AUTHORSHIP ATTRIBUTION IN MODERN GREEK NEWSWIRE CORPORA

George K. Mikros
Department of Italian and Spanish
Language and Literature
University of Athens
Panepistimioupoli Zografou - 15784
Athens, GREECE
+30 210 6511344
gmikros@isll.uoa.gr

ABSTRACT

The aim of this research is to describe tools and methods applied in the stylistic analysis of Modern Greek newswire corpora. We present a new method of selecting Author-Specific Words that can be used as reliable authorship discriminators. Furthermore, we explore the impact of different kinds of various other stylometric variables using Discriminant Function Analysis.

Keywords

Authorship Attribution, Stylometry, Discriminant Function Analysis, Lexical Frequency Profiling Methods, Keyword Extraction.

1. WHAT IS A MEANINGFUL "KILLER APP" FOR STYLISTIC TEXT ANALYSIS?

Text style is usually defined as the way we shape our linguistic messages. The perception of stylistic choices is often vague since they function in many linguistic levels simultaneously. For this reason, text style analysis requires a simultaneous investigation of the usage of many different linguistic units at the same time.

The development of a useful and feature-rich stylometric application should incorporate the above prerequisite. Design principles which are essential in this kind of endeavour are: multilinguality, modularity and interoperability. Below, we will try to describe some ideas regarding the functional characteristics of such an application.

The proposed application should have four modules which will cooperate in order to provide effective stylistic analysis for various possible application areas. More specifically:

1. Corpus creation and management module:

One of the most striking problems in stylometry studies is the lack of homogeneity of the corpora examined [1]. This means that any application which will be used for exploring stylistic variation should incorporate a flexible corpus management module. In order to utilize the existing general language corpora the application should have the ability to create virtual sub corpora based on different user-defined selection criteria. Highly homogeneous corpora could be created on user demand based on medium, genre, topic, date and other text metadata. In addition, each text could be further subdivided into smaller entities and virtual corpora could be created based on smaller textual units like paragraphs, chapters, or even user-defined text portions.

Stylistic variation can be effectively used for clustering online documents and improve information retrieval. For this reason, and in order to utilize the vast amount of online texts, a tool for harvesting the web and creating corpora based on specific design criteria should be included.

2. NLP module: As many stylometric studies have suggested, the stylistic profile of a text can be modelled effectively by taking into account its linguistic structure. For this reason, we need a strong NLP module, capable of handling multilingual texts and which will incorporate accurate tokenizer, Part of Speech tagger, parser, lemmatizer, Named Entity recognition tool etc. The linguistic information should be encoded for each text or other user-defined textual portion and be available to all the other modules of the application.

3. Feature finding and counting module:

The feature finding and counting process should be extensible in as many linguistic levels as possible containing at least the sub-domains of phonology, morphology, syntax, vocabulary and discourse.

These features should be counted in all the sub corpora created by the user and their frequency should be dynamically encoded in the document - feature space. Furthermore, most features should be counted in a multidimensional way, producing relative frequencies of a feature related to the presence of other features which belong to the same or different linguistic level (i.e., the frequency of a specific, graphemic variation in functional words, the frequency of a word in the beginning of the sentences etc.).

Another significant aspect which could be incorporated is feature concordance and collocation. By examining the hits of a specific feature we could explore its linguistic environment and count the neighbouring features. In addition, many features which cannot be automatically counted could be manually annotated and their frequency could be added to the rest of the research variable.

4. Statistical analysis module: The document – feature datasheet should be used to train various statistical and machine learning algorithms. The user could run multiple classification and clustering tasks varying experimentation with different algorithms and their parameters. The results should be cross-validated (using standard procedures such as 10 – fold, Leave one out etc.) and could also be subjected to human evaluation. Documents could be plotted in multidimensional spaces according to various dimension-reducing techniques (PCA, Factor Analysis etc).

2. TOOLS USED FOR STYLISTIC ANALYSIS OF MODERN GREEK TEXTS

A number of the above mentioned modules have already been developed in order to analyse stylistic variation in Modern Greek texts:

1. Corpus Management and Feature counting:

This tool aims to offer an integrated suite for managing and counting features in corpora. The corpus management module needs raw text files and a corpus description file which includes the path of the corpus and the metadata of the text files. In order to create subcorpora based on specific metadata values, we need to define the appropriate column which contains the specific values. Then we select the "Subcorpus creation" function and we have a breakdown of the number of the texts and their size regarding the chosen metadata.

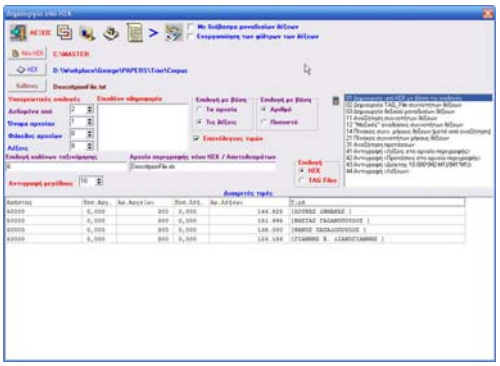

Figure 1: Subcorpora selection window.

In the above example we have a breakdown of our initial corpus based on author information. In the above screenshot (Figure 1) for example, we can adjust the values of Text Size and Number of Texts in order to create a balanced subcorpus in terms of text size or number of texts.

2. Minotavros, (Tool for creating corpora from the Web):

"Minotavros" is a tool which can create a corpus from any website if the user provides its structure. The collected texts along with their metadata are stored in folders in an automated way depending on user-defined criteria (e.g. the author of the text, the size of the text etc.).

The texts that the user gets from every downloaded web page can be automatically transformed to raw text files if the user selects to remove html/xml tags.

3. Ellogon (NLP platform): Ellogon is a multi-lingual, cross-platform, general-purpose language engineering environment, developed from the Institute of Informatics and Telecommunications, NCSR "Demokritos" in order to aid researchers in computational linguistics that produce and deliver language engineering systems [3]. Ellogon offers an extensive set of facilities, including tools for processing and visualizing textual/HTML/XML data and associated linguistic information, support for lexical resources (like creating and embedding lexicons), tools for creating annotated corpora, accessing databases, comparing annotated data, or transforming linguistic information into vectors for use with various machine learning algorithms. Ellogon belongs to the category of referential or annotation based platforms, where the linguistic information is stored separately from the textual data, having references back to the original text. Based on the TIPSTER data model, Ellogon provides infrastructure for:

- Managing, storing and exchanging textual data as well as the associated linguistic information.
- Creating, embedding and managing linguistic processing components (Figure 2).

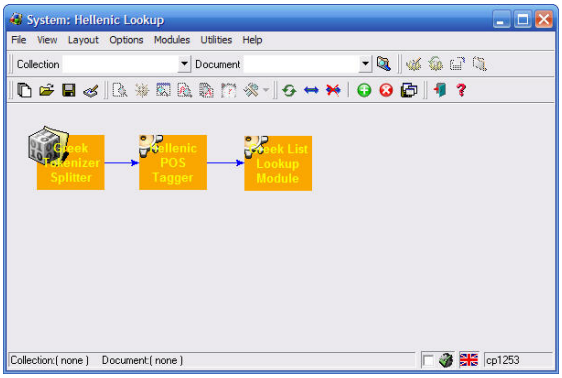

Figure 2: Connecting linguistic processing components

- Facilitating communication among different linguistic components by defining a suitable programming interface (API).
- Visualizing textual data and associated linguistic information.

The creation of Ellogon components can be easily done through the Ellogon GUI which supports C++ and Tcl [3]. The Ellogon GUI offers a specialized dialog where the user can specify various parameters of the component he/she intends to create, including its pre/post-conditions. Then Ellogon creates the skeleton of the new component that will handle all the interaction with the Ellogon platform. If the language of the component is C++, a Makefile for compiling the component under Unix will also be created. Besides creating a skeleton, Ellogon tries to facilitate the development of the component by allowing the developer to edit the source code and re-load the specific component into Ellogon from its GUI.

3. Methodology

3.1 The authorship corpus

In order to explore quantitatively authorship "fingerprints", we devised an experimental methodology similar to [4] and [5]. Instead of stimulating writing for a specific topic we collected already published articles from four authors in a high circulation Greek newspaper "TA NEA" using "Minotavros". In order to achieve high levels of homogeneity the collected texts satisfied the following criteria:

- Same newspaper (TA NEA), which ensures common normalization conventions and common ideological background and attitudes towards Modern Greek language variation.
- Same topic (Internal Politics) and genre (Informational Texts).

- Same number of articles for each author (300).

In total, 1200 texts were written by four different authors with tradition in journalistic writing and stable presence, in the last decade, in the political column of the specific newspaper.

The specific corpus aims to form a difficult challenge for stylometric analysis. It is highly homogeneous regarding the sociopragmatic dimensions of the linguistic production involved and additionally contains small size texts which are untypical of most texts used in stylometry. In particular, 41.8% of the texts have less than 500 words and this poses a further difficulty in the authorship attribution since most stylometric variables exhibit authorship quantitative patterns in larger text sizes [6].

3.2 Stylometric variables

We used different sets of variables in order to enhance authorship attribution, which involve:

1) Lexical "richness" variables (Yule's K, Standardized TTR, Lexical Density, Percentage of hapax and dis-legomena, Ratio of Dis- to Hapax legomena, Relative entropy) - 7 variables.

2) Character level measures (Frequency of the letters and punctuation) - 38 variables.

3) Word level measures (Average word length per text in letters, Word length distribution, Part of Speech frequency and ratios) - 29 variables.

4) Sentence level measures (Average length of sentences in words, Standard deviation of sentence length per text) - 2 variables

5) 80 most Frequent Function Words – 80 variables.

6) Diglossia (Kathareyousa [K] and Dimotiki [D]) related variables (Noun Endings in "–is" [D] / "-eos" [K] and the relative percentage of them, Relative pronouns "pou" [D]/ "opoios" [K] and the relative percentage of them, Relative percentage of D and K prepositions) – 6 variables.

7) 80 most distinctive Author-Specific Words – 80 variables.

The first four sets of variables (lexical "richness", character, word, sentence level measures and most Frequent Function Words) have been extensively used in stylometric studies and their discriminatory power has been well documented. In addition, we have devised two other sets of variables which can enhance authorship attribution.

The first (number 6 in the above variable list) is language specific and relates to the diglossia problem of Modern Greek language, i.e., the parallel co-existence of two forms of written varieties (Dhimotiki [D] and Kathareyousa [K]) with different phonological, morphological and syntactical rules.

The second (number 7 in the above variable list) is language independent and relates to the detection of suitable lexical variables that can maximize the authorship discrimination.

In the present study, we will use the 80 most Frequent Function Words (FFW) of the corpus. Furthermore, we will introduce a new method for automatically selecting the most characteristic Author-Specific Words (ASW) and will compare their discriminatory power to the standard most Frequent Function Words methodology.

This method has been applied previously in automatic text categorization [7] and has been proven superior to any other lexical selection method. It is based on frequency profiling and has already been used in English for different research purposes ([8], [9]). The procedure we propose is explained briefly as follows (Figure 3):

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.

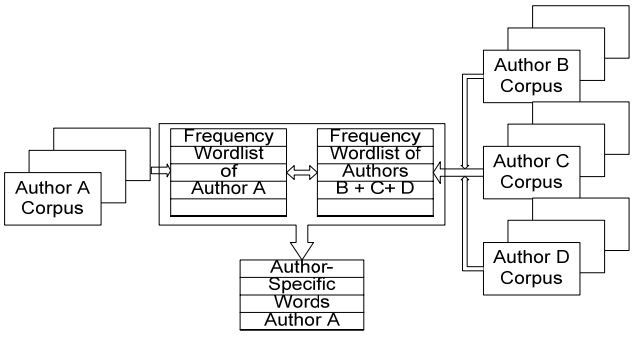

Figure 3: The ASW methodology: Extracting ASW for the first of the four candidate authors.

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. Since the texts were unequal in size we normalized ASW frequency in 1000 word text size.

4. RESULTS

4.1 ASW vs. FFW

The statistical model we used was Discriminant Function Analysis (DFA) and was calculated using different sets of the above mentioned stylometric variables. The first experiment aimed to point the usefulness of the ASW methodology compared to the "bag of words" approach in lexical variables selection. For this reason we conducted a number of DFA varying our sample size (number of texts per author) with a step of 50 texts each time and calculated the cross-validated classification precision for both methods (ASW and FFW). Furthermore, we conducted a series of DFA varying the number of the lexical variables included in the analysis with a step of 5 words per author. The comparison of both methods in sample size and words per author experiments is shown in the following figure (Figure 4):

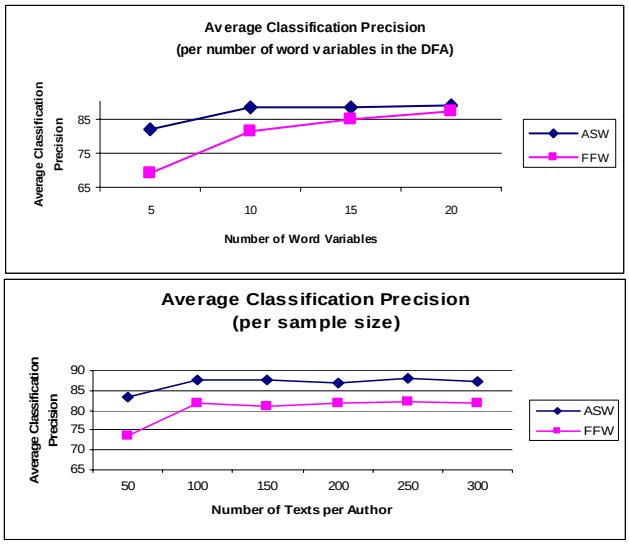

Figure 4: Comparison of author classification using ASW and FFW

The above graphs reveal that ASW outperforms FFW in all conditions and achieves high classification precision even with small samples of training sets and few words per author (which is the most probable situation in most real-life authorship attribution problems). Its performance is stabilized in 100 texts and 10 words per author. Even when we use the whole training corpus (300 texts and 20 words per author), FFW achieves 81.7% average authorship classification precision which is still lower than the 83.4% precision achieved by ASW with the minimum settings of the experiment (50 texts and 5 words per author).

4.2 Relative importance of the stylometric variables

The best authorship classification precision (95%) was obtained using all the stylometric variables including ASW and excluding FFW. The classification plots based on the combination of the 3 discriminant functions produced by our model are shown in

Figure **5**:

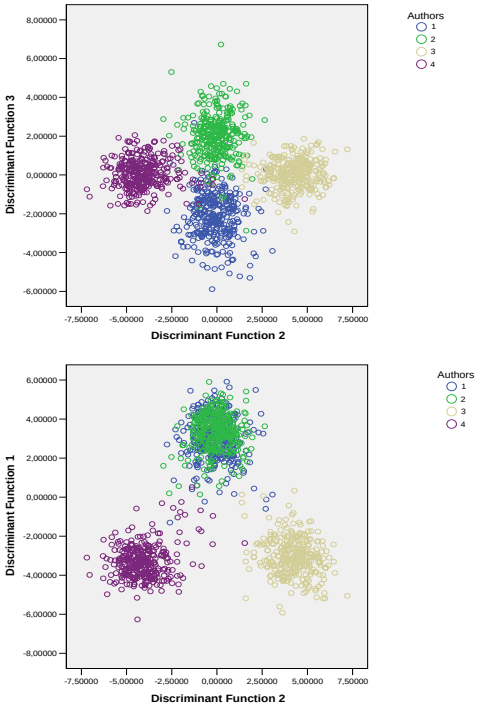

Figure 5: Scatter plot of DF 1 - 3

The above scatter plots display the discriminatory power of each of the 3 discriminant functions (DF) in relation to the 4 authors. DF1 discriminates Author 3 from Author 4, DF2 discriminates Authors 3 & 4 from Authors 1 & 2, and DF3 discriminates Author 1 from Author 2.

In order to investigate further the contribution of the stylometric variables in the discrimination task we rotated the canonical discrimination functions using VARIMAX. The rotated structure and the five strongest correlations of each variable with the 3 Discriminant Functions can be found in Table 1:

Table 1: The rotated structure matrix which contains the five strongest within-group correlations of each predictor variable with each DF.

Stylometric Variables	DF1	DF2	DF3
Frequency of Full stops	0.4963	0.1617	0.0602
Frequency of Nouns	-0.2651	0.0821	-0.0417
ASW (no 41)	0.2270	0.0891	0.0273
ASW (no 21)	-0.2258	0.1354	-0.1897
Frequency of Adverbs	0.2009	0.1555	-0.0721
ASW (no 61)	-0.0138	-0.2568	0.0550
% Hapax legomena	0.2314	-0.2430	-0.0545
Sentence length	-0.1209	0.2359	0.1084
Ratio of Dis to Hapax legomena	-0.2011	0.2035	0.0776
Frequency of letter "]"	0.1309	-0.1879	0.0457
Frequency of Pronouns	-0.0076	0.0883	0.2252
ASW (no 4)	-0.0433	-0.0091	0.2159

ASW (no 26)	-0.0889	0.0896	-0.2146
ASW (no 23)	-0.0876	0.0870	-0.2005
Frequency of Diglossia related word "opoios"	-0.1177	0.0934	0.1719

The above data reveal that stylistic information is conveyed simultaneously in many linguistic levels. The 5 most discriminating variables per DF belong to different levels of linguistic description. The importance of ASW is manifested clearly since appear systematically in all three DF. Furthermore, sociolinguistic information seems to be useful for stylistic analysis, since it appears as the 5th most discriminating variable in the DF3, the DF which discriminates Author 1 from Author 2.

The above results lead us to the conclusion that any stylistic application which wishes to address more general issues regarding the linguistic variation and its communicative functions should be able to retrieve information from the whole spectrum of linguistic structure. Authorship "genome" is highly idiosyncratic, and each author's style is different from another's in an unpredictable way. In order to have a precise authorship attribution in a large set of possible authors, we should use the most extensible list of stylometric variables we can obtain since we don't know a priori which variables are useful to the discrimination of the specific authors set.

5. CONCLUSIONS

In the present study we used a number of tools and evaluated different methods and variables in authorship attribution using Modern Greek newswire corpus. We developed ASW extraction, an automated process which can retrieve Author-Specific Words, and can increase the authorship attribution precision in contrast to other lexical methods (Frequent Function Words). This method is language independent and can be used as a generic method of keyword extraction applied also in text categorization [7].

In addition, by examining the diglossia variables, we investigated the extent to which sociolinguistic variation incorporates in stylistic variation. Our results reveal that this kind of linguistic variation carries stylistic information and can be exploited in order to enhance authorship attribution.

The precision in authorship attribution using our DFA model was 95%. The specific result, taking into consideration the high homogeneity of the studied corpus, is considered satisfactory.

6. ACKNOWLEDGMENTS

The present study has been funded and published in the framework of the research programme **PYTHAGORAS I** which is co-funded by the European Social Fund (75%) and National Resources (25%) - Operational Program for Educational and Vocational Training II (EPEAEK II).

7. REFERENCES

[1] Rudman, J. The State of Authorship Attribution Studies: Some Problems and Solutions. *Computer and the Humanities, 31,* (1998), 351–365.

[2] Koutsis, I., Kouklakis, G., Mikros, G., and Markopoulos, G. MINOTAVROS: A tool for the semi-automated creation of large corpora from the Web. *Proceedings from The Corpus Linguistics Conference Series* (Birmingham, UK, 14-17 July 2005), Vol. 1, 2005. Available in http://www.corpus.bham.ac.uk/PCLC/minotavros.doc.

[3] Petasis, G., Karkaletsis, V., Paliouras, G., Androutsopoulos, I., and C. D. Spyropoulos, C. D. Ellogon: A New Text Engineering Platform. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002),* (Las Palmas, Canary Islands, Spain, May 2002). 2002, vol. I, 72 – 78.

[4] Baayen, H., van Halteren, H., Neijt, A., Tweedie, F. An experiment in authorship attribution. *Proceedings of JADT 2002* (St. Malo 2002). 2002, 29-37.

[5] Van Halteren, H., Baayen, R.H., Tweedie, F.J., Haverkort, M. and Neijt, A. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics, 12,* (2005), 65–77.

[6] Ledger, G., and Merriam, T. Shakespeare, Fletcher, and the Two Noble Kinsmen. Literary and Linguistic Computing, 9, (1994), 235-248.

[7] Mikros, G. Statistical approaches in automatic text categorization of Modern Greek texts: a preliminary investigation of stylometric variables and statistical methods [in Greek]. In *6th International Conference of Greek Linguistics* (Rhethimno, Greece, 18-21 September 2003) available in CD-ROM, 2003.

[8] Rayson, P., Leech, G. and Hodges, M. Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics, 2* (1997), 133 - 152.

[9] Granger, S. and Rayson, P. Automatic profiling of learner texts. In S. Granger (ed.), *Learner English on Computer*, Longman: London and New York, 1998, 119-13.