# An Overview of the Traditional Authorship Attribution Subtask

Patrick Juola
Duquesne University, Juola & Associates
juola@mathcs.duq.edu
pjuola@juolaassoc.com

# Authorship Attribution

- Long-standing problem in many disciplines
  - Plagiarism detection
  - History/literature studies
  - Journalism and law
    - et cetera
- Statistical ("nontraditional") approach commonplace
- "Traditional" vs. new applications (e.g. authorship profiling for criminal behavior)

# Differences since PAN2011

- New and wonderful city
- Number and size of documents decreased
- Different genre
- Lack of automatic markup
- New sub-sub-sub-sub-* task : clustering

# Problems presented

- Authorship Attribution
  - Three pairs of problems (A,B), (C,D), and (I,J)
  - Each pair contains works by the same author, the difference being the first problem is closed class while the second is open; task is to identify author
- Author Clustering
  - Two problems E, and F
  - They consist of a number of documents with joint authorship; task is define what sections were written by different authors

# Authorship Identification

- A/B
  - 3 authors, ~5000 words/sample, 6/10 test documents
- C/D
  - 8 authors, ~10000 words/sample, 8/17 test docs
- G/H discarded
- I/J
  - 14 authors, ~100,000 words/sample, 14/16 test docs

# Authorship Clustering

- E
  - 3 30-paragraph “documents,” intermixed by paragraphs from 2/3/4 separate authors
- F
  - 4 20-paragraph “documents,” single intrusive section from single other author
  - One document had no intrusion

# Evaluating

- Each document or paragraph was independently judged as right or wrong.
- E was harder since clusters needed matching
  - Hand-judged based on best match
  - Still possible for low score if participant determined wrong number of clusters
- Scored :
  - Average correct per problem
  - Total number of documents correct

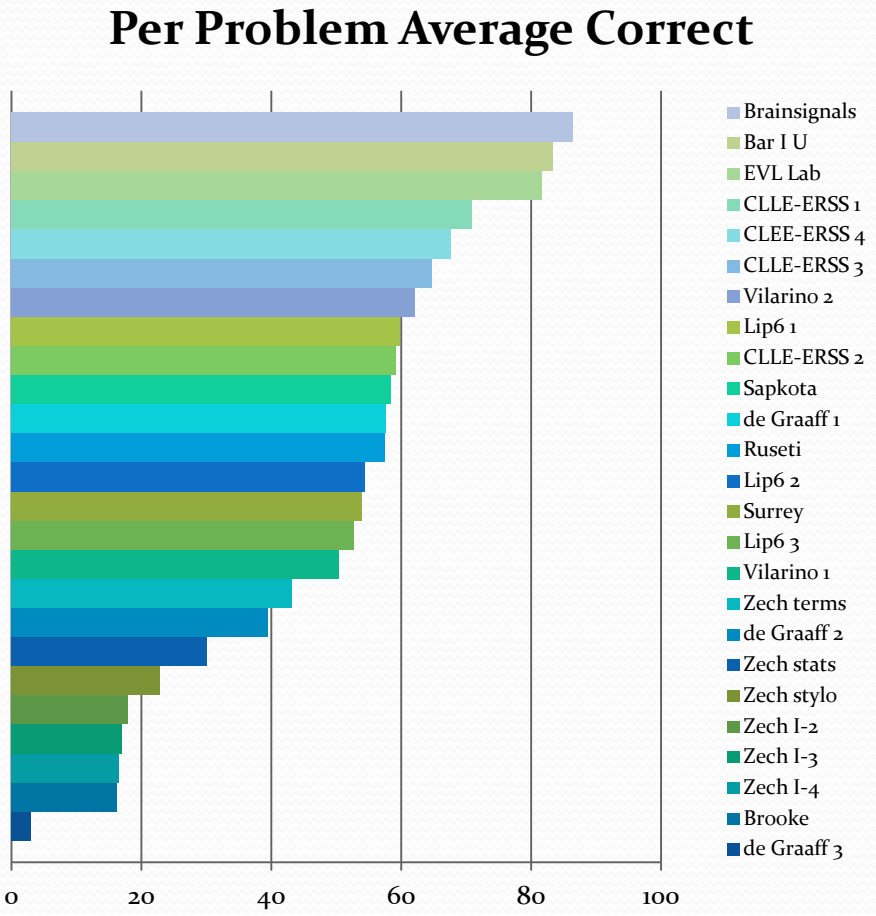
# Evaluating (cont)

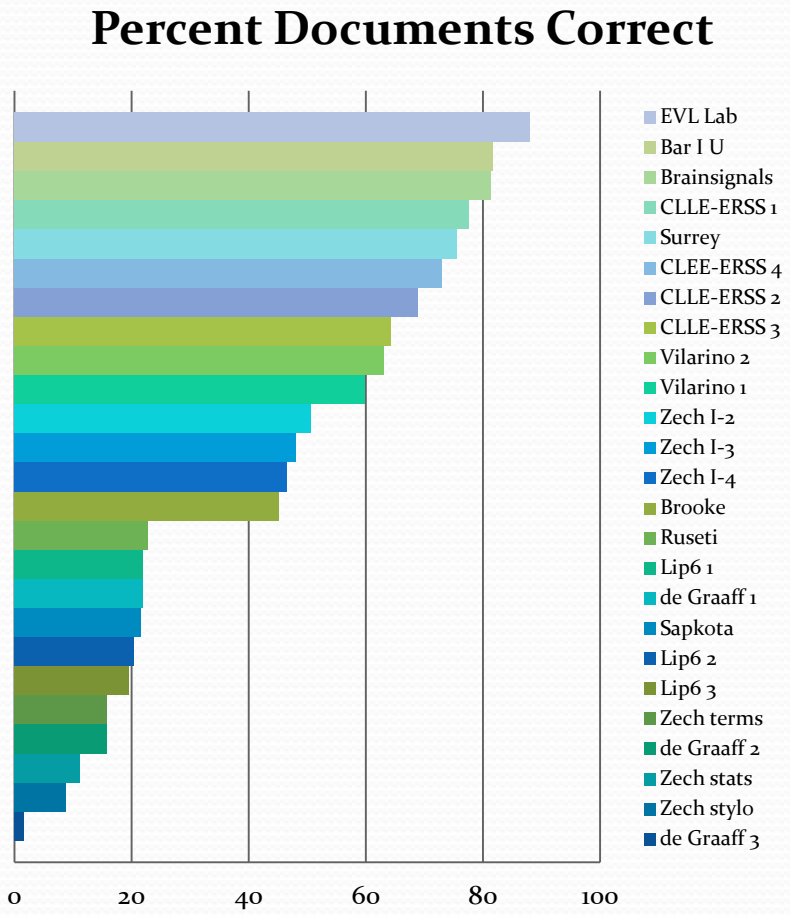
- Example :
  - Ground truth: 1.. 15, 16..30
  - Submitted: 1, 2, 3, 4, 5 … 29, 30
    - Matching red-green and blue-black yields 16/30 correct
    - Matching red-black and blue-green yields 14/30 correct
  - Scored as 16/30 (~53%)

# Participants

- Twelve teams
- Twenty-five submissions
  - Some partial submissions (e.g. only E/F)
- Full league table in proceedings

# Summary of Results



# Congratulations to:

- Brainsignals (Fraunhofer FIRST Berlin, Germany; University of Bucharest, Romania)
- Bar-Ilan University, Israel
- EVL Lab (Duquesne University, USA)

# Proposed plan for 2013

- Simplified/streamlined task
  - Matched document pairs – same author?
  - All answers yes/no
  - Multilingual across pairs (which languages?)
  - Software submission with automatic grading
    - Possible option for manual participation (for "traditional" forensic linguists if interested)
  - Other details to be determined, contact me if you have opinions on genre, size, &c.