Categorising Texts by Using a Three-Level Functional Style Description

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

Dept. of Electrical Engineering and Computer Science
Div. of Telecommunications and Information Technology
University of Patras
26500, Patras, GREECE
Email: michos@wcl.ee.upatras.gr

Abstract

The presented work is strongly motivated by the need of modelling functional style (FS) as well as categorising unrestricted texts in terms of FS in order to attain a satisfying outcome in style processing. Towards this end, it is given a three-level description of FS that comprises: (a) the basic categories of FS, (b) the main features that characterise 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 faces a 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 categorisation. This approach is language independent, statistically and empirically-driven, and can be used in various applications including text categorisation, natural language generation, style verification in real-world texts, and recognition of style shift between adjacent portions of text.

1 Background

Style is the main factor, besides the propositional content, that modifies the listener's reactions. The more important the style is for text understanding, the less computational approaches that handle it there are. Indeed, most of the research to date in computational stylistics has been the development of so-called *style checkers*. Although there are quite a few of them (e.g. RightWriter, CRITIQUE, etc.), they use neither a vocabulary of style, nor a structured representation of stylistic rules. 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, the interpretation of the results is still done by humans.

Most of these systems, if not all, they do not essentially understand what they do. STYLISTIQUE [2] and PAULINE [3] are two additional systems that try to obtain a deep understanding of style in order to achieve better results in machine translation and text generation respectively. Hence, style has been used so far in computational linguistics for

improving the output of either machine translators or text generators. There are no known approaches taking advantage of its features in applications such as text categorisation.

Many linguists claim that there are two distinct types of style: the *literary* style and the *functional* style. The term function has been used by many scholars of style in order to express different things. Firstly, [4] has proposed a list of factors along with a list of functions that correspond to these factors and dominate a concrete text. Another meaning of FS is the functional efficacy: a style is functional if it works efficiently in a given situation [5]. In the presented work we use this term as the Prague school and many Russian scholars [6] 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 categorisation in terms of FS so far.

Our work is strongly motivated by the need of modelling FS as well as categorising unrestricted texts in terms of FS. In order to achieve this purpose, we have relied on both statistical analysis of large Greek text corpora and empirical methods. Our final aim is the development of a computational system that will be able to identify texts of different FS as well as recognise FS shifts between adjacent portions of a concrete text.

In this paper we present an approach to text categorisation that is based on a three-level description of FS. In the next section the three-level description of FS is briefly outlined. For a more detailed presentation the interested reader can look for [7]. This section ends up with the way unrestricted texts can be identified in terms of FS as well as the selection of appropriate stylometrics to achieve the intended results on text categorisation. Then, in section 3 we give an example of application of the above approach to a sample text and make an estimation of its possible FS. Finally, in section 4 some conclusions are drawn and future directions to a complete computational implementation of these findings are given.

2 Our Approach

2.1 Three-level FS description

In order to model FS as better as possible we have adopted a hierarchical description that is composed of the following levels (see Fig. 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 characterise 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:

- *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).
- *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.

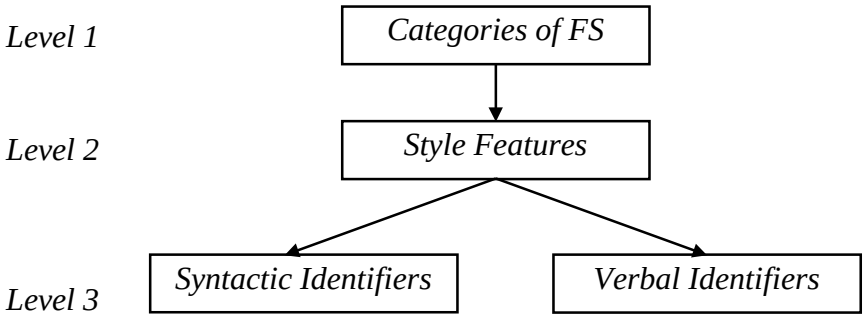

Figure 1. A three-level FS description.

Three points should be mentioned here. Firstly, it is obvious that both a morphological and syntactic analysis of the text at hand must be available. Secondly, 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. Thirdly, 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 the Greek language 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.

2.2 FS identification

Generally, by checking the style markers in a text we are able to draw conclusions about the effect that 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 characterised by the large 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 characterised 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 characterised 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 characterised 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 recognised 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 categorisation was based on three main factors: (a) the empirical selection of the style markers, (b) the statistical processing of medium-sized Greek 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. The previous Greek text corpora were tagged (morphologically and syntactically) and were taken from the ESPRIT-860 project [10]. These texts have come from Greek newspapers, official documents of the European Community in Greek, and some selective literature texts from Greek writers who do not use any kind of idiomatic language.

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 the Greek language as it was derived from the statistical analysis of the aforementioned Greek text corpora. 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.

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

Table 1. Style markers norms for the Greek Language.

2.4 Text categorisation 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 lesser 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. 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. An Example

With the view of clarifying further the above methodology to text categorisation 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 analysed 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, the syntactic ones. The verbal ones have been calculated manually.

From the morphological and syntactic analysis of the sample text we calculated the set of the values of the style markers. 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.

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	+			–

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

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 summarised as follows:

Formality:	8/13 ≈ 62%	(between 50% and 65%)
Elegance:	5/7 ≈ 71%	(>65%)
Syntactic Complexity:	7/9 ≈ 78%	(>65%)
Verbal Complexity:	0/6 ≈ 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. Conclusions and Future Work

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 a novel, formal description of FS that makes the problem of FS identification in unrestricted texts more amenable to computational solution. It is hoped that this research will lead to a system sophisticated enough to cope with various applications including text categorisation, 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%). 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 the problems that faces a 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 categorisation. 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 precise 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] DIMARCO C. & HIRST G. (1993), "*A Computational Theory of Goal-Directed Style in Syntax*", Computational Linguistics, vol. 19, no. 3, pp. 452-459.

[3] HOVY E.H. (1990), "*Pragmatics and Natural Language Generation*", Artificial Intelligence, vol. 43, pp. 153-197.
[4] JACOBSON R. (1960), "*Linguistics and Poetics*", Sebeok, pp. 350-377.
[5] ENKVIST N.E. (1973), "*Linguistic Stylistics*", The Hague: Mouton.
[6] RIESEL E. (1971), "*Stil und Gesellschaft*", Lange-Roloff, pp. 357-365.
[7] 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.
[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*".