# Authorship attribution using Discriminant Function Analysis: Exploring literary stylistic variation in five Modern Greek novels

*George K. Mikros*

*University of Athens - Greece*

# Aims of the study

- Authorship attribution in 5 Modern Greek novels (4 authors).
- Specific research questions:
  - Is an arbitrary portion of a novel, carrier of authorship information?
    - How many words do we need for each novel segment?
    - How many novel segments do we need?
  - Are function words the only lexical source of authorship information?
  - Can the extraction of specific content words [author-specific words (ASW)] be used effectively in authorship attribution?

# Authorship “genome”

- Stylometric attempts to detect authorship have a long standing history starting from the Biblical studies and expanding to modern texts.
- A wide variety of statistical methods has been employed from machine learning to neural networks and multivariate techniques.
- The basic assumption is that each writer possess a idiosyncratic way of using his/her linguistic competence and this can be traced using quantitative methods.

# Problems in stylometry studies

- Following Rudman (1998) some of the most striking problems in stylometry studies are due to the lack of the homogeneity of the corpora examined. In particular:
  - The improper selection, unavailability or fragmentation of the texts.
  - The text normalization that often applies from the editor or the publisher causing serious distortion in the writer's style.
  - The cross validated texts should be controlled for genre, topic, date and medium when comparing to the training texts.

# Corpus selection criteria

- 5 novels from
  - 4 widely known modern Greek writers from the same publishing house
    - Same normalization conventions
- Matesis
  - The mother of the dog [47929 words]
- Michailidis
  - Murders [72831 words]
- Milliex
  - From the other side of the time [78077 words]
  - Dreams [9796 words] – Test novel
- Xanthoulis
  - The dead liqueur [28602 words]

# Experimental methodology

- Slice each novel in text segments of varying size (50 – 100 – 200 – 500 words).
- Create 4 different datasets for each size (50 words texts, 100 words text etc.).
- Subdivide each dataset further using random sampling in the text segments and creating 4 extra subcategories (20%, 40%, 60%, 80% of text segments compared to the original dataset).
- Calculate in each dataset 3 different feature groups:
  - Stylometric variables (ST)
  - Frequent function words (FFW)
  - Author - Specific words (ASW)
- Use Discriminant Function Analysis (DFA) in order to obtain authorship classification for each text segment in each dataset.

# Size of experimental datasets

| | Sample size (% of text segments) | | | | |
|---|---|---|---|---|---|
| **Text segment size** | ***20%*** | ***40%*** | ***60%*** | ***80%*** | ***100%*** |
| ***50 words*** | 979 | 1901 | 2788 | 3773 | 4736 |
| ***100 words*** | 498 | 929 | 1453 | 1896 | 2367 |
| ***200 words*** | 209 | 491 | 718 | 955 | 1181 |
| ***500 words*** | 87 | 191 | 309 | 380 | 471 |

# Feature Sets

- **Stylometric variables (ST)**
  - **Lexical "richness"**
    - Yule's K
    - Lexical Density
    - % of Hapax- and Dis- legomena
    - Dis- / Hapax - legomena
    - Relative entropy
  - **Character level measures**
    - Frequency of characters
  - **Word level measures**
    - Average word length (per text)
    - Word length distribution
- **80 most frequent function words (FFW)**
- **80 most distinctive author specific words (ASW)**

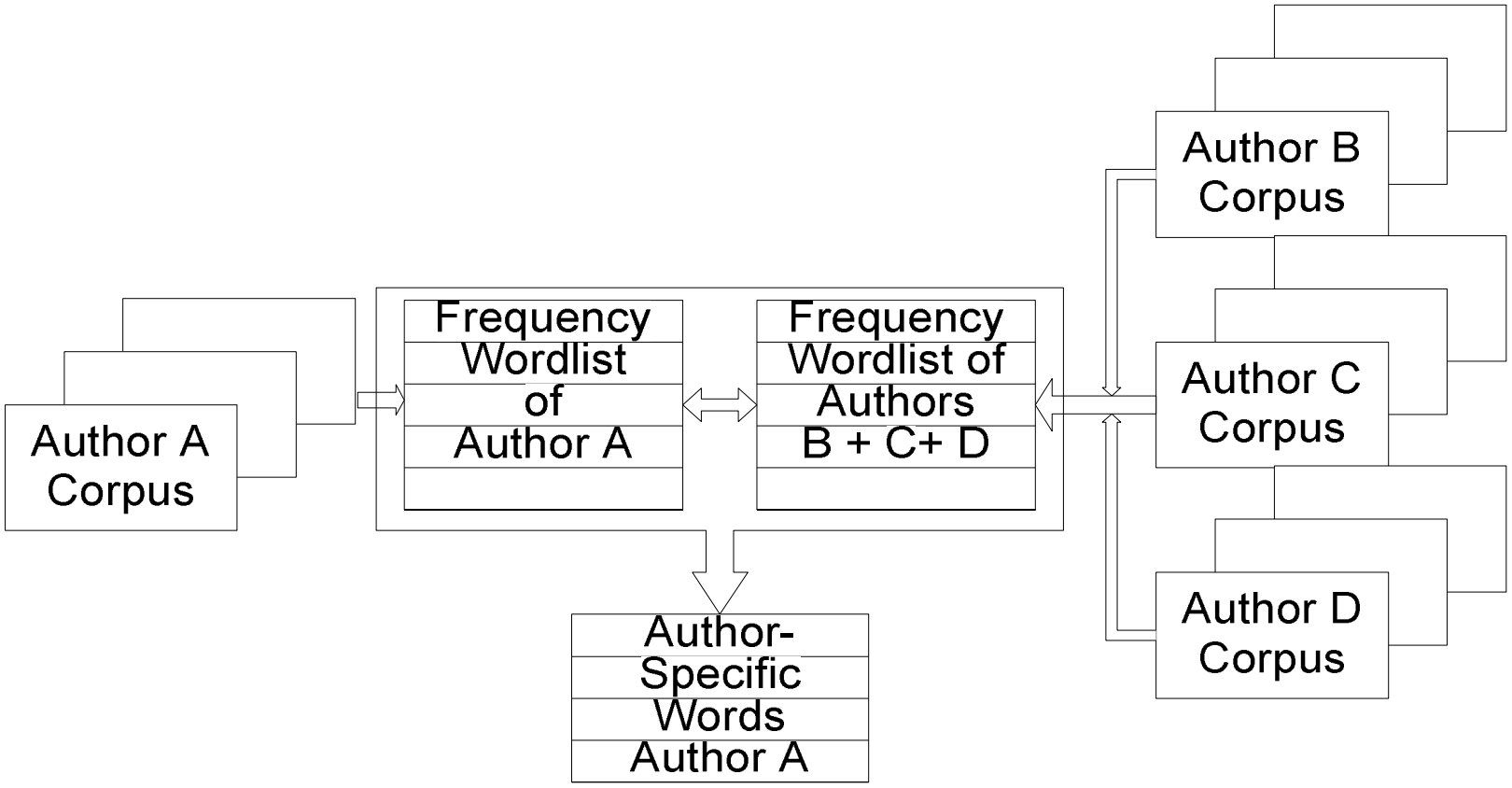
# Function vs Content words in authorship attribution studies

- Function words high frequency is considered as reliable authorship characteristic since it is beyond the conscious control of the author.
- However many recent studies have found evidence that content words carry stylistic information suitable for authorship attriburion (Baayen et al. 2002, Hoover 2004, Lancashire 1999, Schler et al. 2006).
- Content words in authorship attribution are selected with various methods:
  - Word distinctiveness ratio (Ellegård 1962)
  - Mutual information (Luyckx & Daelemans 2005)
  - Information gain on classification categories (Schler et al. 2006)

# Author Specific Words (ASW) method

- In Mikros (2003) we introduced a frequency profiling method in order to select content words for using them as discriminating variables in text categorization tasks. The procedure we proposed is explained briefly as follows:
  1. Selection of the training corpus.
  2. Formation of homogeneous sub corpora regarding the author of the included texts.
  3. Creation of frequency wordlists (FWL) for each of the sub corpora (for example Author A FWL, Author B FWL, Author C FWL, and Author D FWL).
  4. Comparison of each FWL with the unified FWL of the remaining authors, i.e., comparison of Author A FWL with the FWL which has been created joining Author B, Author C, and Author D FWLs
  5. Extraction of the k most frequent words that exhibit maximum discriminating power. The extraction is performed using Log Likelihood measure.
  6. Repetition of the procedure (stages 4 & 5) by deploying the remaining combinations of the available FWL comparisons.
  7. Extraction of n words (in the previous example 4 X k) which can be used as Author-Specific lexical variables in an authorship attribution training set.
- For the needs of our study we performed this methodology and we extracted 80 ASW (20 words per author). For every one of these words we calculated its frequency in each text of the corpus.

# Extracting ASW for the first of the four candidate authors



# Discriminant Function Analysis

- In order to explore the discriminating power of the selected variables in authorship attribution we used Discriminant Function Analysis (DFA).
- DFA involves deriving a variate, the linear combination of two (or more) independent variables that will discriminate best between a priori defined groups. Discrimination is achieved by setting the variate's weight for each variable to maximize the between-group variance relative to the within-group variance (Hair et al., 1995). If the dependent variables have more than two categories DFA will calculate k-1 discriminant functions, where k is the number of categories. Each function allows us to compute discriminant scores for each case for each category, by applying the formula:

$$D_{jk} = \alpha + w_1 x_{1k} + w_2 x_{2k} + \ldots + w_i x_{ik}$$

Where

$D_{jk}$= Discriminant score of discriminant function j for object k.

$a$= intercept

$w_i$= Discriminant weight for the independent variable i

$x_{ik}$= Independent variable i for object k

- For the needs for our study we used the step-wise method.

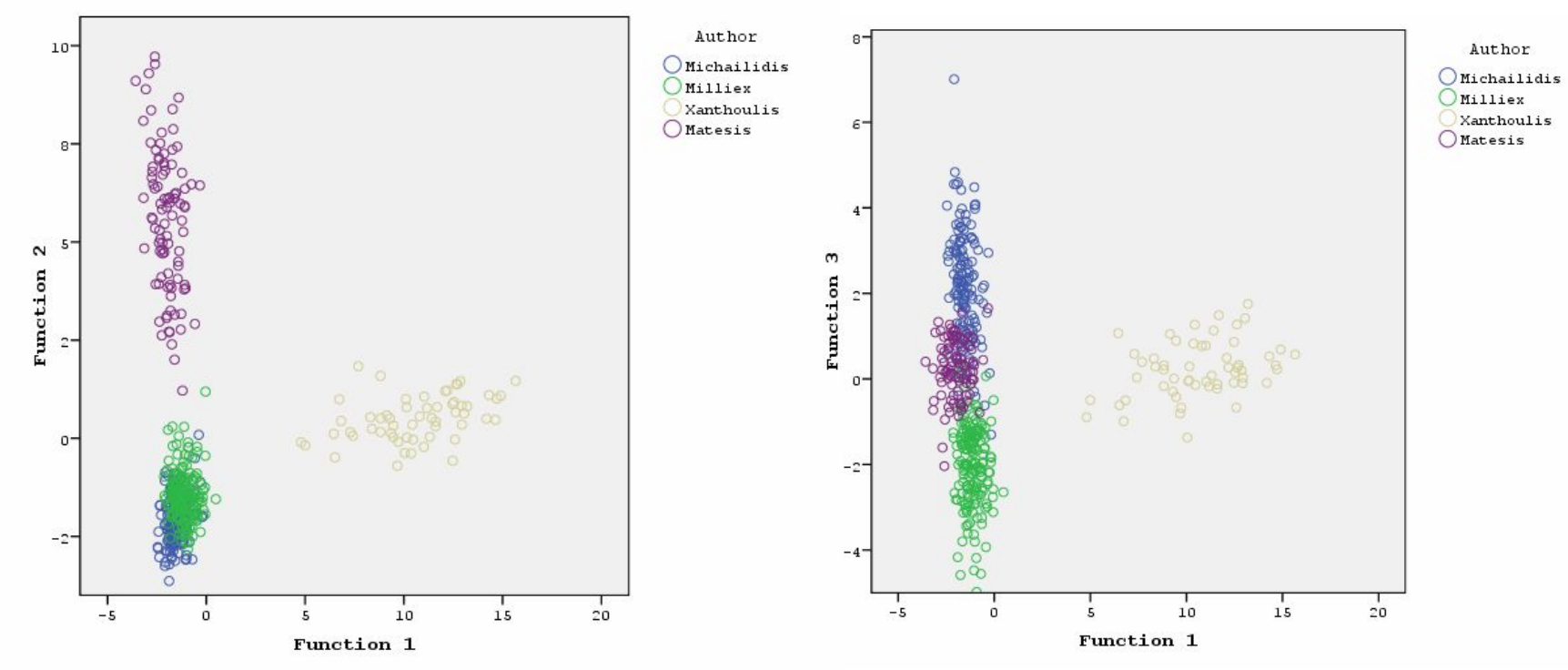
# Validation

- In order to validate the DFA results we used 2 different methods:
  - ***U*-method**: it is based on the "leave-one-out" principle (Huberty, Wisenbaker, Smith, 1987). Using this method, the discriminant function is fitted to repeatedly drawn samples of the original sample. Estimates k-1 samples, eliminating one observation at a time from a sample of k cases.
  - **Test novel**: For one author (Milliex) we used a second novel, not included in the training data, in order to evaluate the classifier's accuracy in unforeseen data from the same author. The specific test resembles more closely in real life authorship attribution problems.

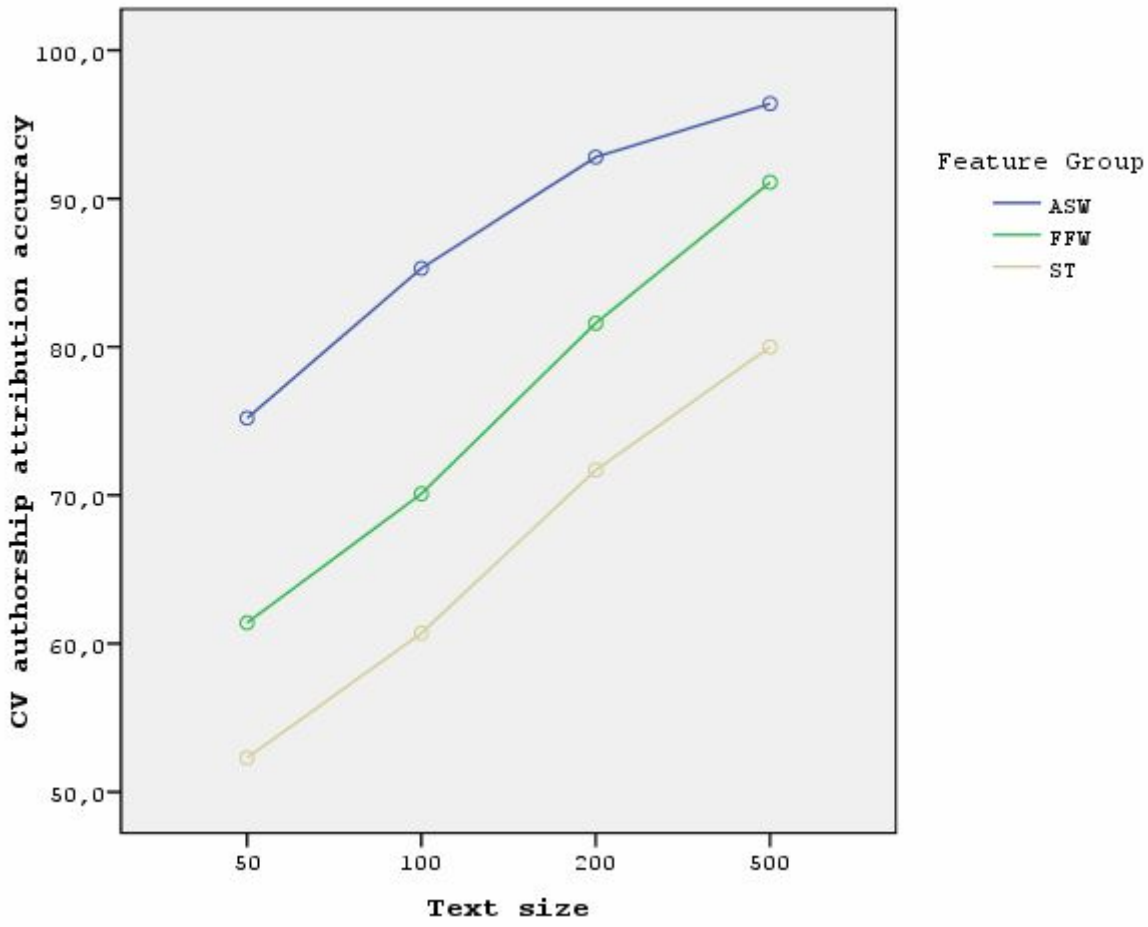
# Best classification confusion matrix (ASW method, 500 words samples)

| | | Matesis | Michailidis | Milliex | Xanthoulis | Total |
|---|---|---|---|---|---|---|
| Count | Matesis | **92** | 0 | 3 | 0 | 95 |
| | Michailidis | 0 | **134** | 11 | 0 | 145 |
| | Milliex | 0 | 2 | **172** | 0 | 174 |
| | Xanthoulis | 0 | 0 | 2 | **55** | 57 |
| % | Matesis | **96,8** | 0 | 3,2 | 0 | 100,0 |
| | Michailidis | 0 | **92,4** | 7,6 | 0 | 100,0 |
| | Milliex | 0 | 1,1 | **98,9** | 0 | 100,0 |
| | Xanthoulis | 0 | 0 | 3,5 | **96,5** | 100,0 |

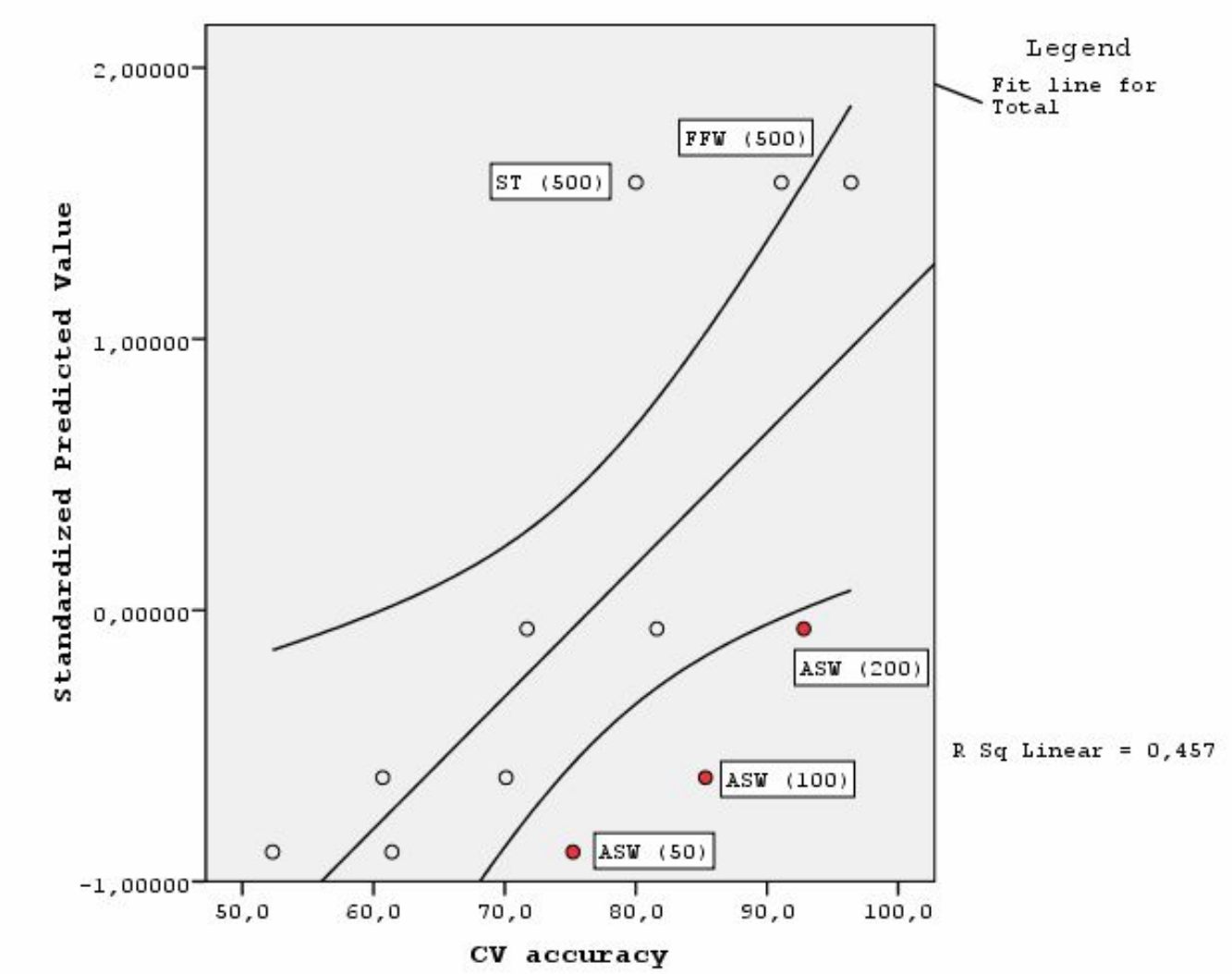
# Best classification (ASW method, 500 words samples)



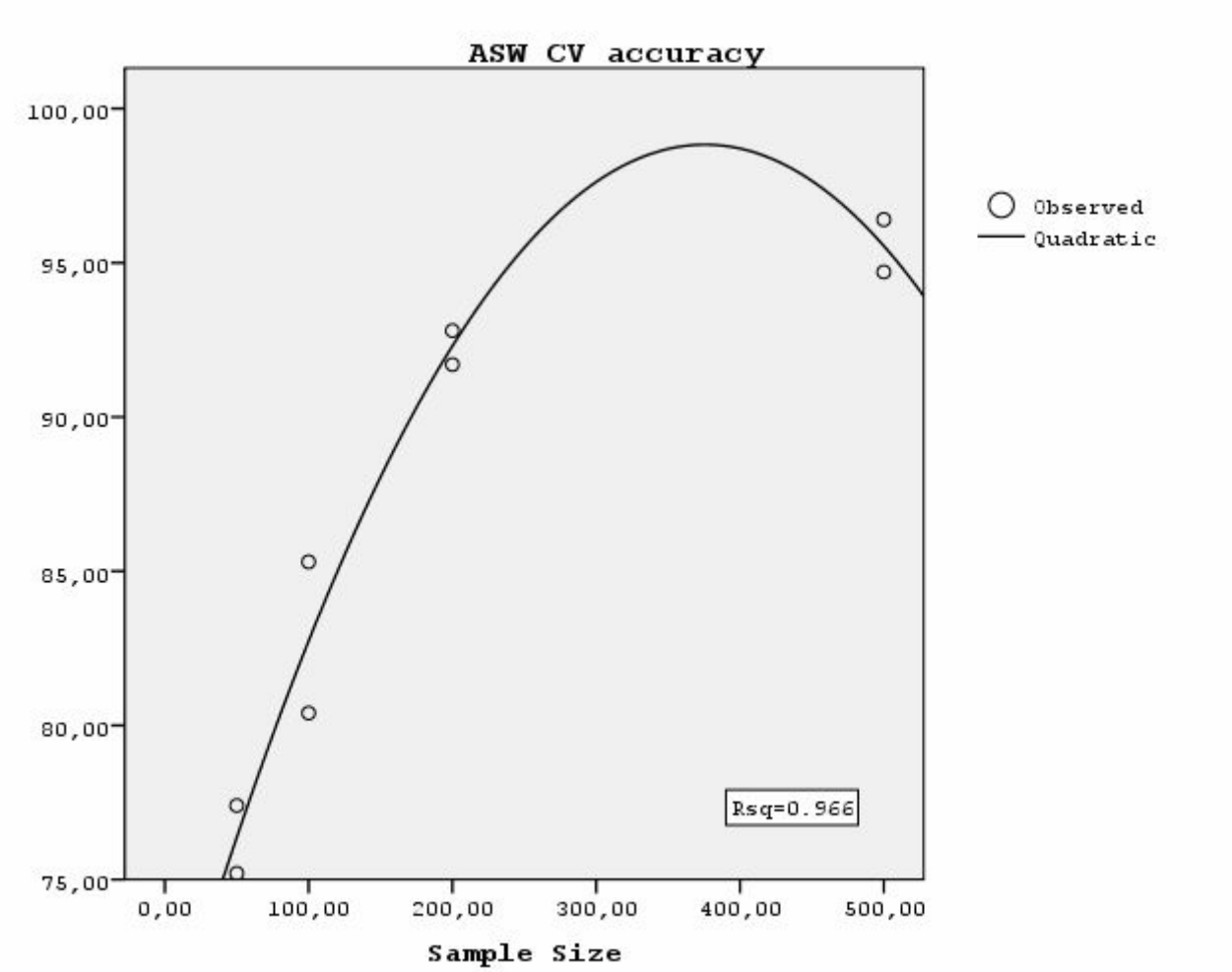
# Comparison of the feature sets over different text sizes



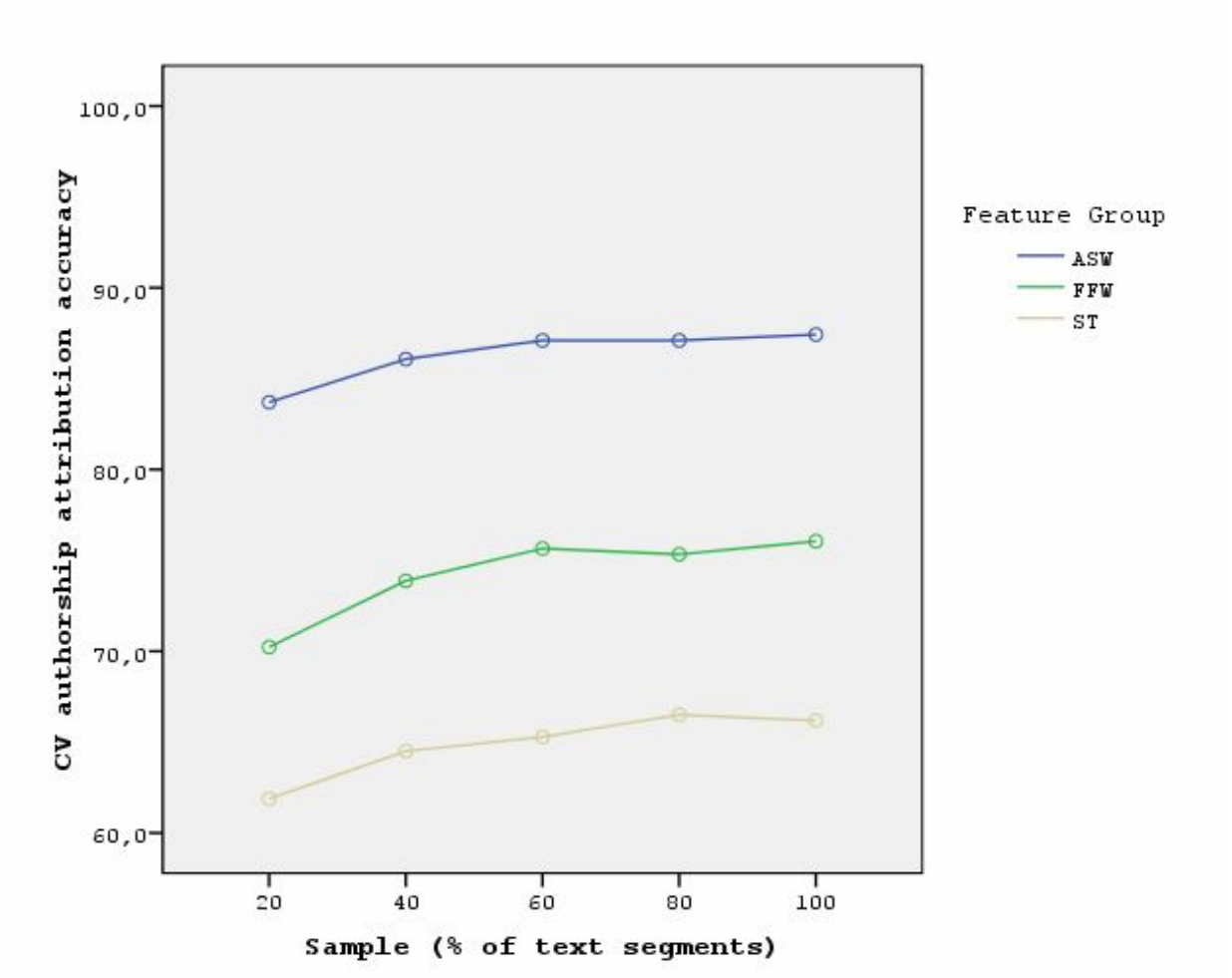
# Relation of classification accuracy and text size



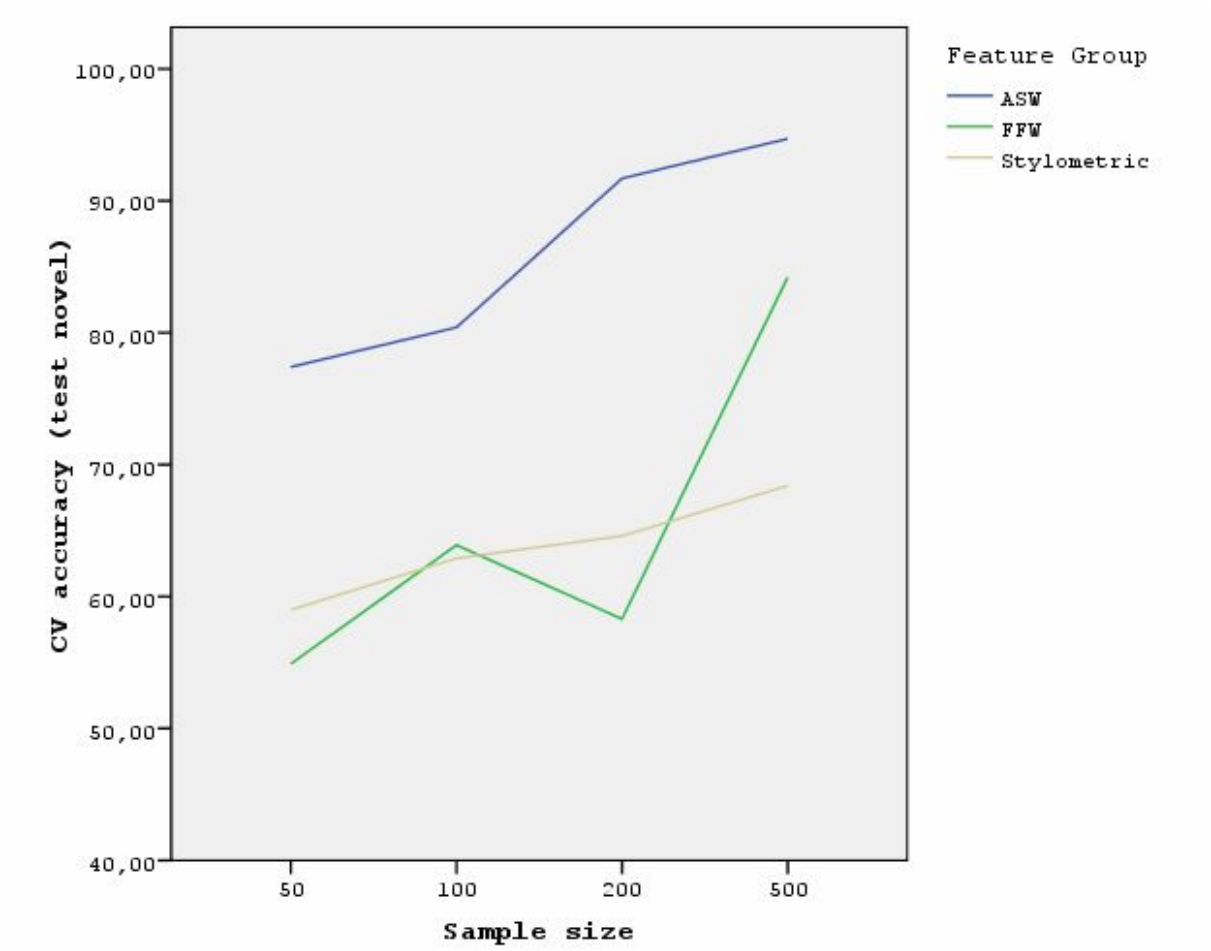
# Relation of ASW classification accuracy and text size



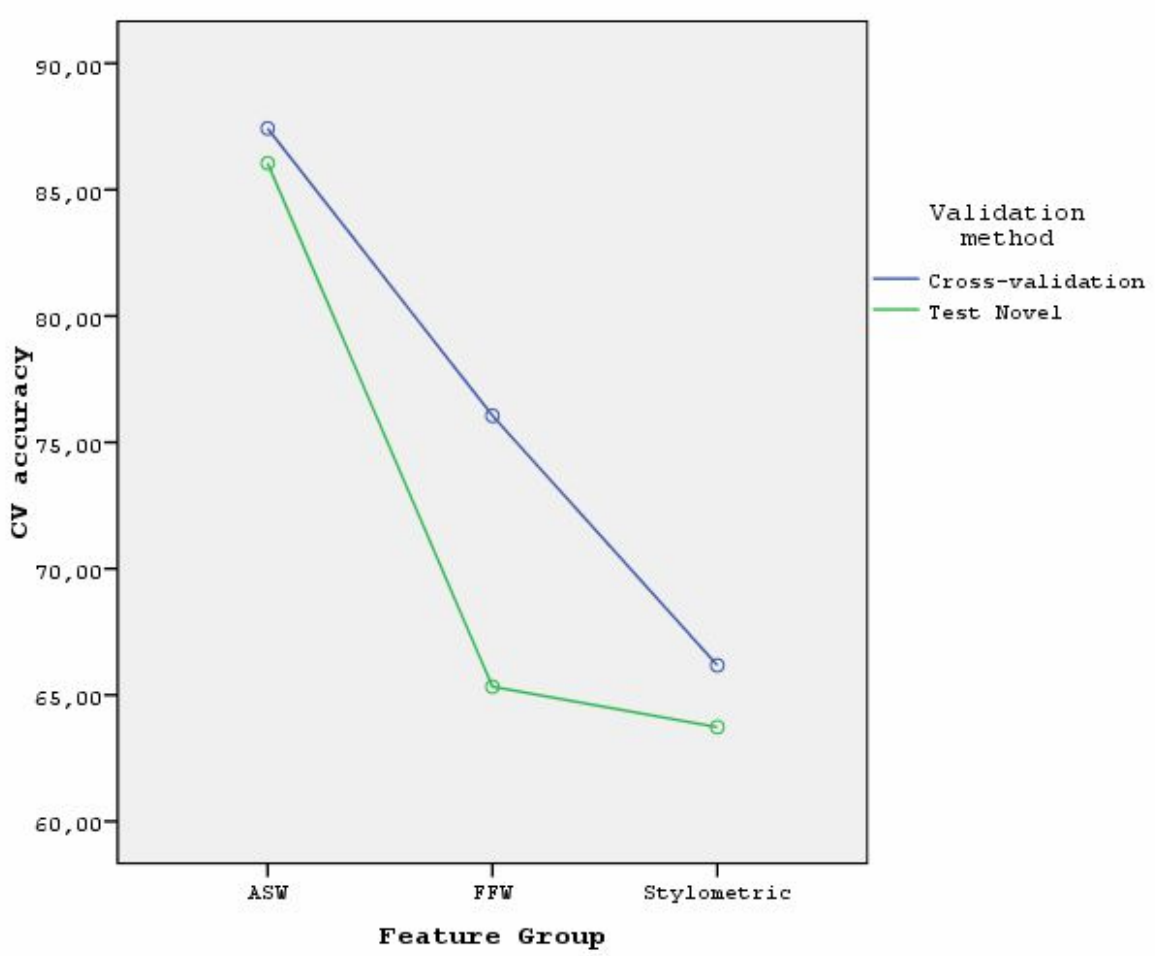
# Comparison of the feature sets over different sample sizes



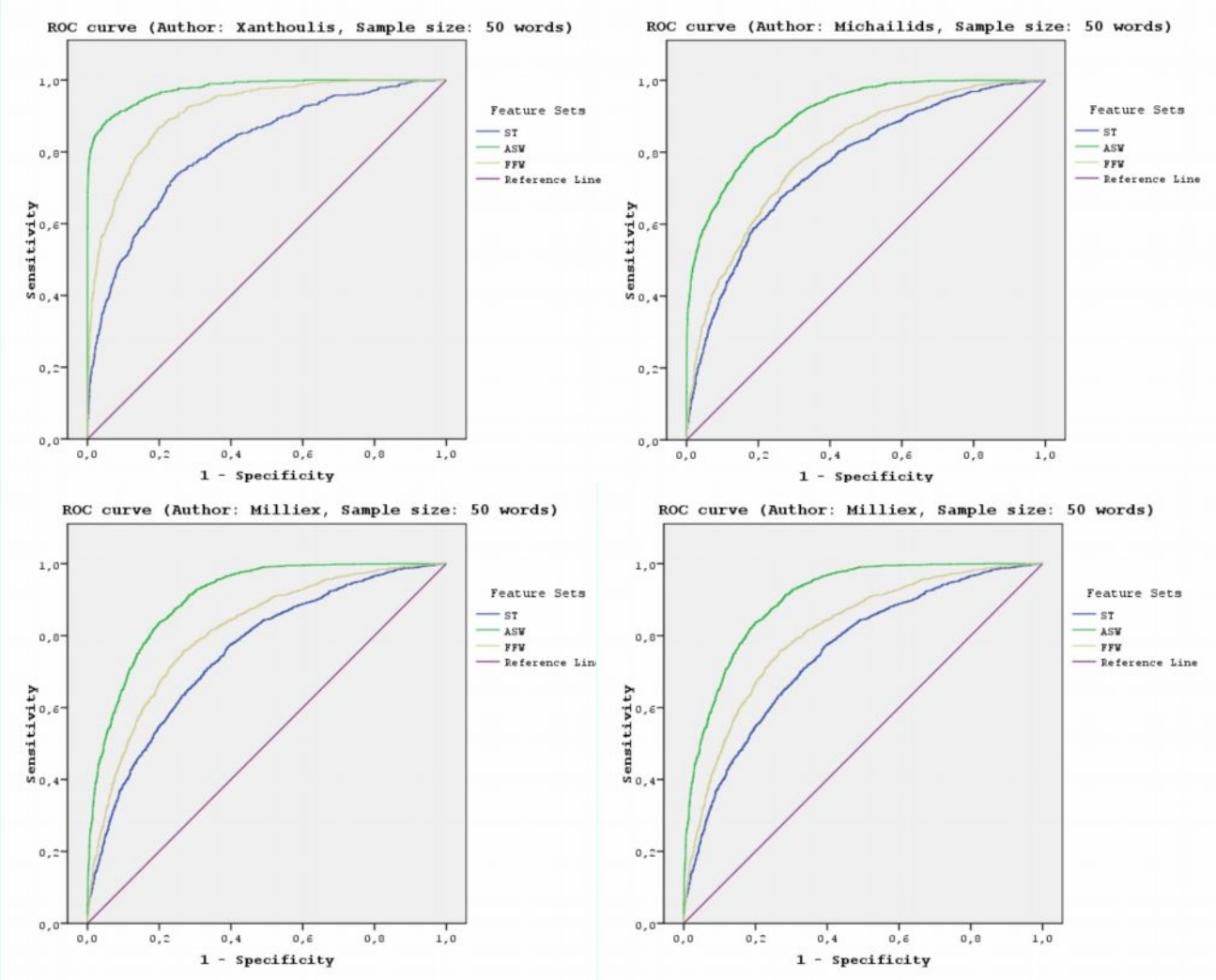
# Comparison of the feature sets over different text sizes in test novel



# Comparison of the validation methods over the three feature sets



# ROC curves for each author in 50 words samples



# Conclusions

- Authorship "genome" exists even in small text samples of 50 words especially if we take into consideration the frequency patterns of content words.
- The size of a text segment exhibits linear correlation with the precision of authorship attribution. However, if we examine only the ASW features the best fit is obtained through a quadratic function.
- Variations in sample size (from 20% to 100%) didn't affect the authorship attribution accuracy in a statistically significant way.
- ASW method outperforms the FFW and ST methods in all experimental conditions and performs especially well in small text sizes.

# Acknowledgments


- The project is co-funded by the European Social Fund and National Resources – (EPEAEK II) PYTHAGORAS