ALIAS Methods for HIATUS

An Introduction to ALIAS Technology for Forensic Authorship Identification

Carole E Chaski PhD
ALIAS Technology LLC
US 1 302-856-9488
cchaski@aliastechnology.com
19 January 2022

1. ALIAS Technology Firsts and Future

ALIAS Technology has pioneered in authorship id

- We have pioneered a series of “firsts in the field”….
  - First arrest warrant using author id for probable cause (1993)
  - First federally funded grant in authorship identification methods (1995 - 2004)
  - First vetted database of ground truth authorship (1995)
  - First testing of authorship id methods on ground truth data (1996 – 2001)
  - First development and use of lineup technique (1997)
  - First method fully admitted as scientific evidence in US courts (2001)
  - First web-accessible, multi-functional software (2009) …..
- ***But we aren’t finished yet!***
- HIATUS offers the opportunity to pioneer (a) multilingual SynAID and (b) author privacy, building on a solid foundation of ALIAS SynAID.
- We look forward to teaming with an experienced partner.

2. Supporting HIATUS Goals
A. Validated Authorship Identification

- ALIAS SynAID is a validated method for forensic authorship identification
  - **correct identification of known documents from 95% to 100%** in various validation tests and cases (Chaski 2001, 2005, 2013, 2021, 2022)
- SynAID has been validated by independent research teams
  - Spain, Canada, US
- ***Civil justice:*** value of civil cases is over US$ 75 billion
- ***Security:*** value of insider threat/corporate security is in US$ Millions
- ***Criminal justice:*** value in major felony cases for prosecution and defense is priceless

B. Reliable Authorship Identification

- SynAID is the only current method that provides a **known accuracy rate for each set of documents** in an investigation
- SynAID is used to classify questioned documents when known accuracy is high
- SynAID provides a **probability of authorship for each document** that is tested
- SynAID is the only current method that has developed **standard operating protocols for data quantity and data quality.**

C. Use on Different Registers, Topics and Genres

- SynAID is the only forensic authorship id method that **operates with high reliability on documents across registers, topics and genres**.
- SynAID was developed and tested on **a database of vetted authors who produced writing samples across registers, genres and topics**.
  - Ten writing samples for each author (not scraped from web)
  - Multiple Registers: formal and informal
  - Multiple Topics: eight different topics
  - Multiple Genres: three different genres

D. Explainable, Human-Interpretable Approach

- SynAID is grounded in linguistics: its core algorithms are both **sophisticated and explainable**.
- SynAID's approach combines syntactic theory, psycholinguistics, and neurolinguistics with statistical classifiers.
- SynAID's current statistical classifiers are explainable as traditional statistics (not invisible layers of neural nets).
- One Federal judge commented during an evidence hearing: **"I'm admitting this evidence because it makes sense."**

E. Multi-Lingual Approach

- SynAID’s approach is **theoretically multi-lingual** and this can be tested.
- Since all languages have syntax, SynAID can use a particular algorithm for each language.
- Particular language algorithms can be developed.
- SynAID’s current parser is multi-lingual.

F. Author Privacy

- ALIAS Tech differentiates **related concepts** (Chaski 2017, 2018).
  - Authorship
  - Plagiarism
  - Style
  - Textual similarity
  - Imitatability of features
  - Edited Text
- ALIAS Tech's approach to author privacy can build on these differences.

3. ALIAS Technology Team for HIATUS

Team Members and Skillsets

- **Carole Chaski**, PhD Linguistics, (Brown U), MEd Psychology of Reading (U Delaware), English and Ancient Greek (Bryn Mawr College)
  - Computational linguistics, syntax, language variation, psycholinguistics
  - Database design and scripting, design of experiments
- **Cristina Aggazzotti**, PhD Linguistics (Harvard), MS Decision Sciences (UCL UK), Mathematics and Linguistics (U Southern California)
  - Computational linguistics, formal semantics, Bayesian quantitative analysis
  - Machine learning
- **Support Coding Staff**: Java, Python, Database Scripting, REST API
- **Support Linguists**: PhDs in Linguistics with expertise in Spanish, Korean, Russian, Arabic, Urdu (can be expanded as needed)