An Empirical Text Categorizing Computational Model Based on Stylistic Aspects

S. E. Michos, E. Stamatatos, N. Fakotakis, and G. Kokkinakis

Dept. of Electrical and Computer Engineering
Div. of Telecommunications and Information Technology
University of Patras

Abstract

The presented work is strongly motivated by the need of categorizing unrestricted texts in terms of functional style (FS) in order to attain a satisfying outcome in style processing. Towards this aim, it is given a three-level description of FS that comprises: (a) the basic categories of FS, (b) the main features that characterize each one of the above categories, and (c) the linguistic identifiers that act as style markers in texts for the identification of the above features. Special emphasis is put on the problems that faced the computational implementation of the aforementioned findings as well as the selection of the most appropriate stylometrics (i.e., stylistic scores) to achieve better results on text categorization. This approach is language independent, empirically-driven, and can be used in various applications including grammar and style checking, natural language generation, style verification in real-world texts, and recognition of style shift between adjacent portions of text.

1. Introduction

Though style is the main factor, besides the propositional content, that modifies the listener's reactions, there are few computational approaches that handle it. Indeed, most of the research to date in computational stylistics has been the development of the so-called *style checkers* (i.e., systems that provide a naive and superficial stylistic analysis looking for such stylistic deficiencies as unbalanced punctuation, excessive size of sentences, use of archaic words, etc.).

On the other hand, several attempts have been made for achieving a statistical analysis of style by counting certain words or phrases (i.e., the so-called *style markers*) in texts and comparing the results to a relative norm in order to decide what type of style the text is [1]. However, these systems are not able to interpret the results. This work must be done by humans.

Another interesting and thought-provoking approach in style variation makes use of a multi-dimensional methodology in order to chart the various ways in which language varies [2]. Although the above methodology is very useful due to its cross-linguistic and diachronic orientation, it presents serious technical obstacles and seems time-consuming to implement.

Moreover there are few sophisticated systems that take advantage of style features in order to attain better results in applications such as text generation (e.g., PAULINE) [3] or machine translation (e.g., STYLISTIQUE) [4]. Nevertheless, there are no known systems based on stylistic information for text categorization. STYLISTIQUE can identify the stylistic goals of the writer by choosing a goal from three dimensions (clarity-obscurity, concretness-abstraction and staticness-dynamism) but it's not clear how these dimensions vary between different text categories as well as that they are an adequate set of dimensions for text categorization.

On the other hand, many linguists claim that there are two distinct types of style: the *group* style and the *idiosyncratic* style of anyone writer. Group style can be further subdivided into two major types: *literary* style and *functional style* (FS). The term function has been used by many scholars of style in order to express different things. In the presented work we use this term as the Prague school and many Russian scholars [5] do. Hence, FS is the quantitative and qualitative use of language in a specific social relationship for a specific communication aim. It is usually encountered in texts where the personal style of the author is overshadowed by the functional objectives. Typical categories of FS are the scientific and the journalistic one. However, to the best of our knowledge, there are no computational approaches dealing with text categorization in terms of FS so far.

In this paper we present a text categorizing computational model for Modern Greek (MG) that is based

0-8186-7686-8/96 $05.00

on stylistic information, namely a three-level stylistic description of FS. Our work is strongly motivated by the need of categorizing unrestricted texts using as little information as possible. In order to achieve this purpose, we have relied on the statistical analysis of large MG text corpora as well as on empirical methods. Finally, with the view of making the required information available, we were based on already existing systems for morphological and syntactic analysis.

In the next section the three-level description of FS is briefly outlined. For a more detailed presentation the interested reader can look for [6,7]. This section ends up with the way unrestricted texts can be categorized in terms of FS as well as the selection of appropriate stylometrics to achieve the intended results on text categorization. Then, in section 3 we present the computational implementation of our model by giving the model requirements, an overview of the computational model and a clarifying example. An early evaluation of it follows in section 4. Finally, in section 5 some conclusions are drawn and future work directions are given.

2. Background

2.1. Three-level FS description

In order to model FS as well as possible we have adopted a hierarchical description that is composed of the following levels (see Figure 1):

Level 1

This level comprises the five basic categories of FS, that is *public affairs* style, *scientific* style, *journalistic* style, *everyday communication* style and *literary* style. Although the definition of a complete set of FS categories seems to be an unsolved problem, it is stressed here that this classification conforms with what many scholars call a potential and logical set of FS categories [8].

Level 2

This level includes the main features that characterize each one of the above categories, that is *formality*, *elegance*, *syntactic complexity* and *verbal complexity*.

Level 3

This level is composed of the linguistic identifiers that act as style markers in texts for the identification of the above features. These identifiers are divided into verbal and syntactic ones and are given below:

(a) ***Verbal identifiers***: idiomatic expressions like "ρίχνω λάδι στη φωτιά" (add fuel to the fire) or "πηγαίνω κατά διαβόλου" (go by the board), "sophisticated" expressions like "επ' άπειρον" (in perpetuity) or "γνήσιο τέκνο" (true-born issue), scientific terminology like "ισοζύγιο" (balance) or "πληκτρολόγιο" (keyboard), "formal" words like "άρση" (lifting) or "μεταστροφή" (swing) or "εμφαντικά" (emphatically), poetic words like "άτι" (steed) or "ξεροβόρι" (icy wind), abbreviations like "ΗΠΑ" (USA) or "ΕΚ" (EC) or "ΟΗΕ" (UN).

(b) ***Syntactic identifiers***: number of words per sentence, number of conjunctions per sentence, number of sentences per paragraph, verbs-nouns ratio, verbs at third person-verbs ratio, nouns at genitive case-nouns ratio, subordinate-main sentences ratio, adjectives-nouns ratio, adverbs-verbs ratio, active-passive voice ratio.

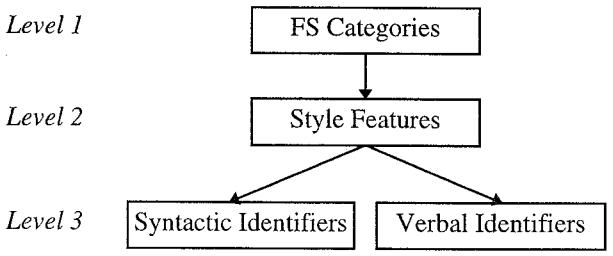

Figure 1. A three-level FS description.

Three points should be mentioned here. First, it is obvious that both a morphological and syntactic analysis of the text at hand must be available. Second, the above description would be more accurate if a semantic and/or pragmatic analysis of texts could also be available. In this case, it could be expanded to include also semantic and/or pragmatic identifiers. Nevertheless, the aim of this work is to deal with unrestricted texts, so such an effort seems unrealistic regarding the excessive computational cost that yields. Third, in order to obtain as language-independent results as possible from such a description, we attempted to build the set of style markers as generally as possible. So, intrinsic elements of MG such as the use of special verbal endings that could be comprised in the third level, have been ruled out. Surely, for getting better results it could be useful to apply the three-level description to a specific language by incorporating such special elements. It must be noted that each language has its own set of words and expressions that compose the verbal identifiers. Thus, if a word or an expression is characterised as idiomatic in one language, its translation into another language might not be idomatic at all.

2.2. FS identification

Generally, by checking the style markers in a text we are able to draw conclusions about the effect that they have on the four style features and finally make an estimation of the text FS category. The linguistic identifiers of the third level act as style markers for the style features of the second level as it is explained below:

Formality

Regarding the verbal identifiers, formal texts are characterized by the heavy use of "formal" words and "sophisticated" expressions as well as the infrequent presence of abbreviations and idiomatic expressions. Concerning the syntactic identifiers, the following style markers have been detected in formal texts: great number of words per sentence, small number of sentences per paragraph, great number of conjunctions per sentence, low verbs-nouns ratio, high nouns at genitive case-nouns ratio, high verbs at third person-verbs ratio, predominance of the passive voice over the active one and high subordinate-main sentences ratio.

Elegance

From the verbal point of view elegant texts are characterized by many idiomatic expressions and poetic words. From the syntactic point of view these texts have been observed to possess high adjectives-nouns ratio, high adverbs-verbs ratio, low verbs-nouns ratio, high verbs at third person-verbs ratio and predominance of the active voice over the passive one.

Syntactic complexity

Syntactically complex texts are characterized by great number of words per sentence, great number of sentences per paragraph, great number of conjunctions per sentence, low verbs-nouns ratio, high nouns at genitive case-nouns ratio, high verbs at third person-verbs ratio, high adjectives-nouns ratio, high adverbs-verbs ratio and high subordinate-main sentences ratio.

Verbal complexity

Verbally complex texts are characterized by many "sophisticated" expressions, plenty of scientific terminology, many "formal" words, a lot of abbreviations and poetic words and few idiomatic expressions.

Then, after having recognized the degree of effect of the four style features in a given text, the identification of its FS can be based on the following set of estimation rules:

Public affairs style: Formal and syntactically complex to a large extent, elegant and verbally complex to a small extent.

Scientific style: Formal and verbally complex to a large extent, elegant and syntactically complex to a small extent.

Journalistic style: Elegant and syntactically complex to a large extent, verbally complex and formal to a small extent.

Everyday communication style: Formal, elegant, syntactically complex and verbally complex to a small extent.

Literary style: Elegant to a large extent, formal, syntactically complex and verbally complex to a small extent.

The presented approach to text categorization was based on three main factors: (a) the empirical selection of the style markers and style features, (b) the statistical processing of large MG text corpora of about 100,000 words, and (c) the empirical assessment of the statistical results with the view of identifying FS in unrestricted texts as impartially as possible.

2.3. Determination of style markers norms

Expressions like "great number of conjunctions per sentence" or "low verbs-nouns ratio" are referred to the comparison of the text's number of conjunctions per sentence and text's verbs-nouns ratio to the corresponding ones of the language norms. It has proved that such linguistic quantities are very similar among languages. For example, for English and French the conjunctions are approximately 4% and 3% of the words respectively, while the verbs-nouns ratio is approximately 0,6 and 0,5 respectively [9].

In Table 1 we give the set of style markers norms for MG as it was derived from statistical analysis of large tagged MG text corpora taken from the ESPRIT-860 project [10] (it should be mentioned here that the texts were selected to belong to all FS categories). This set can be easily ported to other languages with slight modifications of its values. It has also to be noted that some values especially those referring to verbal identifiers are approximate since it is not yet possible to have an acceptable average for them.

Table 1. Style markers norms for MG.

Style Markers	Norm
number of words per sentence	15
number of conjunctions per sentence	0,6
number of sentences per paragraph	5
verbs-nouns ratio	0,5
verbs at third person-verbs ratio	0,6
nouns at genitive case-nouns ratio	0,25
subordinate-main sentences ratio	1,5
adjectives-nouns ratio	0,3
adverbs-verbs ratio	0,4
active-passive voice ratio	1,5
idiomatic expressions	0,02
"sophisticated" expressions	0,01
scientific terminology	0,01
"formal" words	0,05
poetic words	0,01
abbreviations	0,02

2.4. Text categorization methodology

According to the previous stylistic description, if the detected value of a style marker is different from that of its norm, then this style marker may have a positive or

negative effect on a certain style feature. For example, if the active-passive voice ratio has been found to be greater than the norm, then this style marker has a positive effect on the elegance and a negative one on the formality as it can be derived from the descriptions of these two features in section 2.2.

Additionally, a style feature is considered to be "to a small extent" if the percentage of the style markers that have a positive affect on it is smaller than 50% (<50%). Furthermore, a style feature is considered to be "to a large extent" if the corresponding percentage of the style markers that have a positive affect on it is greater than 65% (>65%). If the previous percentage is between 50% and 65% (50%-65%), then this percentage is ambiguous and cannot lead to a valid estimation of the feature impact.

Finally, the estimation on the FS category of a given text is made by employing the set of the estimation rules of the section 2.2. Needless to say that every time we have four measured percentages that equal the number of four style features of a FS category, this category compose the estimation. Therefore, if at least three of the above percentages are unambiguous (i.e., <50% or >65%), we look for the estimation rule that best matches the results. If there are two of them, we do make an estimation but this estimation cannot lead to a definite FS category. In this case, a further analysis of the given text is needed in order to draw a more precise conclusion of its FS category. On the other hand, if at least two of the percentages are ambiguous, an estimation is no longer feasible. Again in this case a further analysis of the given text is needed in order for an estimation to be feasible. Obviously, in several cases the extraction of a valid estimation is a quite difficult process, especially when the size of the text is too small.

3. Implementation

3.1. Requirements

As it has been mentioned in the previous section, in order to develop a system that will implement the three-level description of FS, a morphological and a syntactic analyzer should be available.

- The former must be able to provide verbal information (e.g., "formal" word, abbreviation, poetic word, etc.) besides the pure morphological information (e.g., part-of-speech, case, number, etc.) for each word of the text.
- The latter must be able to recognize predefined expressions (e.g., "sophisticated" ones, idiomatic ones, etc.), calculate syntactic quantities (e.g., number of words per sentence, number of sentences per paragraph, etc.), and provide syntactic information (e.g., main sentences, subordinate sentences, etc.) for every sentence of the text.

3.2. Overview of the model

The presented model is the integration of three distinct modules as it is shown in Figure 2. These modules are described below:

(a) The Morphological Analyzer (MA) is a two-level processor based on a PC-KIMMO description of MG. Its lexicon contains about 30000 words. For a detailed presentation the interested reader can look for [11].

(b) The Syntactic Parser (SP) is a computational model that is able to parse unrestricted texts of 'quasi free' word order languages such as MG. For a detailed presentation the interested reader can look for [12,13].

(c) The Stylistic Analyzer (SA) is the module that implements the presented method for text categorization based on stylistic information.

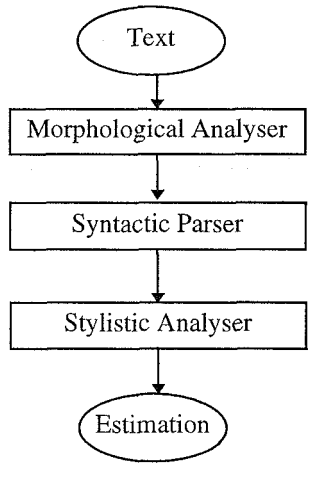

Figure 2. Overview of the model.

So, when the morphological and syntactic processing of the text have been carried out, all the required style markers values are available and the SA is able to make an estimation about the FS of the text based on the three-level stylistic description.

It has to be underlined that the two modules, MA and SP, already existed and were not designed especially for stylistic analysis. Hence, in order to adopt them on our model several modifications had to be done. So, the MA has been extended to include the required verbal information. On the other hand, the SP has improved with the incorporation of a submodule that recognizes characteristic expressions and the addition of functions that calculate the required style markers values.

Table 2. Results of the analysis of the sample text.

Style Markers	Value	Deviation (%)	Formality	Elegance	Syntactic Complexity	Verbal Complexity
number of words per sentence	27,7	+85	+		+	
number of conjunctions per sentence	1,17	+95	+		+	
number of sentences per paragraph	2,74	-45	+		–	
verbs-nouns ratio	0,59	+18	–	–	–	
verbs at third person-verbs ratio	0,79	+27	+	+	+	
nouns at genitive case-nouns ratio	0,27	+8	+		+	
subordinate-main sentences ratio	2,3	+53	+		+	
adjectives-nouns ratio	0,62	+101		+	+	
adverbs-verbs ratio	0,57	+43		+	+	
active-passive voice ratio	1,53	+2	–	+		
idiomatic expressions	0,05	+150	–	+		–
"sophisticated" expressions	0,008	-20	–			–
scientific terminology	0,01	0				–
"formal" words	0,04	-20	–			–
poetic words	0	-100	+	–		–
abbreviations	0,001	-95	+			–

3.3. A clarifying example

With the view of clarifying further the above methodology to text categorization in terms of FS we give in this section a detailed example of identification of the FS category of a text based on it. We have used a text of 3500 words taken from a newspaper that has been analyzed in the framework of the ESPRIT-860 project. It has to be noted that this analysis provided only a part of the aforementioned set of style markers. All the rest have been calculated manually.

After the morphological and syntactic analysis of the sample text, the set of the values of the style markers was available. The results, the corresponding deviations from the norm values as well as their effect on each style feature are shown in Table 2. Note that the symbols (+) and (-) stand for positive and negative effect on a certain feature respectively.

Taking into account the results of this table we calculated the percentages of the style markers that have a positive effect on each style feature. These can be summarized as follows:

Formality: $8/13 \approx 62\%$ (>50%, >65%)

Elegance: $5/7 \approx 71\%$ (>65%)

Syntactic Complexity: $7/9 \approx 78\%$ (>65%)

Verbal Complexity: $0/6 \approx 0\%$ (<50%)

From the observation of these percentages, we can conclude that the sample text is elegant and syntactically complex to a large extent and verbally complex to a small extent. Regarding the formality of this text we cannot make a valid estimation of this feature impact since its percentage has been found to be ambiguous. Finally, the estimation rule that best matches these results is that of the *journalistic style* since at least three of the above percentages are unambiguous (i.e., elegance, syntactic complexity, and verbal complexity).

4. Evaluation

In Table 3 there are shown the analysis results and the estimations the model produced for five sample texts, each one of them belonging to a different FS category. In spite of the small size of these sample texts (about 210-841 words), the model managed to identify correctly the FS of 3 texts (i.e., public affairs, scientific and everyday communication).

Table 3. Analysis results for 5 sample texts.

FS Category	Words	Formality (%)	Elegance (%)	Syntactic Complexity (%)	Verbal Complexity (%)	Estimation
Public Affairs	841	77	57	78	17	*Public Affairs*
Scientific	500	77	29	56	67	*Scientific*
Journalistic	320	77	43	67	17	*Public Affairs*
Everyday Commu/tion	210	23	43	11	0	*Everyday Communication*
Literary	395	31	71	56	0	*Literary or Journalistic*

Moreover, the estimation for the literary text led to two FS categories (i.e., literary or journalistic) and only for one sample text (i.e., the journalistic one), the estimation is not correct. This was due to the unusual high percentage of the formality of this text.

However, in order to attain as better results as possible, it has been estimated that the number of words in the text must be at least 500.

5. Conclusion

Stylistic aspects, though necessary in deep understanding of language, have been neglected in computational linguistics research. These problems had been too vague and ill-defined to be dealt with by computational systems. However, in this work, we have presented an empirical model based on a formal description of FS that makes the problem of text categorization more amenable to computational solution. It is hoped that this research will lead to a system sophisticated enough to cope with various applications including grammar and style checking, natural language generation, style verification in real-world texts, and recognition of style shift between adjacent portions of text (e.g., paragraphs).

It can be understood that the more the deviation of a linguistic identifier is from the norm, the more significant its effect is on the estimation process. For instance, a text that has a verbs-nouns ratio equal to 0,2 (i.e., deviation from the norm = 60%) is considered more formal than another one that has 0,3 (i.e., deviation from the norm = 40%). For all these, its obvious that the deviation of the style marker value from the norm must be taken into account by the model by using a weights mechanism. However, in those cases that the percentage of the deviation of a linguistic identifier from its norm is sufficiently small, if not negligible, we are looking for some threshold values that will ensure the correct evaluation of our results.

Short-term research is currently focused on some problems that faces the computational implementation of the aforementioned findings as well as the selection of more precise and appropriate stylometrics to achieve better results on text categorization. Towards this direction, the extraction of the most appropriate language norms for all the presented style markers on one hand and the formulation of the most accurate estimation rules on the other hand are the key points for the successful completion of the above research.

References

[1] CLUETT R. (1990), "*Canadian Literary Prose: A Preliminary Stylistic Atlas*", ECW Press.

[2] BIBER D. (1995), "Dimensions of Register Variation: A cross-linguistic comparison", Cambridge University Press.

[3] HOVY E.H. (1990), "*Pragmatics and Natural Language Generation*", Artificial Intelligence, vol. 43, pp. 153-197.

[4] DIMARCO C. & HIRST G. (1993), "*A Computational Theory of Goal-Directed Style in Syntax*", Computational Linguistics, vol. 19, no. 3, pp. 452-459.

[5] RIESEL E. (1971), "*Stil und Gesellschaft*", Lange-Roloff, pp. 357-365.

[6] MICHOS S. E., STAMATATOS E., FAKOTAKIS N. & KOKKINAKIS G. (1996), "*Identification of Functional Style in Unrestricted Texts Based on a Three-Level Stylistic Description*", Proceedings of the AISB 1996 Workshop on Language Engineering for Document Analysis and Recognition, Brighton.

[7] MICHOS S. E., STAMATATOS E, FAKOTAKIS N. & KOKKINAKIS G. (1996), "*Functional Style Recognition in Real-World Texts*", Proceedings of the 10th International Symposium on Theoretical and Applied Linguistics, Thessaloniki.

[8] RIESEL E. (1963), "*Stilistik der deutschen Srache*", 2nd Edition, Moskau.

[9] DERMATAS E. & KOKKINAKIS G. (1995), "*Automatic Stochastic Tagging of Natural Language Texts*", Computational Linguistics, vol. 21, no. 2, pp. 137-163.

[10] TECHNICAL ANNEX OF THE ESPRIT-860 PROJECT (1986), "*Linguistic Analysis of the European Languages*".

[11] SGARBAS K., N. FAKOTAKIS & G. KOKKINAKIS (1995), "*A PC-KIMMO Based Morphological Description of Modern Greek*", Literary and Linguistic Computing, Vol. 10, No. 3, Oxford University Press, New York.

[12] MICHOS S. E., N. FAKOTAKIS & G. KOKKINAKIS (1994), "*A Novel method for Parsing Complex Sentences in Syntactically-Free Languages*", Proceedings of the 6th IEEE International Conference on Tools with Artificial Intelligence, New Orleans.

[13] MICHOS S. E., N. FAKOTAKIS & G. KOKKINAKIS (1995), "*A Novel and Efficient Method for Parsing Unrestricted Texts of Quasi Free Word Order Languages*", International Journal on Artificial Intelligence Tools, Vol.4, No. 3, pp. 301-321.