STYLOMETRIC PROFILING OF THE GREEK LEGAL CORPUS

Georgios Brousalis
University of Athens, Greece
gbrousalis@phil.uoa.gr

Georgios Mikros
University of Athens, Greece
gmikros@isll.uoa.gr

Georgios Markopoulos
University of Athens, Greece
gmarkop@phil.uoa.gr

ABSTRACT

This paper deals with the application of corpus analysis to Greek law texts in order to illustrate their stylometric profile. Due to some vagueness or rigidity elements, these texts are often criticised for constituting an opaque "legalistic" language (Tiersma 1999: 139-41). For the purposes of this study we designed a Greek Legal Corpus which we then juxtaposed to the Hellenic National Corpus[1]*. Using criteria, such as lexical "richness", word and sentence length, and part-of-speech frequencies, we look for linguistic features which may affect the precision and comprehensibility of the legal language (Bhatia 2010).*

Keywords: corpus processing, stylometry, text classification, legal language, statistical analysis, POS tagging

1. Introduction

The general aim of this research is to study the possibilities of applying corpus analysis to legal texts. We focus especially on the language of Greek legislation and its stylometric profile. After classifying some general properties and lexico-grammatical features of the laws, we create and process a corpus of Greek law texts. Before referring to the process and findings, some useful distinctions should be drawn.

1.1 Definitions

Many terms are used when referring to legal discourse, depending on both the scientific aspect and the corresponding legal system. In this paper, we adapt the following distinctions:

Forensic texts: According to Olsson (2004: 5) any text or item of spoken or written language is potentially a forensic text, provided it is somehow implicated in a legal or criminal context. In this respect, even a parking ticket could become a forensic text. Therefore, the term "forensic" has a much wider sense than what we are interested in.

Language of the Law: As Gibbons suggests (2003: 15) "the language of the law can be broadly divided into two major areas – the codified and mostly written language of legislation and other legal documents, such as contracts […] and the more spoken, interactive and dynamic language of legal processes […]".

Legal language: the formal, formulaic, rigid, "legalistic" language of legal documents, such as statutes, court decisions, contracts, wills, etc., very close to what Tiersma calls "operative legal documents" (1999: 139).

Legislative texts: the written sources of a legal order, which in our case is the Greek legal order.

[1] http://hnc.ilsp.gr/

In Z. Gavriilidou, A. Efthymiou, E. Thomadaki & P. Kambakis-Vougiouklis (eds), 2012,
Selected papers of the 10th ICGL, pp. 167-176. Komotini/Greece: Democritus University of Thrace.

1.2 The Greek legal order

“Legal order” refers to the aggregation of all the written and non-written rules which regulate the external behaviour of the members of a community. With respect to the Greek legal order, the non-written rules correspond to the mores and customs of the Greek community, while the written legal sources constitute a quite complex, but hierarchically organised system of laws.

In particular, the written part of the Greek legal order consists of the so-called “special Law”, i.e. the Greek Constitution, the European Union Law and the International Law, as well as the “typic” and the “substantial” laws, which constitute the “common Law”. In this project we have excluded EU and International Law, because our research interest is restricted to untranslated texts originally redacted in Greek.

2. Properties of the legal language

2.1 The desirable characteristics

Researchers representing different scientific fields and studying different legal systems, conclude that there is a set of properties which legal language should present. Among them, Knapp (1991: 4-10) listed some properties which should be observed in the legal language and notions: accuracy, consistency, discernibility, unemotionality, intelligibility, unambiguity, constancy. Recently, Bhatia (2010: 38) suggested some desirable characteristics of legislative language, which are clarity, precision, unambiguity and all-inclusiveness, defined in terms of comprehensibility, accessibility and transparency. Similarly, Panaretou (2009: 58) has observed a “legal paradox”: the laws must be precise, clear and unambiguous, but, at the same time, general, widely applicable and inclusive. In order to combine the characteristics listed above, the legal texts may result in being either rigid or vague, or rather both.

2.2 Lexico-grammatical features of the laws

The rigidity or the vagueness, which often characterises the language of legislative texts, is usually related to some linguistic properties of these texts, namely the lexical and grammatical choices made by the authors. Tiersma (1999: 203-10) and Panaretou (2009: 75-119) have made a thorough description and classification of such features listed as follows:

- **Technical vocabulary**, e.g. *ενάγων* (= plaintiff); *λιπομαρτυρία* (: the offense of defaulting witness); *ελευθεροκοινωνία* (= pratique); *υπερθεματιστής* (= tenderer); *καταπίστευμα* (= trust); *αναιρεσείων* (= appellant).
- **Archaic, formal, formulaic, unusual words and forms**, e.g. *ανήκεστος βλάβη* (= irreparable damage); *ο κρινόμενος* (<participle> “the judged”); *μονιμοποιητέος* (<verbal adjective> “someone (employee) who must become permanent, the *permanent-able”).
- **Binomials – multinomials**, e.g. *οι διάδικοι ή οι νόμιμοι αντιπρόσωποί τους ή οι δικαστικοί τους πληρεξούσιοι* (= the parties or their legal representatives or their judicial attorneys); *πραγματικά ή νομικά ελαττώματα* (= actual or legal defects); *οι ενέργειες και οι παραλείψεις* (= the actions and the omissions).
- **Impersonal constructions and overuse of passives** (in typic laws), e.g. *κυρώνεται* (“something is ratified”); *απαγορεύται* (“it is prohibited”); *υποχρεώνεται* (“he/she is obliged”).
- **Preference to nouns instead of verbs & overuse of verbal nouns** (combined with nominalisation) e.g. *διαλειτουργικότητα* (= interopability), *πραγματοποίηση* (= realisation), *απασχολησιμότητα* (= employability).
- **Modal verbs**: πρέπει (“must, shall”, normally expressing deontic necessity), **μπορεί**: (“may, can”, normally expressing deontic possibility, but also expressing epistemic possibility only when describing the facts).
- **Long and complex sentences**
- **Overuse of the conjunctions**: *και*, *ή*, *είτε...είτε* (: “and”, “or”, “either…or”), e.g. “Για τους σκοπούς της παρούσας σύμβασης, ο όρος διεθνής μεταφορά σημαίνει οιαδήποτε μεταφορά στην οποία με βάση τη συμφωνία μεταξύ συμβαλλομένων μερών, ο τόπος αναχώρησης *και* ο τόπος προορισμού, ανεξαρτήτως εάν υπάρχει *ή* όχι διακοπή της μεταφοράς *ή* μεταφόρτωση, βρίσκονται *είτε* εντός των εδαφών δύο συμβαλλομένων κρατών, *είτε* εντός του εδάφους ενός και μόνον συμβαλλομένου

κράτους…"(N. 3006. K.No.B. 50 (2002): 789).

Our research was based on quantifying some of the above features and using them in order to define the stylometric profile of the Greek Legal Corpus.

3. Corpora and Features

The aim of this study can be summarised in three different but highly related research questions:

1. What is the stylometric profile of the Greek law texts and how can this be compared with the stylometric profile of Greek language corpora?
2. How significant is the impact of each stylometric group in the discrimination of legal texts?
3. Which specific stylometric features related to the legal language could provide reliable genre discrimination?

In order to investigate the abovementioned research questions we used both the Hellenic National Corpus (HNC) and the Greek Legal Corpus (GLC). HNC is currently the biggest written corpus of Modern Greek and consists of many different genres and topics (Hatzigeorgiu et al., 2000) [2]. GLC is a collection of contemporary Greek legislative texts containing:

- the current Greek Constitution
- the codified Civil and Criminal Law

as well as randomly selected balanced samples of

- 'typic' laws
- Presidential Decrees
- Ministerial Decrees
- Decentralised Government Acts

The basic descriptive statistics of the two corpora are displayed in the following table (Table 1) :

Corpora	**Texts (N)**	**Average text length**	**Median**	**St. Deviation**	**Minimum text length**	**Maximum text length**
HNC	45.691	755	494	2.374	10	166.576
GLC	1.594	1.224	991	2.284	504	45.333

Table 1: Descriptive statistics for HNC and GLC

Both corpora underwent Part-of-Speech tagging using the Tree Tagger developed by Schmid (1994). The Tree Tagger is a probabilistic tagger, where transition probabilities are estimated using a decision tree and provides better results than HMM- and Trigram taggers. This tagging technique can achieve up to 96% accuracy on the data. The training of the Greek parameter file, used in our study, is based on a 500.000 words tagged corpus. The data have been tagged for all grammatical categories.

In order to define the stylometric profile of the HNC and GLC, we measured six broad sets of stylometric features which contain both lexical and sublexical units. Each set groups a number of variables which function complementary and all together approximate a specific textual construct. Although the listing is not exhaustive, it contains most of the variables that have been employed in modern stylometric research and we consider them as sociolinguistically neutral. All the features used in this study are the following:

Lexical "richness"

- *Yule's K*: Vocabulary richness index that exhibits stability in different text sizes (Tweedie & Baayen, 1998).

[2] We acknowledge the fact that the HNC developed by ILSP is not a balanced corpus of Modern Greek with sufficient size to reflect linguistic generalizations. However, the aim of this study was to identify specific properties of the Greek legal language and use them for the stylometric profiling of Greek legal texts.

- *Lexical Density*: The ratio of functional to content words frequencies in the text, also known as *Functional Density* (Miranda & Calle, 2007).
- *% of Hapax- and Dis legomena*: The percentage of words with frequency 1 and 2 in the text segment.
- *Dis-/Hapax- legomena*: The ratio of dis legomena to hapax legomena in the text segment, indicative of authorship style (Hoover, 2003).
- *Relative entropy*: It is defined as the quotient between the entropy of the text and its maximum entropy multiplied by 100. Maximum entropy for a text is calculated if we assume that every word appears with frequency 1 (Oakes, 1998: 62).

Word length

- Average word length (per text) measured in letters.
- Word length distribution: The frequency of words of 1, 2, 3 … 14 letters long normalized in 1,000 words sample.

Sentence length

- The average sentence length measured in words.
- The standard deviation of the sentence length

Part-of-Speech frequencies

- The frequency of each Part of Speech tag expressed as percentage of the text size.

4. Statistical analysis

In order to examine in detail the way each of the six variable sets relates to the legal genre, we used *Multiple Analysis of Variance* (MANOVA). MANOVA is a multivariate statistical analysis which is specifically designed to analyze the effect of one or more categorical independent variables on two or more continuous dependent variables. Although the problem could be tackled with multiple univariate tests, the overall a-level error (Type I) will be inflated and the probability to reject the null hypothesis when it is true is increased. MANOVA controls against Type I error and offers an omnibus test of significance that takes into account the effect of the independent variable(s) to all the dependent variables simultaneously (Weinfurth, 1995). Furthermore, MANOVA is particularly useful if the dependent variables are conceptually related and there is a moderate intercorrelation between them. Since each variable group contains stylometric variables that attempt to measure the quantitative expression of a specific textual construct, we expect a certain amount of redundancy. MANOVA takes into account this shared common information and tests the effect of the independent variable(s) in a multivariate way (i.e. taking all dependent variables at once). Another reason why a multivariate approach is preferable in our data is that it can detect differences when groups differ on a system of variables (Huberty & Morris, 1989). MANOVA finds a linear composite of the dependent variables that maximizes the separation of the categories that form the independent variable.

A non-significant MANOVA result means that the specific set of dependent variables examined simultaneously do not differ across the categories of the independent variable and no further analysis should be made. A significant MANOVA however indicates that at least one dependent variable differs significantly across the categories of the independent variable. In the relevant literature most researchers perform univariate tests (t tests in our case) with adjusted alpha level (Bonferroni correction) for each of the dependent variables in order to detect which variable is different between the categories of the independent variable (Hair Jr, Anderson, Tatham, & Black, 1995; Stevens, 2002). This procedure however has been criticized (Bray & Maxwell, 1982; Huberty & Morris, 1989) among others for confusing the univariate with the multivariate research questions. Since genre and language structure interact in complex and multilevel ways we chose to further explore significant MANOVAs with *Discriminant Function Analysis* (DFA). Conducting DFA following a significant multivariate effect allows the researcher to investigate in detail the linear composites of the dependent variables and to determine their structure as well as the weights of each dependent variable (Meyers, Gamst, & Guarino, 2006).

5. Results

5.1 Feature group importance

In our data we performed separated MANOVAs for each one of the six stylometric groups by using the author's gender as the independent variable. The multivariate statistic we calculated was HotellingT^2 which is the multivariate counterpart of the univariate t statistic. Furthermore, partial η^2 was calculated indicating the percentage of the variance explained by the combined dependent variables. The following histogram displays the importance of each feature group based on η^2:

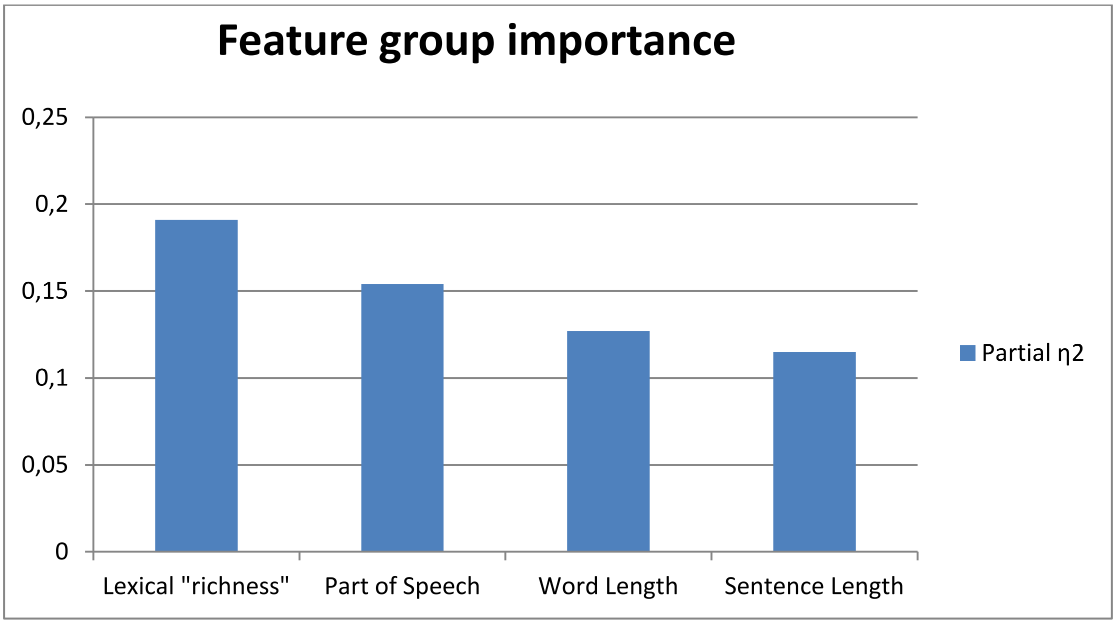

Figure 1 Feature group importance using Partial η^2

All stylometric groups had a multivariate statistically significant effect in the discrimination of the legal texts from the texts belonging to the HNC. The group that accounts for the biggest amount of variance is Lexical "richness" (19.1%) followed by Part of Speech frequencies (15.4%), Word Length (12.7%) and Sentence Length (11.5%). For each of the four variable groups that HotellingT^2 was found statistically significant we performed DFA in order to further explore the linear composite structure and to assess each dependent's variable contribution to legal genre discrimination.

5.2 Lexical "richness" DFA

The Discriminant Function Analysis using Lexical "richness" variables as independent and text genre as dependent showed that 3 variables exhibit a statistically significant difference between the GLC and the HNC. In order to assess the importance of each variable to the genre discrimination, we examined the factor structure coefficients (absolute values), i.e. the correlations between the variables in the model and the discriminant functions. Depending on the coding of the genre factor, coefficients were positive or negative indicating whether a specific variable relates to the legal genre or the general language. The structure coefficients are displayed in the following histogram:

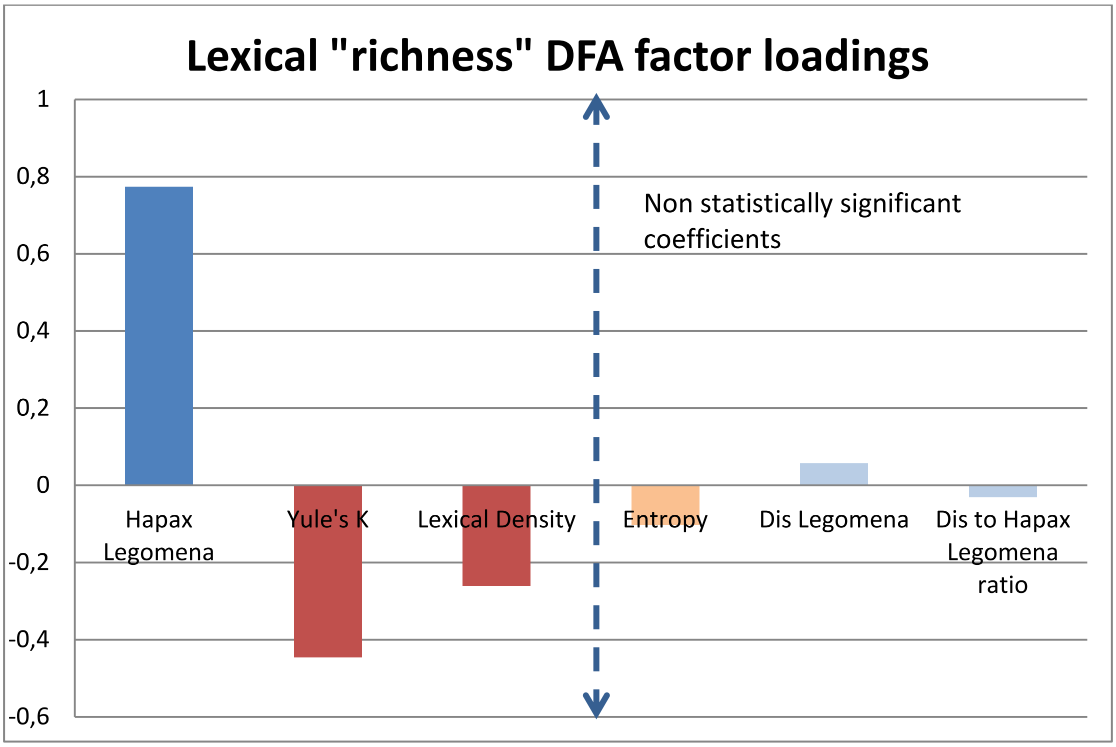

Figure 2 DFA factor loadings using Lexical "richness" features

The inspection of the factor loadings reveals that hapax legomena relate to the general language (positive coefficient) and Yule's K and Lexical Density relate to the legal language. Yule's K is a measure of lexical repetition which increases as the vocabulary of the text tend to be formulaic. Indeed, the mean of Yule's K in HNC is 74.13 while in GLC is 93.38, which means that GLC texts use vocabulary that is repeated frequently. Lexical density is the second variable that can be considered as legal stylometric index. Legal texts have a mean Lexical Density of 1.39 while HNC has a mean of 1.1. From this we can conclude that the legal texts present higher ratio of content to function words than the texts from general language. This higher ratio can be attributed to the high terminology load that characterizes the specific genre. Interestingly, Yule's K correlation with the GLC can be utilized further to understand why hapax legomena correlate with HNC. Legal texts have small portion of hapax legomena (24%) compared to the HNC (43%) since their vocabulary is highly repeated resulting in high values of Yule's K. From this perspective, we can conclude that hapax legomena is not a characteristic stylometric variable of the general language, but their decreased presence can be correlated with the legal genre as an indirect effect of the vocabulary repetition.

5.3 Part of Speech DFA

The DFA using Part of Speech variables as independent and text genre variables as dependent also showed that 3 variables differ significantly between the GLC and the HNC. Using the same analysis, we present the structure coefficients of each variable in the following histogram:

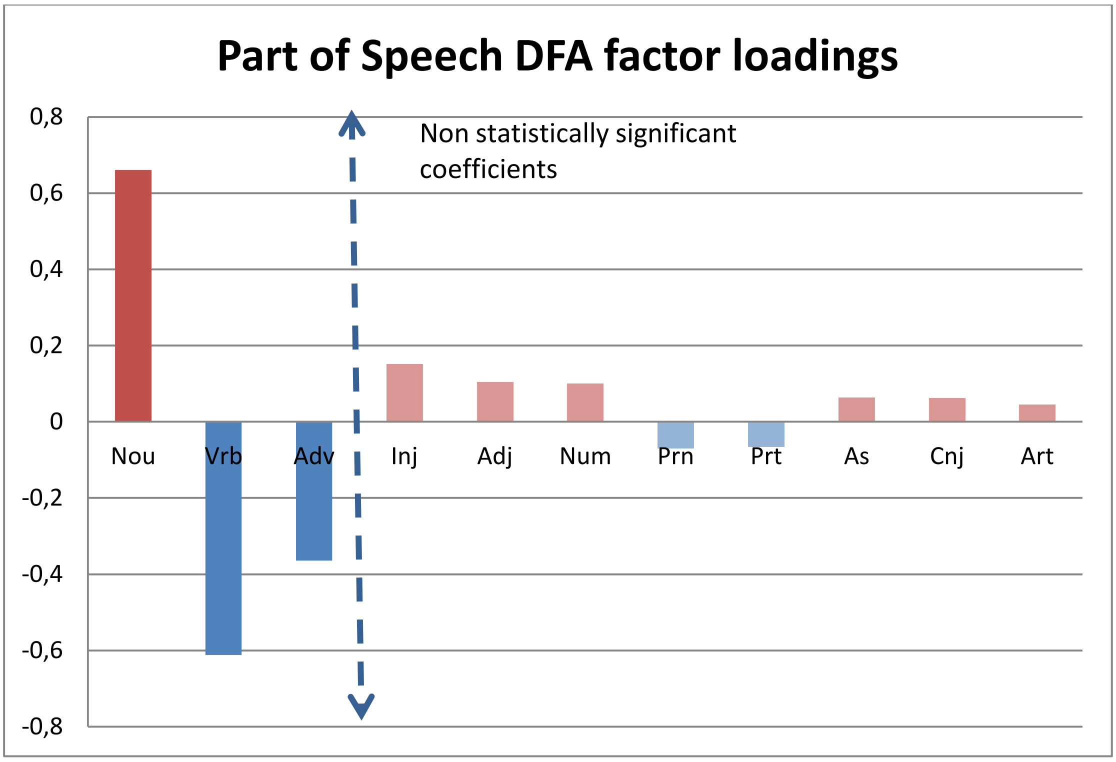

Figure 3 DFA factor loadings using Part of Speech features

The analysis of the factor loadings reveals that the legal texts have a higher frequency of nouns (26.6%) compared to the HNC (21.5%). On the other hand, HNC presents significantly higher percentages of verbs (11.15%) and adverbs (5%) compared to the GLC (8% and 3.8% respectively).

The statistical data correlate closely with some features of the legal language as so far observed. Such features are the preference to nouns instead of verbs plus the overuse of verbal nouns combined with the abundance of impersonal constructions. This deviation could be blamed for the rigidity and vagueness, which legal texts are often criticised for.

5.4 Word Length DFA

The DFA using Word Length variables as independent and text genre variables as dependent showed that 11 variables differ significantly between the GLC and the HNC. We present the structure coefficients of each variable in the following histogram:

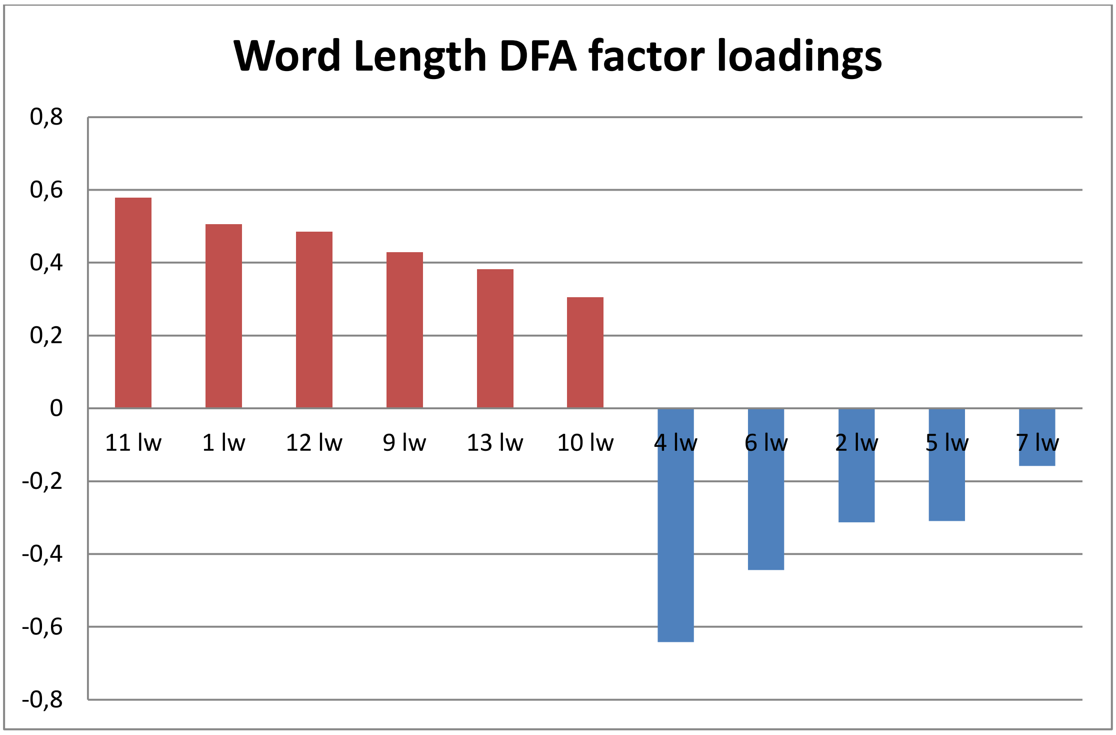

Figure 4 DFA factor loadings using Word Length features

The factor loadings of the word length spectrum reveal a strong correlation between legal genre and word length categories. In our analysis, all words greater than or equal to 9 characters (up to 13 characters) correlate with legal genre. Furthermore, 1 letter words have a higher frequency in legal texts, an observation which could be indirectly linked to the grammatical category of these words, which belongs almost always to articles. Since nouns have a significantly higher frequency in the GLC, we expect that an analogous increase could occur in articles. Another factor which could amplify the importance of one letter words concerns the tokenization rules used in the processing of the corpora. Since we used a simple regular expression tokenizer, we could not disambiguate named entities, acronyms and document structures in the texts. In many legal texts we encountered, for instance, a numbering structure which our tokenizer interpreted as one-letter tokens. The same occurred with abbreviations such as '*ν.*' which stands for '*νόμος*' (law) and is used very frequently as a quick reference to previous legislation codes.

5.5 Sentence Length DFA

The DFA with Sentence Length variables as independent and text genre variables as dependent showed that both variables of sentence length (Average Sentence Length and Standard Deviation of Sentence Length) differ significantly between the GLC and the HNC. We present the structure coefficients of each variable in the following histogram:

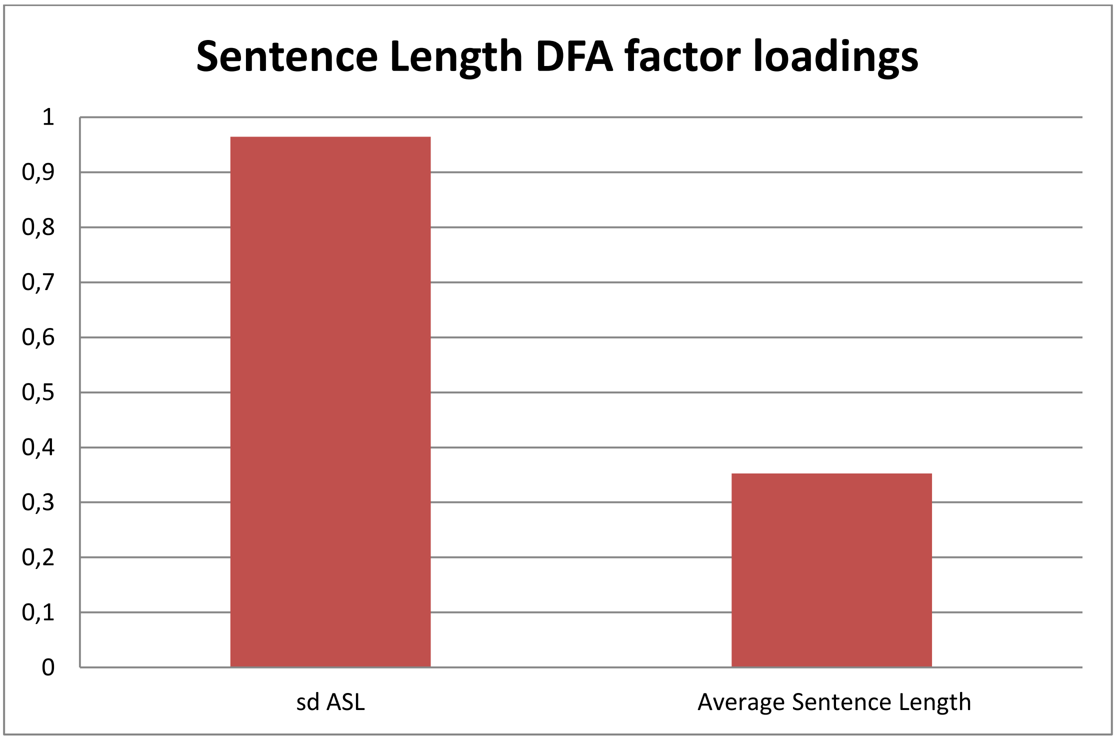

Figure 5: DFA factor loadings using Part of Speech features

Both Average Sentence Length and Standard Deviation of Sentence Length (sd ASL) correlate significantly with legal genre. The sd ASL index exhibits high factor loading, which means that texts belonging to the legal genre contain on average not only bigger sentences but also sentences whose length varies significantly between them.

6. Conclusions

Legal texts have a distinct and highly recognizable stylometric profile. All groups of stylometric measures in the GLC were found to show statistically significant differences with respect to the corpus of the general language (HNC). More specifically, the most distinctive stylometric characteristics of the legal texts are:

- Both, the ratio of content to function words (Lexical Density) and the vocabulary repeatability (Yule's K) are systematically higher in legal texts compared to texts in other genres.
- High frequency of occurrence in a specific grammatical category, i.e. nouns.
- High word length (over 8 characters).
- High sentence length with measurements that span across a wide range (high standard deviation).

The above findings confirm previous theoretical studies of the legal genre and quantify many of the qualitative observations attested there. Namely, factors such as the formulaic language, the preference to nouns and impersonal constructions, the use of technical vocabulary, and the length and complexity of sentences characterising the Greek law texts, all agree with the abovementioned statistics, as well as with the descriptions which relate such features to the lack of precision and transparency of the legal texts.

7. Acknowledgments

This study is funded by the Special Account for Research Grants. We would also like to thank Dr Effie Georgopoulou for providing us the Modern Greek parameter file in order to use TreeTagger in our research.

References

Bhatia, V. K. 2010. “Legal writing: specificity. Specification in legislative writing: accessibility, transparency, power and control.” In M. Coulthard and A. Johnson (eds.). *The Routledge Handbook of Forensic Linguistics*. Routledge: London and New York, 37-50.

Bray, J. H., & Maxwell, S. E. 1982. “Analyzing and Interpreting Significant MANOVAs.” *Review of Educational Research, 52*(3), 340-367. doi: 10.3102/00346543052003340

Gibbons, J. 2003. *Forensic Linguistics. An Introduction to Language in the Judicial System.* Oxford: Blackwell.

Hair Jr, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. 1995. *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ, USA: Prentice-Hall.

Hatzigeorgiu, N., Gavrilidou, M., Piperdis, S., Carayannis, G., Papakostopoulou, A., Athanasia, S., Iason, D. 2000. “Design and implementation of the online ILSP Greek Corpus.” In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperdis & G. Stainhaouer (Eds.), *Proceedings of the second International Conference on Language Resources and Evaluation (LREC 2000)* (Vol. III, pp. 1737-1742). Athens, Greece: ELRA.

Hoover, D. 2003. “Another perspective on vocabulary richness.” *Computers and the Humanities, 37*, 151-178.

Huberty, C. J., & Morris, J. D. 1989. “Multivariate analysis versus multiple univariate analyses.” *Psychological Bulletin, 105*(2), 302-308.

Knapp, V. 1991. “Some Problems of Legal Language.” *Ratio Juris.* Vol. 4 No. 1 March 1991, 1-17.

Meyers, L. S., Gamst, G., & Guarino, A. J. 2006. *Applied multivariate research. Design and interpretation*. Thousand Oaks, CA: Sage.

Miranda, G. A., & Calle, M. J. 2007. “Function words in authorship attribution studies.” *Literary and Linguistic Computing*, *22*(1), 49-66.

Oakes, M. P. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Olsson, J. 2004. *Forensic Linguistics: An Introduction to Language, Crime and the Law.* London: Continuum.

Schmid, H. 1994. “Probabilistic Part-of-Speech Tagging Using Decision Trees.” *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Stevens, J. P. 2002. *Applied multivariate statistics for the social sciences* (4rth ed.). Hillsdale, NJ: Erlbaum.

Tiersma, P. 1999. *Legal Language*. Chicago: University of Chicago Press.

Tweedie, F. J., & Baayen, H. R. 1998. “How variable may a constant be? Measures of lexical richness in perspective.” *Computers and the Humanities*, *32*(5), 323-352.

Weinfurth, K. P. 1995. “Multivariate Analysis of Variance.” In L. G. Grim & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 245-276). Washington, DC: American Psychological Association.

Παναρέτου, Ε. 2009. *Νομικός Λόγος. Γλώσσα και Δομή των Νόμων.* Αθήνα: Εκδόσεις Παπαζήση.