simulate language processing as would be required to complete similar tasks in real life.

Validation in the context of Forensic Linguistics


Linguists employ instruments of measurement to collect data
Draws inferences about the likelihood that a suspect may or may not be verified as author
Interpretation has high-stakes consequences
Instrument is valuable (or not) to address language-related questions in forensic context
Questions arise in a high-stakes forensic context that relate to language and salience of language use

How can DH contribute to FL/ FL contribute to DH?

- Facilitate
- Enrich
- Contribute the a robust argument for the validity of linguistic measurement

More to read

- Kestemont et al (2021) present an overview of cross-domain authorship verification tasks;
- Longhi (2020) addresses the use of digital humanities and linguistics to help with terrorism investigations;
- Hunyadi, Arabi and Toth (2003) consider the contribution of Forensic lInguistics to Humaities Computing;
- Zamecnik and Lackova (2021) proposes a methodological basis for building digital humanities education on a Linguistic Background.

Concluding remarks

1. Validity is the most fundamental quality of measurement instruments in FL-context.
2. One size does not fit all
3. Contextual considerations need to be taken into account:
   1. Text length
   2. Genre
   3. Multiple authorship
   4. Multilingual authors

4. The need for interdisciplinary exploration: DH, linguistics and law.

Invitation and thank you

Future research

- Continued validation is needed
  - same data sets, different linguists
  - different data sets; same methodology
  - Refinement: use of log-likelihood calculator, Atlas.ti & Stylo
  - Database of authorship cases in South Africa

- Need for theoretical validation framework as point of reference in FL-context
  - Kane’s (1992; 2010; 2011; 2016) conceptualization of a validation argument

Bibliography

Chapelle, C. A. Validity in language assessment. *Annual Review of Applied Linguistics*, vol. 19, 1999, pp. 254–272., DOI:10.1017/S0267190599190135.

Fulcher, G. 2013. Practical language testing. *Practical Language Testing*. 1-352. DOI: 10.4324/9780203767399.

Kane, M.T. 1992. An argument-based approach to validity. *Psychological Bulletin,* 112(3): 527-535.

Kane, M.T. 2010. Validity and fairness. *Language Testing,* 27(2): 177-182. DOI: 10.1177/0265532209349467.

Kane, M.T. 2011. Validity score interpretations and uses: Messick lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing.* 29(1): 3-17.

Kane, M.T. 2016. Explicating validity. *Assessment in Education: Principles, Policy & Practice* 23(2): 198-211; DOI: 10.1080/0969594X.2015.1060192

Kane, M.T. & Wools, S. 2019. Perspectives on the Validity of Classroom Assessments. *Classroom assessment and educational measurement.* Routledge

Kotzé, E.F. & J. Bezuidenhout. 2007. In search of the author ... 'linguistic fingerprints' as identifying evidence, in De Rebus, September: 22-25.

Kotzé, E.F. 2010. Author identification from opposing perspectives in forensic linguistics, *Southern African Linguistics and Applied Language Studies*, 28:2, 185-197, DOI: 10.2989/16073614.2010.519111

Lazaraton, & Taylor, L. 2007. Qualitative Research Methods in Language Test Development and Validation. Fox, J. et al. *Language Testing Reconsidered*. Chapter 6 Ottawa: Les Presses de l'Université d'Ottawa | University of Ottawa Press, 2007. (pp. 113-129) Web. <http://books.openedition.org/uop/1570>.

Messick, S. 1989. Validity. In R. L. Linn (Ed.), *Educational measurement*. 3rd ed., pp. 13-103. Macmillan.

Popham, J. 2017. *Classroom assessment: What teachers need to know*. Eighth edition. Pearson.

NWU®

NORTH-WEST UNIVERSITY
NOORDWES-UNIVERSITEIT
YUNIBESITI YA BOKONE-BOPHIRIMA

Thank you

Karien van den Berg

North-West University

karien.hattingh@nwu.ac.za