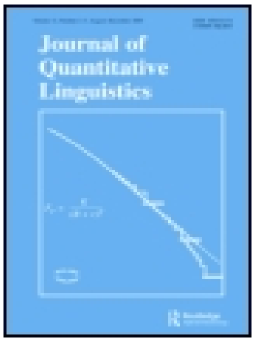


Journal of Quantitative Linguistics

Routledge
Taylor & Francis Group

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: https://www.tandfonline.com/loi/njql20

The Discriminativeness of Internal Syntactic Representations in Automatic Genre Classification

Mingyu Wan, Alex Chengyu Fang & Chu-Ren Huang

JOURNAL OF QUANTITATIVE LINGUISTICS
https://doi.org/10.1080/09296174.2019.1663655

The Discriminativeness of Internal Syntactic Representations in Automatic Genre Classification

Mingyu Wan[a], Alex Chengyu Fang[b] and Chu-Ren Huang[a*]

[a]School of Foreign Languages, Peking University, Beijing, China; [b]Department of Linguistics and Translation, City University of Hong Kong, Hong Kong, China

ABSTRACT

Genre characterizes a document differently from a subject that has been the focus of most document retrieval and classification applications. This work hypothesizes a close interaction between syntactic variation and genre differentiation by introspecting stylistic cues in functional and structural aspects beyond word level. It has engineered 14 syntactic feature sets of internal representations for genre classification through Machine Learning devices. Experiment results show significant superiority of fusing structural and lexical features for genre classification ($F_{\Delta max.}$ = 9.2%, sig. = 0.001), suggesting the effectiveness of incorporating syntactic cues for genre discrimination. In addition, the PCA analysis reports the noun phrases (NP) as the most principle component (66%) for genre variation and prepositional phrases (PP) the second. Particularly, noun phrases with dominant structures of prepositional complements and pronouns functioning as a subject are most effective for identifying printed texts of high formality, while prepositional phrases are useful for identifying speeches of low formality. Error analysis suggests that the phrasal features are particularly useful for classifying four groups of genre classes, i.e. unscripted speech, fiction, news reports, and academic writing, all distributed with distinct structural characteristics, and they demonstrate an incremental degree of formality in the continuum of language complexity.

1. Introduction

1.1. Automatic Genre Classification and Its Challenges

Automatic Genre Classification (AGC) is one of text classification applications in data mining which aims at automatically categorizing the genre class of a document for facilitating genre-related document retrieval (Fabrizio, 2002). As genre characterizes a document differently from a subject or a topic that has been the focus of most document retrieval and classification applications, AGC becomes a more challenging issue than conventional text classification (Lim, Lee, & Kim, 2005). For instance, the

CONTACT Mingyu Wan pku.clara@gmail.com

*Present affiliation: Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China.

documents grouped by a subject 'life story' will vary widely in their styles such as conversation, fiction, news articles, academic writing and so on. A news article about a person's life story might show some formal and factual reportage, while a fiction describes it in narratives.

Swales (1990) defines genre as a class of communicative events where there is some shared set of linguistic forms serving certain communicative purposes. As such, a genre is considered as a stylistic view of a document, which demonstrates some distinct characteristics that a subject or a topic cannot simply represent, including the categorical, formulative and functional aspects of the expressions and they are far more complex than the document content (Biber, 1992, 1995; Miller & Charles, 1991).

A challenge of AGC comes from the vague, arguable and unsolved definitions of genre in the existing literature (Biber, 1992; Halliday, Matthiessen, & Matthiessen, 2014; Martin, 1984) or the state-of-the-art applications (Lee & Myaeng, 2002; Scaringella, Zoia, & Mlynek, 2006; Sigtia & Dixon, 2014). Biber (1995) introduces the idea of predicting the function of a text by the structural patterns which were instantiated from the quantitative approach according to which distributional patterns vary with the text function. Miller and Charles (1991) then stated the weak contextual hypothesis that structural differences reflect functional ones while similar functions tend to be manifested by similarly structured texts. There is a many-to-many relation of structure and function, and learning text types by exploring text structures is a nontrivial task, as neither can we deterministically infer a unique function based on observing some text pattern, nor is the same function always manifested by the same pattern (Mehler, Geibel, & Pustylnikov, 2007).

However, due to practical reasons, existing studies of AGC have extensively adopted topic-related surface features for genre classification. For example, Bekkerman and Allan (2004), Fürnkranz (1998), Mikolov, Chen, Corrado, and Dean (2013), Tan, Wang, and Lee (2002) used BOWs, n-grams, skip n-grams, or word-to-vectors (*e.g.* one-hot representations) as feature vectors, which can be constructed automatically from large-scale unstructured datasets. Needless to say, modelling surface features is easy, fast and effective, yet with decent performance. The challenge is, the surface feature technology usually lacks intrinsic explanations to genre-specific properties and is encountering a dilemma of further breakthroughs with the full-fledged ML techniques (Liu & Wan, 2019; Manning, 2011; Wan & Liu, 2018). The AGC technology will need to investigate sophisticated linguistic cues to tease out the characteristics of a genre. This calls for a well-defined corpus in terms of genres and an examination of genre-specific representations by seeking cross-discipline breakthroughs between computer scientists, domain experts and knowledge engineers (Hou & Huang, in press, Hou, Huang, Ahrens, & Lee, 2019; Hou, Huang, Do, & Liu, 2017a; Hou, Huang, & Liu, 2017b; Wan et al., 2019).

In the present paper, we focus on the feature engineering of internal syntactic features for AGC with ML and NLP techniques by using a well-structured English corpus – The ICE-GB corpus (International Corpus of English, the British component), which provides this study with rich linguistic information and well-defined texts categories of unambiguous genre classes (cf. Section 3.1).

1.2. Objectives

In this paper, our goal is to employ fine-grained syntactic features for genre classification by modelling a comprehensive collection of syntactic features of internal structures by using the ICE-GB corpus. We hypothesize that the internal syntactic structures, which encode rich linguistic cues of structural, categorical and functional information of constituents (cf. Section 4.2.1) might be more indicative of a text genre (style) than surface features, such as content-bearing words, function words, punctuations marks or just BOWs. We aim to achieve the following five objectives in this work:

(1) to test fine-grained syntactic features for automatic genre classification;
(2) to examine the discriminativeness of various internal syntactic features for genre classification;
(3) to understand the interaction between the various syntactic features and text genres;
(4) to provide an explanation to genre classification applications in terms of syntactic enquires;
(5) to lay a foundation for future syntactic stylometry and complexity studies.

The remain of paper is organized as follows. Section 2 reviews the most relevant work in AGC during the past few decades and highlights the uniqueness and significance of this work. Section 3 introduces the data/corpus and its parsing scheme for a general understanding of the syntactic information. Section 4 describes the methodology including the ML methods and feature engineering of the internal representations. Section 5 shows the experimental setup, classification results, error analysis, discussion, etc. Section 6 concludes the current work and prospects the future work.

2. Related Work

AGC involves continuous and multi-disciplinary endeavours of many great works from both qualitative and quantitative perspectives (Biber, 1988, 1992, 1995; Hou et al., 2019; Karlgren, 2000, 2004; Karlgren & Cutting, 1994; Kessler, Nunberg, & Sch¨utze, 1997; Lee & Myaeng, 2002; Nanni, Costa, Lumini, Kim, & Baek, 2016; Stamatatos, Fakotakis, & Kokkinakis, 2000). However, many recent genre classification applications are related to music genres (Nanni et al., 2016; Scaringella et al., 2006; Sigtia & Dixon, 2014) or web documents (Lim et al., 2005) where main arguments of genre identification are deliberately avoided. Existing AGC technologies treat genre classification much alike a conventional text classification task and are lack of feature vectors of genre-related characteristics. For example, commonly adopted features are simply derived from textual properties, such as the number of sentences, the number of a certain word, etc. Although the documents can be grouped successfully according to their subjects, there is a big difference in styles among the documents in a cluster. It is crucial that a document can be represented by feature space that is close to the attribute of a genre, that is, selecting features that can make a clear distinction among the genres is the core of automatic genre classification.

In recent decades, feature engineering attracted many researchers in the field of document classification. Broadly speaking, feature selection of AGC can be categorized into two main streams of technologies: one led by computer scientists who aim at advancing Machine Learning (ML) and Feature Selection (FS) algorithms based on surface features; the other led by knowledge engineers, such as linguists, who aim at constructing cleaner and finer-structured data resources for empirical observation of genre-related characteristics by generalizing language rules with expert knowledge. Notably, computational linguists possess both computational skills and linguistic expertise in structuring data for automatically modelling discriminative linguistic features with various kinds of linguistic annotations (*e.g.* acoustic-prosodic tags, part-of-speech tags, syntactic dependencies, semantic relations, speech acts, etc.) (Chen et al., 1999; Fang, 1996; Liu & Huang, 2016; Liu & Wan, 2019; Neergaard & Huang, 2019; Wang, Huang, Yao, & Chan, 2019). These investments provide useful knowledge-enriched resources/features for world-wide and cross-lingual AGC (or NLP in a wider scope) applications.

The aim of feature engineering is to find alternatives to the bag-of-words approach (Biber, 1995; Kessler et al., 1997; Karlgren, 2000; Lee & Myaeng, 2002; Wolters & Kirsten, 1999). Although lexical features are selective with respect to text content, this IR model generally disregards text structure. Now, modelling document structure comes into reach of machine learning. Some approaches even show that structural patterns allow to classify texts in the absence of any lexical information (Lindemann & Littig, 2006; Pustylnikov, 2006). For instance, Fang and Cao (2015) achieved promising results in genre classification by focusing on frequencies of fine-grained POS tags. In addition, various types of features have been proposed for the automatic classification of text genres. Table 1 summaries some most related works in terms of feature engineering.

Early studies of AGC have already experimented some simple linguistic features. A representative work would be Karlgren and Cutting (1994), where they adopted a small set of 20 simple features using discriminative analysis. These features include a few lexical counts (*e.g.* 'Me' count, 'I' count, 'It' count), POS counts (*e.g.* Adverb, Noun, Preposition), and some derivative features (*e.g.* Character count, Long word count, Type-Token Ratio). The result showed a slight improvement of using POS features compared to the lexical and derivative features. They also found that the classification performance decreases substantially with the increasing of genre classes: the respective error rates are 4%, 27% and 48% in the cases of 2, 4 and 15 genres.

Wolters and Kirsten (1999) adopted content words, function words, lemmata, and POS information for domain and genre classification with the use of a KNN classifier. They conducted feature selection by excluding all lemmata that occur in less than 10 sources for the reduction of feature space. Besides, gain ratio was used for further feature selection. Their experimental results showed 1) a negative correlation between the feature space and the general performance. 2) a positive correlation between POS features and recall. 3) an increase of precision when POS information is encoded.

Stamatatos et al. (2000) employed a syntactic parser for extracting grammatical features. Unlike the previous studies, they use the features extracted from a phrasal level and an analytical level, such as the ratio of NPs (noun phrase) to the total number of chunks, the average number of words included in NP and morphological ambiguities or syntactic ambiguities, and so on. The results showed that using grammatical features is better than the one using the high-frequency words. As for textual styles, they constructed a corpus of 10 genre classes by downloading the documents from websites.

Lee and Myaeng (2002) presented a methodology for genre classification by using word statistics. They attempt to find the most genre-revealing terms by computing the goodness value of the terms with *df* and *tf*. Deviation formula and discrimination formula were also used. The similarity-based framework was adopted to compare with the Naive Bayes classifier approach. Web documents were collected for their experiment. Seven genre classes were gathered: reportage, editorial, technical paper, critical review, personal homepage, Q&A, and product specification. Half of the collected documents were used for training and the remaining half were used for testing. They found that the *df* ratio always resulted in a better performance than *tf* or *tf* ratios, although the ordinary *tf* values produced the best result when Naive

Table 1. The main features adopted in previous studies.

Studies	Feature Class				Feature Set	Top Feature
	Lexical cues	POS cues	Derivative cues	Structural cues		
Karlgren and Cutting (1994)	✓	✓	✓		lexical, POS and derivative counts	POS
Wolters and Kirsten (1999)	✓	✓			content words, function words, lemmata, and POS information	POS
Stamatatos et al. (2000)	✓			✓	high-frequency words, ratio of NPs, no. of words included in NP, morphological or syntactic ambiguities	Structural
Lee and Myaeng (2002)			✓		*tf, df* and ratios	*df*
Fang and Cao (2015)	✓	✓			BOW, impoverished and fine-grained POS tags	fine-grained POS tags
Hou and Huang (in press), Hou et al. (2017a, 2017b, 2019)		✓	✓	✓	average word length distribution, sentence length distribution, POS distribution	Structural

Bayes was employed. Some genre classes were found to be more difficult for classification because they tended to share some subject-specific terms.

Fang and Cao (2015) used a fine-grained POS tag set (487 tags) in comparison with BOW and impoverished POS tag set (36 tags), by using the NB classifier. They found that fine-grained POS tags would significantly help improve the classification performance, compared with the other two baselines, and the advantage is even larger when the genre classes became more granulated. Their experimental results have well-supported their claim that well-annotated linguistic interpretations are more useful cues for the detection of different genre classes in comparison to impoverished features.

A series of papers by Hou and Huang (in press) and Hou et al. (2017a, 2017b, 2019) dealt with syntactic features for text classification relating to Chinese register/stylometrics. Specifically, they investigated linguistic characteristics of Chinese register based on the Menzerath – Altmann Law and Text Clustering. Their studies showed that syntactic features, such as the power relations between a linguistic unit and its constituents, and distributional relations among different POS tags, can be effective features for textual classification and are not necessarily more complex or difficult to obtain than BOW.

The above studies tend to suggest the usefulness of non-lexical features for genre classification that involve the structural and grammatical information. However, related studies of modelling fine-grained syntactic features for AGC are under-researched. A limited number of studies are found to use syntactic categories of coarse granularity and some attempts to make use of syntactic relations. In this work, we propose to engineer syntactic features of finer granularity for AGC, with a focus on the internal structures of constituents. We also analyse the interaction of internal syntactic structures with respective genre categories through ML devices, so as to tease out the syntactic characteristics of text genres.

3. Data and Its Syntactic Parsing Scheme

3.1. The ICE-GB Corpus

The ICE-GB corpus (International Corpus of English, the British component) is adopted as the training and testing dataset. The ICE project was launched by Professor Sidney Greenbaum at the Survey of English Usage, University College London. This project, participated by 20 national and regional teams, aims at the grammatical description of English in countries and regions where it is used either as a first or an official language. The

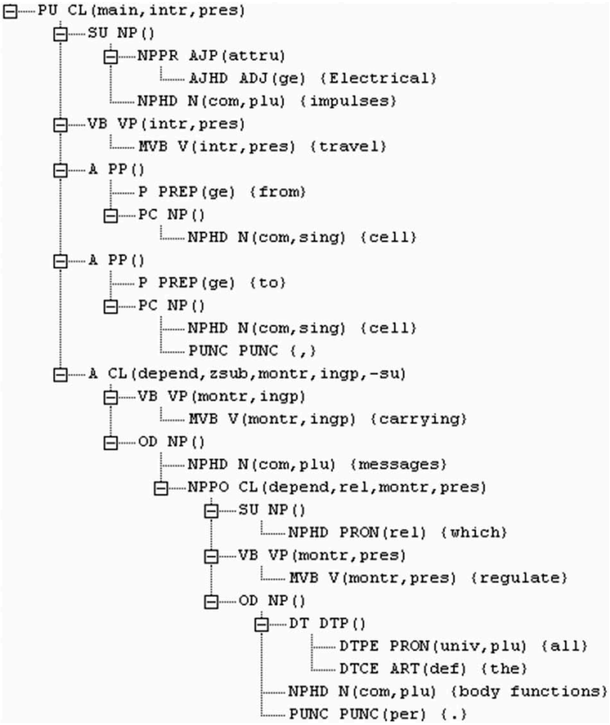
British component of the corpus (ICE-GB) consists of 300 texts of transcribed speech and 200 texts of written samples, of 2,000 tokens each, generally dated from the period 1990–1994. Text collections were sampled randomly to represent daily usage of English with balanced and well-defined genre categories in a hierarchical taxonomy (cf. Appendix A). The spoken section, which contains 60% of the total corpus in terms of words, is divided between dialogues and monologues. The dialogues range from private direct and distanced conversations to public situations such as broadcast discussions and parliamentary debates. The written samples are divided into two initial categories: non-printed and printed. The former is a collection of university essays and letters of correspondence. The latter has four major divisions: informational, instructional, persuasive, and creative. This categorization of genres indicates how native English users deal with the language in daily life in a broader sense, and therefore provides a benchmark for defining genres which can be projected from different contextual and cultural factors. As the ICE-GB corpus contains the most representative genre structures and fine-grained linguistic information for the English language, we are able to construct a comprehensive set of fine-grained internal structural features for AGC, as detailed in the following sections.

3.2. The Parsing Scheme

ICE-GB has been grammatically tagged, syntactically parsed and manually checked. The parsing scheme indicates a full analysis of the phrase structures and assigns syntactic functions to these constituents by following the comprehensive grammar of English (Quirk, Greenbaum, Leech, & Svartvik, 1985). The ICE-GB parsing schema includes 32 syntactic categories (including the main word classes), 58 syntactic functions, and 117 grammatical features. In this paper, we focus on the syntactic categories and functions. Details of the tags are attached in Appendices B and C.

Tagging was implemented automatically by the Survey Parser (Fang, 1996) at around 87% accuracy, and manual validation was conducted for ensuring a 97% accuracy of the syntactic tags in the published dataset. Each node in the tree is labelled with up to three types of information: word class (POS tag), syntactic category, syntactic function, as well as grammatical features (*e.g.* transitivity, tense, aspect). An example syntactic tree in ICE-GB is shown below:

Example 1. '*Electrical pulses travel from cell to cell, carrying messages which regulate all the body functions.*' <W2B-023-004 >[1]



```
⊟PU CL(main,intr,pres)
    ⊟SU NP()
        ⊟NPPR AJP(attru)
            AJHD ADJ(ge) {Electrical}
        NPHD N(com,plu) {impulses}
    ⊟VB VP(intr,pres)
        MVB V(intr,pres) {travel}
    ⊟A PP()
        P PREP(ge) {from}
        ⊟PC NP()
            NPHD N(com,sing) {cell}
    ⊟A PP()
        P PREP(ge) {to}
        ⊟PC NP()
            NPHD N(com,sing) {cell}
            PUNC PUNC {,}
    ⊟A CL(depend,zsub,montr,ingp,-su)
        ⊟VB VP(montr,ingp)
            MVB V(montr,ingp) {carrying}
        ⊟OD NP()
            NPHD N(com,plu) {messages}
            ⊟NPPO CL(depend,rel,montr,pres)
                ⊟SU NP()
                    NPHD PRON(rel) {which}
                ⊟VB VP(montr,pres)
                    MVB V(montr,pres) {regulate}
                ⊟OD NP()
                    ⊟DT DTP()
                        DTPE PRON(univ,plu) {all}
                        DTCE ART(def) {the}
                    NPHD N(com,plu) {body functions}
                    PUNC PUNC(per) {.}
```

Figure 1. The ICE parse tree for Example 1.

This above example is taken from the fourth sentence in Text 23 of Genre W2B (non-academic natural sciences), with the syntactic tree structure in Figure 1. Each node in an ICE-GB tree comprises two labels: syntactic function and syntactic category. For example, SU,NP() represents 'syntactic subject realized by the noun phrase'. Similarly, NPPR AJP (*attru*) indicates an attributive adjective phrase performing the function of an NP premodifer. The leaf nodes, *i.e.* the lexical items, are enclosed within curly brackets. Example 1 is analysed as a main clause consisting of a subject and a verb, with three adverbials: two realized by the prepositional phrases '*from cell to cell*' and one realized by a non-finite present participial clause carrying messages which regulate all the body functions. Features associated with the adverbial clause indicate that it does not have an overt subordinator (*zsub*), that its main verb is present participial (*ingp*), and that this clause does not have an overt subject *(-su*). The detailed annotation thus indicates explicitly the category names such as the clause and the phrase type as well as their syntactic functions such as subject and adverbial. ICE-GB therefore allows for unambiguous retrieval of different types of clauses and phrases, as well as grammatical features.

4. Methodology

4.1. Machine Learning Device

Modern technologies of ATC include the use of statistical Machine Learning (ML) models such as Naïve Bayes (NB), Support Vector Machines (SVM), Decision Trees (DT) and K-Nearest-Neighbouring (KNN). Recent advances witness the rise of Deep Learning method in the subject domain with the advent of big data and Artificial Neural Networks, such as the Convolutional Neural Network (CNN) and the Recurrent Neural Network (RNN) (Lidy & Schindler, 2016; Sigtia & Dixon, 2014). A typical Machine Learning method, with a supervised mode, usually requires pre-defined labels of text types in the training data, and the construction of feature space for model-fitting and class prediction, as shown in Figure 2[2] below.

Figure 2 demonstrates the pipeline of ATC which involves data pre-processing, feature engineering, model fitting, parameter tuning and class prediction by constructing certain feature sets from the pre-labelled training data and makes predictions on the testing data with the same sets of features modelled from the test data (Bird, Klein, & Loper, 2009). Preliminary experiments on using the collection of classifiers (*e.g.* Naïve Bayes, Support Vector Machine, Decision Tree, Logistic Regression, K-Nearest-Neighbourhood) consistently showed a superior performance of the Naïve Bayes and Support Vector Machine classifiers and they are adopted in the current AGC tasks.

The *Naïve Bayes* (NB) classifier is widely adopted in machine learning due to its simplicity and fast speed of building the model, yet with impressive performances. Bayes classifiers assign the most likely class to a given

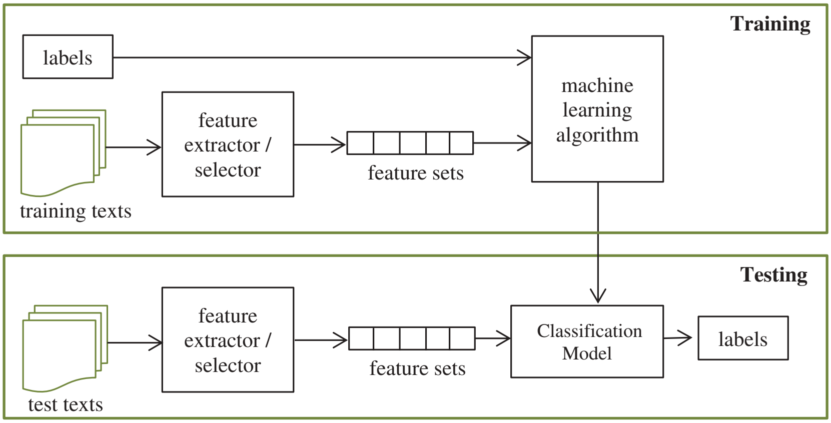


Figure 2. A diagram of machine learning for automatic text classification.

example described by its feature vector. Learning such classifiers can be greatly simplified by assuming that features are independent, that is, $\mathrm{P}(\boldsymbol{X}|C) = \prod_{i=1}^{n} P(X_i|C)$, where $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a feature vector and C is a class. Despite the naïve assumption, the NB classifier is remarkably successful in practice, often competing with much more sophisticated techniques. Li and Jain (1998) indicated that NB is good at dealing with the over-fitting problem and the performance improves with the number of features. Lewis (1992) and Witten, Frank, Hall, and Pal (2016) found that NB requires only a small number of training data to achieve good performance. Rish (2001) demonstrates that Bayes is not directly correlated with the degree of feature dependencies measured as the class conditional mutual information between the features. A better predictor of naive Bayes accuracy is the amount of information about the class that is lost because of the independence assumption.

Another classifier in this work is the ‘support vector machine’-based SMO (*Sequential Minimal Optimization*) model that is realized by John Platt’s pairwise classification model (Platt, 1998) which effectively solved the Quadratic Program (QP) by decomposing it into small sequences of minimal optimizations, as in Equation (1) below:

$$\begin{aligned} &\min_{\vec{\alpha}} \psi(\vec{\alpha}) = \min_{\vec{\alpha}} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j K\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right)\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i, \\ &0 \le \alpha_i \le C, \forall i, \\ &\sum_{i=1}^{N} y_i\alpha_i = 0. \end{aligned} \tag{1}$$

The Lagrange multipliers α_i are computed *via* a quadratic program. The non-linearities alter the quadratic form, but the dual objective function ψ is quadratic in α. In order to make the QP problem above be positive definite, the kernel function K must obey Mercer’s conditions. The Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient conditions for an optimal point of a positive definite QP problem. The QP problem is solved when, for all i:

$$\begin{aligned} &\alpha_i = 0 \leftrightarrow y_i u_i \ge 1, \\ &0 < \alpha_i < C \leftrightarrow y_i u_i = 1, \\ &\alpha_i = C \leftrightarrow y_i u_i \le 1 \end{aligned} \tag{2}$$

where u_i is the output of the SVM for the ith training example. The KKT conditions can be evaluated on one example at a time, which will be useful in the construction of the SMO algorithm.

SMO is also enabled for multi-class classification by using the ‘one *vs.* one’ algorithm. It is a very powerful classifying model that shows outstanding performance than many state-of-the-art classifiers. Joachims (1998) explained that SVM uses overfitting protection to ensure its well performance for dealing with features of high dimensionality, and it does not require parameter tuning to achieve high accuracy, as he puts it 'With their ability to generalize well in high dimensional feature spaces, SVMs eliminate the need for feature selection, making the application of text categorization considerably easier.'

Among the various ATC tasks, Automatic Genre Classification (AGC) specifically aims for the detection of a text style among candidates of predefined genres. Genre-related linguistic investigations and feature engineering techniques would be crucial for a successful genre-based retrieval, which shall be significantly different from a general text classification task.

4.2. Feature Engineering

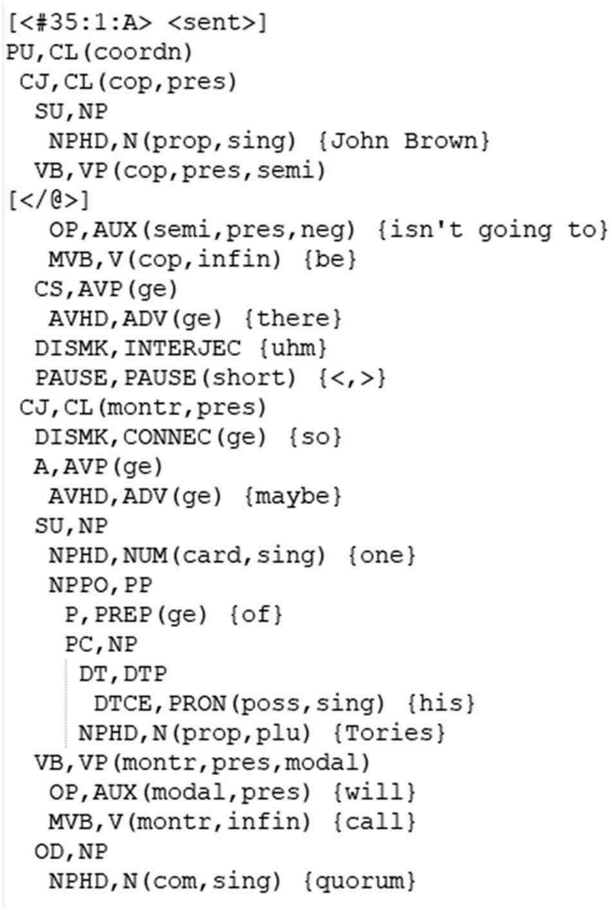
4.2.1. Internal Syntactic Representations

The composition of a complete linguistic unit/a chunk (*e.g.* a sentence) in a language can be described as a series of combination processes in which words combine into phrases and phrases combine to clauses and clauses combine into a complex sentence, if necessary. Each combination process of composing a larger linguistic unit displays certain linguistic forms and functions of shaping its own style. These respective combination processes certainly show important syntactic characteristics for separating texts of different kinds. As such, this paper proposes to investigate the structural representations for genre classification.

Being different from a bag of random word features that disregard syntactic information, internal structural representations encode rich syntactic information that might be very important characteristics for identifying different text genres (Fábregas, 2007). The structural representations in this paper include constituent orders (*e.g.* a simple example of linear order: SVO), syntactic categories (*e.g.* NP, AJP, PP, AVP) and functions (*e.g.* SU, DO, IO, A) for each node of a tree. This paper has innovatively investigated and made use of these structural representations for classifying texts of different genres, with the purpose of identifying indicative syntactic feature sets for AGC, and contribute to the genre classification problems of NLP in a broader sense.

Corresponding to the combination processes for composing a larger language unit, the structural representations of syntactic features in this paper include three macro-categories of information: 1) the clausal structures and 2) the phrasal structures and 3) the non-clauses. Each macro category will also be subdivided into different inner structures showing finer interpretations of linguistic composition and functions, as demonstrated below.

The following example is used to illustration the concept of the internal representation, with its parse tree in ICE-GB shown in Figure 3.



```
[<#35:1:A> <sent>]
PU,CL(coordn)
 CJ,CL(cop,pres)
  SU,NP
   NPHD,N(prop,sing) {John Brown}
  VB,VP(cop,pres,semi)
[</@>]
   OP,AUX(semi,pres,neg) {isn't going to}
   MVB,V(cop,infin) {be}
  CS,AVP(ge)
   AVHD,ADV(ge) {there}
  DISMK,INTERJEC {uhm}
  PAUSE,PAUSE(short) {<,>}
 CJ,CL(montr,pres)
  DISMK,CONNEC(ge) {so}
  A,AVP(ge)
   AVHD,ADV(ge) {maybe}
  SU,NP
   NPHD,NUM(card,sing) {one}
   NPPO,PP
    P,PREP(ge) {of}
    PC,NP
     DT,DTP
      DTCE,PRON(poss,sing) {his}
     NPHD,N(prop,plu) {Tories}
  VB,VP(montr,pres,modal)
   OP,AUX(modal,pres) {will}
   MVB,V(montr,infin) {call}
  OD,NP
   NPHD,N(com,sing) {quorum}
```

Figure 3. The ICE parse tree for Example 2.

Example 2. '*John Brown isn't going to be there uhm, so maybe one of his Tories will call quorum*'

s1a-068: [<#35:1:A> <sent>]

The internal representations of this particular tree will be exhaustively extracted based on the FSM model, as detailed in section 4.2.2., with the following internal structures for Example 2:

- **Clausal internal representations**

(1) PU,CL:CJ,CL-CJ,CL
(2) CJ,CL:SU,NP-VB,VP-CS,AVP
(3) CJ,CL:A,AVP-SU,NP-VB,VP-OD,NP

- **Phrasal internal representations**

(1) SU,NP:NPHD,N
(2) VB,VP:OP,AUX-MVB,V
(3) CS,AVP:AVHD,ADV
(4) A,AVP:AVHD,ADV
(5) SU,NP:NPHD,NUM-NPPO,PP
(6) NPPO,PP:P,PREP-PC,NP
(7) PC,NP:DT,DTP-NPHD,N
(8) DT,DTP:DTCE,PRON
(9) VB,VP:OP,AUX-MVB,V
(10) OD,NP:NPHD,N

Take the first clausal internal representation for instance, the linear form 'PU,CL:CJ,CL-CJ,CL' is interpreted as a clausal parse unit (the mother node of a tree) with two conjoined sub-clauses (the two daughter nodes). The left symbol (PU and CJ) before each comma in each node indicates its syntactic function; the right symbol (CL) after each comma indicates its syntactic category. An internal representation of a mother node is therefore composed by many daughter nodes which are connected by a hyphen. In addition, the position of each node in the linear form shows the structural order of a constituent. The whole linear form hence represents the internal structure of the mother node (PU,CL).

All the enlisted clausal and phrasal structures become overt linguistic cues of a text style in terms of constituent occurrences and embedding forms, which may provide useful predictors for genre classification. It is therefore interesting to see how these transformed linear syntactic features with rich structural information could affect the performance of AGC in comparison to pure word features. In order to automate the identification and construction of such internal feature vectors, we successfully implemented the following FSM-derived model (Wan & Fang, 2018).

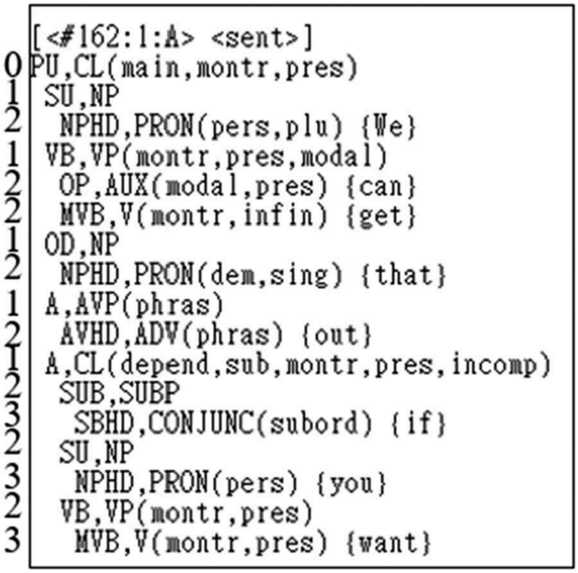
4.2.2. The FSM Model for Feature Construction

4.2.2.1. The Coding Principle. Prior to the construction of the internal syntactic feature is the identification of such representations with the coding procedure. The coding of each constituent in a tree is specially designed for its parsing scheme that can effectively index a certain internal structure, as displayed in Figure 4.

Take a simple example (Example 3):

Example 3: '*We can get that out if you want*' *s1a-006<#162:1:A>*

Its parse tree with embedding codes is shown in Figure 4 below.



```
 [<#162:1:A> <sent>]
0PU,CL(main,montr,pres)
1 SU,NP
2  NPHD,PRON(pers,plu) {We}
1 VB,VP(montr,pres,modal)
2  OP,AUX(modal,pres) {can}
2  MVB,V(montr,infin) {get}
1 OD,NP
2  NPHD,PRON(dem,sing) {that}
1 A,AVP(phras)
2  AVHD,ADV(phras) {out}
1 A,CL(depend,sub,montr,pres,incomp)
2  SUB,SUBP
3   SBHD,CONJUNC(subord) {if}
2  SU,NP
3   NPHD,PRON(pers) {you}
2  VB,VP(montr,pres)
3   MVB,V(montr,pres) {want}
```

Figure 4. The parse tree with embedding codes of Example 3.

The coded numbers in the left side of Figure 4 indicate the embedding levels of all the constituents in the tree. These numbers are very important indicators of indexing the internal structures of a target constituent. As we can see, the tree starts with a root node (PU,CL) with 0 embedding level. Branching after the root node, there are five daughter nodes (SU,NP|VB, VP|OD,NP|A,AVP|A,CL) with value 1 as the embedding level. Each of the nodes can be further embedded until it reaches a terminal node, a word (*e. g*. SBHD,CONJUNC {if}). After all the nodes of the tree are coded with their embedding levels, the identification and extraction of the wanted internal representations can be realized by the following method.

4.2.2.1. The FSM-derived Model. The idea of identifying the internal representations was inspired by the Finite State Machine model (Kishima & Ito, 1998; Selic, Gullekson, & Ward, 1994). It is an abstract machine that can be defined by a list of an initial state, a finite set of changed states and the conditions for each transition at any given time. The FSM can change from one state to another in response to some external inputs. The change from one state to another is called a transition. Figure 5 below illustrates how the internal structures of a NP (noun phrase) can be identified and extracted based on the FSM-derived method. The transition patterns of the FSM model resemble the structural transition states of a target constituent within a parse tree. The following diagram illustrates how the internal structures of a NP (noun phrase) can be identified and extracted based on the FSM-derived method.

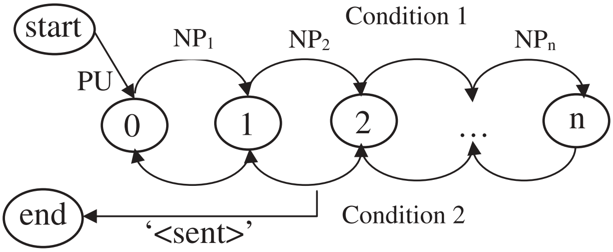


Figure 5. The FSM-derived method for constituent searching.

The identification of all the internal structures of a NP in each tree is triggered by meeting a parsing unit (PU) with an initial status of 0. The status increments to 1 when meeting the first NP and the identification process is activated. At the same time, the embedding number of the mother node is stored for a reference purpose. When condition 1 (a deeper embedding level of NP is spotted) is satisfied, the status continues to increment by 1 until reaching a terminal node. In the meanwhile, different variables will be assigned the respective embedding levels of all the NP mother nodes as reference points of identifying their internal nodes. In contrast, when condition 2 (an equal or smaller embedding level node is spotted) is satisfied, the status would decrease to one of the original status depending on the current embedding code, until it goes back to 0. The whole identification process will end whenever it comes across the sentence end marker (<sent>). The same process will be iteratively implemented until all the texts in the corpus are thoroughly searched. After this process, all the NPs together with their internal nodes in the corpus can be exhaustively identified and stored in lists of variables. The following part introduces how these internal structures can be extracted and computed.

4.2.2.3. The Extraction Process. This part describes how the internal representations of a constituent are extracted and converted into linear structural representations with frequency information through the following flow chart. The current-activated status is chosen as *i* for illustration. The flow chart in Figure 6 shows the procedure of the identification, conversion and construction of the internal representations of a target constituent. The ellipsis represents any other cycle of status from the current status triggered by certain transition condition. After all the syntactic trees have been completely searched, the status may change freely between 0 and n. The extracted structural nodes will then be converted into linear representations and their frequency distribution will be calculated automatically by using the FreqDist module in NLTK. These frequencies serve as the attribute values of the feature sets for the AGC experiment.

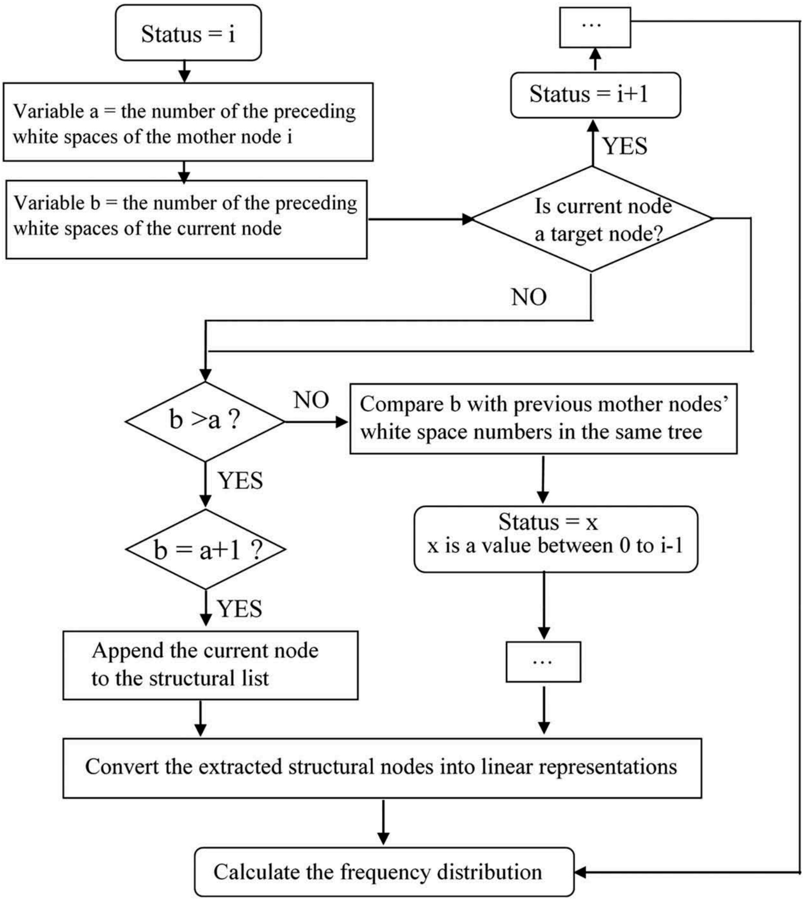

Status = i
Variable a = the number of the preceding white spaces of the mother node i
Variable b = the number of the preceding white spaces of the current node
Is current node a target node?
YES
Status = i+1
…
NO
b >a ?
NO
Compare b with previous mother nodes' white space numbers in the same tree
YES
b = a+1 ?
Status = x
x is a value between 0 to i-1
YES
Append the current node to the structural list
…
Convert the extracted structural nodes into linear representations
Calculate the frequency distribution

Figure 6. The flow chart of the extraction process.

4.2.3. The Feature Sets

In the experiment, we use the bag-of-words as the baseline feature set. Besides, 14 syntactic feature sets are constructed (as shown in the above flow chart) for studying the different roles of structural features for genre classification. Detailed information about the feature sets are shown in Figure 7 below:

BOW: There are 33,832 types of words extracted from the whole corpus, after excluding the stop words and punctuations. These words contribute a feature matrix of 33,833 attributes and 500 instances to the AGC tasks, serving as a baseline of the AGC tasks.

Fused: There are 30,149 types of internal structures extracted from the corpus. Adding the bag-of-words features, the fuse feature set incorporates

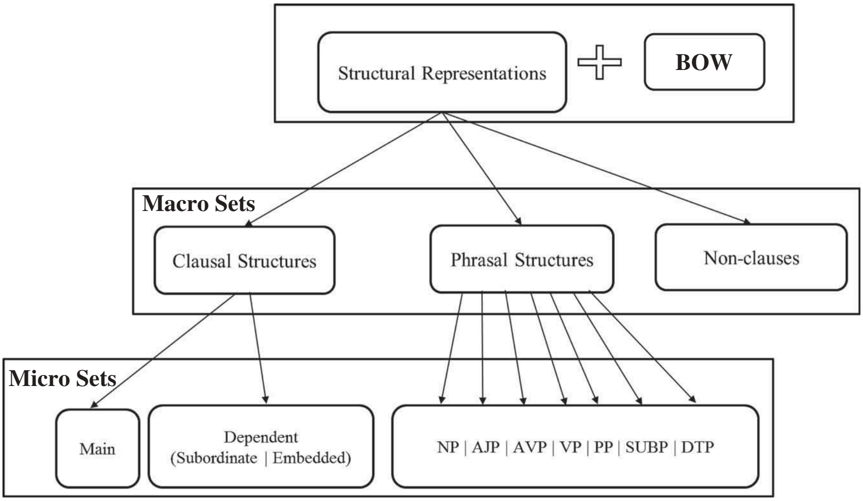


Figure 7. The relation diagram of the 15 feature sets.

both lexical and syntactic information with a feature matrix of 63,982 attributes and 500 instances.

Macro: The macro sets of internal structures include three sets of features: 1) a set of clausal structures (CL: 21,811 types); 2) a set of phrasal structures (Phr: 7,113 types); and 3) a set of non-clause structures (NONCL: 1,225 types).

Micro: The micro sets of internal structures include: 1) three subsets of clausal internal structures: Main clauses (Main: 11,828 types), Subordinate clauses (Sub: 7,015 types) and Embedded clauses (Emb: 3,050 types); and 2) seven subsets of phrasal internal structures: NP (4,602 types), AJP (496 types), AVP (types), VP (475 types), PP (850 types), SUBP (types), and DTP (283 types).

The above syntactic features of triple levels of granularity are constructed to testify the discriminativeness of the various constituents for indicating text genres. Feature vectors are constructed by using the *tf* (term frequency) of the various words or structures. The following graph shows the hierarchical relation of the adopted feature sets.

5. Experiments and Results

5.1. Experimental Setup

The ICE-GB corpus contains 500 text samples which constitute 500 instances for all the classification tasks. Four classification tasks are implemented for each feature corresponding to the four division of genre classes, *i.e.* SW2, SW5, SW11, and SW32, where 'S' stands for 'Speech' and 'W' stands for 'Writing'. The numbers correspond to the genre categories of different granularities (cf. Appendix A). The classification task of predicting the two broad genres (speech vs. writing) is represented by 'SW2′; the five macro-genres (dialogues, monologues, mixed, printed and non-printed) represented by 'SW5′; the eleven micro-genres (*e.g.* private, public, scripted, unscripted) represented by 'SW11′; and the 32 mini-genres (*e.g.* conversations, calls, classroom lessons) represented by 'SW32′. Feature vectors of all the classification tasks are converted according to the term frequencies (*tf*) in each sample text. No feature selection is implemented due to the moderate feature dimension of the 15 feature sets (each feature dimension is less than 70,000), and the two classifiers are both capable of dealing with features of this size. The occurring frequencies can be regarded as normalized data *per se* because of the balanced samples (2,000 words *per* sample) in the corpus. Training and testing are conducted with 10-fold cross-validation. Classification results are reported with evaluation metrics of Precision, Recall, and F-score.

5.2. Results and Discussion

5.2.1. Classification Performance

In the AGC experiment, 15 feature sets have been fitted to two statistical classifiers (NB and SMO), and each feature set has been experimented in four individual tasks (*i.e.* SW2, SW5, SW11 and SW32) for both classifiers. The main classification results in average are shown in Table 2 below.

Table 2. The overall *W Avg. F* of all feature sets in AGC.

		NB				SMO			
W Avg. F[a]		SW2	SW5	SW11	SW32	SW2	SW5	SW11	SW32
BOW		92.40	81.40	69.20	54.00	93.80	79.00	61.50	30.60
Fused		92.60	82.70*	71.90*	54.10	95.20*	83.50**	66.50**	39.80**
Macro	CL	84.70	71.30	53.90	34.30	91.20	72.30	52.00	26.00
	Phr	82.70	74.60	61.80	43.10	89.30	76.70	61.20	39.20**
	NONCL	79.90	70.30	52.80	35.60	77.90	61.80	42.30	25.50
Micro	Main	85.50	74.60	56.90	37.50	91.80	72.50	56.00	29.20
	Sub	81.90	57.80	41.40	24.70	80.30	59.90	42.20	18.10
	Emb	77.30	56.10	40.10	24.20	75.80	56.80	39.40	15.40
	NP	80.30	69.70	58.30	41.20	83.80	71.50	58.30	35.70**
	PP	77.20	63.80	49.40	33.90	79.30	62.10	46.80	27.40
	DTP	79.40	60.40	43.10	25.20	80.70	62.10	46.40	26.80
	AJP	75.00	57.60	44.70	26.20	78.30	60.30	40.60	26.20
	VP	79.60	60.70	44.40	24.50	78.60	58.40	41.60	24.00
	AVP	68.30	55.00	38.80	28.20	76.20	52.60	34.70	22.60
	SUBP	61.70	39.70	25.30	10.50	52.70	37.50	23.20	6.70

a: It represents the weighted average F-score*100 (***W Avg. F***) in predicting the respective genre classes.
_: Underlined scores are those feature sets that outperform the baseline feature set (BOW).
*: Single asterisk marks the significantly superior ***W Avg. F*** with p <0.05 compared to the baseline set.
: Double asterisks mark the significantly superior ***W Avg. F* with p <0.001compared to the baseline set.

The weighted average F-score of each group suggests that syntactic features demonstrate no obvious advantage than the baseline feature unless they are combined. The fused set with both syntactic and lexical features outperforms the BOW set in ALL of the eight classification tasks for both classifiers, and the performance discrepancy is significant ($F_{\Delta max.}$ = 9.2%, *sig.* = 0.001) in the cases with double-asterisks. In addition, NB shows better classification performance than SMO in average, but SMO outperforms NB in SW2 for most feature sets.

Several implications can be derived: 1) Internal syntactic features are not more discriminative than BOW features but a combination of structural and lexical features would significantly help the classification performance. 2) Internal syntactic features tend to outperform the BOW features to a greater extent when the genre classes are more granulated, such as in SW32, showing a more robust property of using complex features in finer genre classification. 3) The positive F-score discrepancy of the fused set over the baseline set is greater for the SMO model which seems to echo Joachims (1998)'s claim about the capability of support vector machines in dealing with high-dimensionality features.

The classification results of using the macro sets of internal structures show that clausal, non-clausal and phrasal structures in separate do not compete with lexical features for genre classification, except for the case of using the phrasal set in SW32. The results in Table 2 generally suggest that phrasal structures are more discriminative than the other structural features for genre classification. The results of the micro sets of internal structures strengthen the observation that individual syntactic features are not more useful than lexical features, except for the NP phrase in SW32 (*W Avg. F* = 35.70, *sig.* = 0.000), which indicates the outstanding prediction power of noun phrasal structures for genre classification.

5.2.2. Discriminativeness Rank and Analysis

In this section, the discriminativeness rank of the 15 feature sets for AGC will be sorted out according to the mean F-score of each feature set among the S2, SW5, SW11, SW32 tasks, so as to provide a reference for stylometry studies of syntactic investigations. In addition, we aim to find out whether the AGC results correlate to the feature sets' number of tokens (or tokens of structural forms), types (or types of structural forms), and type-token ratio (TTR, or structural TTR) to provide a possible account for the varied discriminativeness of the various syntactic structures. Data are shown in Table 3 below. The '$F_{\_Mean}$' column lists the mean F-scores of the 15 feature sets with averaged scores of the two classifiers. The table is ranked according to the '$F_{\_Mean}$' values in descending order.

The discriminative rank shows the usefulness of phrasal features as a whole (with noun phrases as a distinct micro-set), clausal features as a

Table 3. Discriminativeness rank of the 15 feature sets.

Rank	Feature Sets	Token	Type	TTR	$F_{\_Mean}$
1	**Fused**	997,499	42,686	4.280	0.489
2	**BOW**	510,464	33,832	6.630	0.456
3	**Phr**	830,647	7,113	0.860	0.410
4	**NP**	317,556	4,061	1.280	0.379
5	**Main**	76,403	11,831	15.480	0.343
6	**CL**	145,935	21,811	14.950	0.313
7	**PP**	104,014	849	0.820	0.280
8	**NONCL**	20,917	1,225	5.860	0.278
9	**AJP**	62,021	495	0.800	0.265
10	**DTP**	118,603	281	0.240	0.258
11	**VP**	140,058	474	0.340	0.243
12	**Sub**	42,556	7,020	16.500	0.231
13	**AVP**	67,391	50	0.070	0.225
14	**Emb**	24,756	3,054	12.340	0.200
15	**SUBP**	20,321	60	0.300	0.097

whole (with main clauses as a distinct micro-set), and the prepositional phrases for AGC in addition to the distinct performance of the fused set. The Pearson Correlation Test between $F_{\_Mean}$ and variables of Token, Type, TTR shows that features of high tokens and types tend to produce better classification results for the AGC task ('token vs. F': correlation = 0.799, *sig.*= 0.000; 'type vs. F': correlation = 0.700, *sig.*= 0.001). However, it does not show much correlation with the TTR (correlation = 0.0337, *sig.*= 0.542) of the features, which represents the diversity of the lexical or syntactic features. The statistical results seem to suggest a general and simplified principle that more tokens or types of features could promise a more successful classification result. But this does not apply to all cases such as VP and DTP. The discriminativeness of a feature set for ATC is however a complex problem that needs further investigation. In order to position the distinct internal structure of the phrasal features, *esp*. the noun phrases, a further analysis to the phrasal set will be conducted in Section 5.2.3, to account for its outstanding performance in AGC.

On the basis of the experimental results, we come to a few implications: Fine-grained internal features that encode rich syntactic information is proved to be useful for AGC in combination with lexical features. Although most sub-types of the internal structures show equal or inferior performance compared to BOW, the phrasal features, esp. noun phrases demonstrate great discriminativeness for differentiating multiple genres classes. Main clauses alone have also shown a close performance to BOW in discriminating text genres. Model-fitting-wise, syntactic features tend to correlate positively to the SMO model, while lexical features to the NB model. For example, the comprehensive set outperforms BOW with a significant p-value for the SMO model, while the advantage is minor for the NB model; phrasal structures and noun phrase structures are comparable to lexical features for the SMO model, but they are less outstanding when the NB model is adopted. This could probably be accounted for by the independence degree, dimensionality and sparseness of the different kinds of feature vectors, since lexical features are higher dimensional, sparser and more mutually-independent, while syntactic features are less sparse and less mutually independent.

5.2.3. PCA Analysis to the Subsets of Phrasal Features

This section aims to provide a Principle Component Analysis (PCA) to the seven phrasal feature sets in terms of F-score performance in relation to the five macro genre classes. Results are taken from the SW32 of the Phr set with the SMO classifier. Appendix D contains the original data. This PCA analysis shall find out the most salient components for differentiating the various genre classes, as well as displaying the correlation of the seven types of phrases with the respective genres. Table 4 below shows the statistics about the PCA analysis.

The above data in Table 4 clearly show that the noun phrases (NP) and prepositional phrases (PP) are the two principle components (67.4%) which contribute the most to the differentiation among the different genre classes. It indicates the salient roles of the two phrasal features for genre classification, with NP being the most principle one. The following plot in Figure 8 further shows the correlation between the several phrases and the two main components which are grouped by the five macrogenre classes.

Figure 8 presents three groups of data despite that we input five classes of genres, where the Mixed and Non-printed documents are quite sparse and there are too few points to calculate an ellipse. From a general view of the plot, the Printed writing genre positively correlate with PC1 (NP), showing that noun phrases are effective to identify high formality texts; Monologues and Dialogues show no clear correlation with PC1 (NP), suggesting that noun phrases are not so effective in telling apart speech genre types. However, the case for PC2 (PP) seems opposite: Printed writing materials show no obvious correlation with prepositional phrases, while Monologues and Dialogues tend to be negatively related to prepositional phrases. The analysis hence implies that noun phrases are most effective for identifying printed text types with high formality, while prepositional phrases are useful features for identifying speaking text types with low formality. Similar analysis could also be conducted for the three clausal types (main clauses, subordinate clauses, and embedded

Table 4. Principle component analysis of the seven phrases.

	PC1(NP)	**PC2(PP)**	PC3(DTP)	PC4(AJP)	PC5(VP)	PC6(AVP)	PC7(SUBP)
Importance of components:							
Standard deviation	2.1580	0.8077	0.75100	0.60875	0.56391	0.49077	0.44409
Proportion of Variance	**0.6653**	**0.0932**	0.08057	0.05294	0.04543	0.03441	0.02817
Cumulative Proportion	0.6653	0.7585	0.83905	0.89199	0.93742	0.97183	1.00000

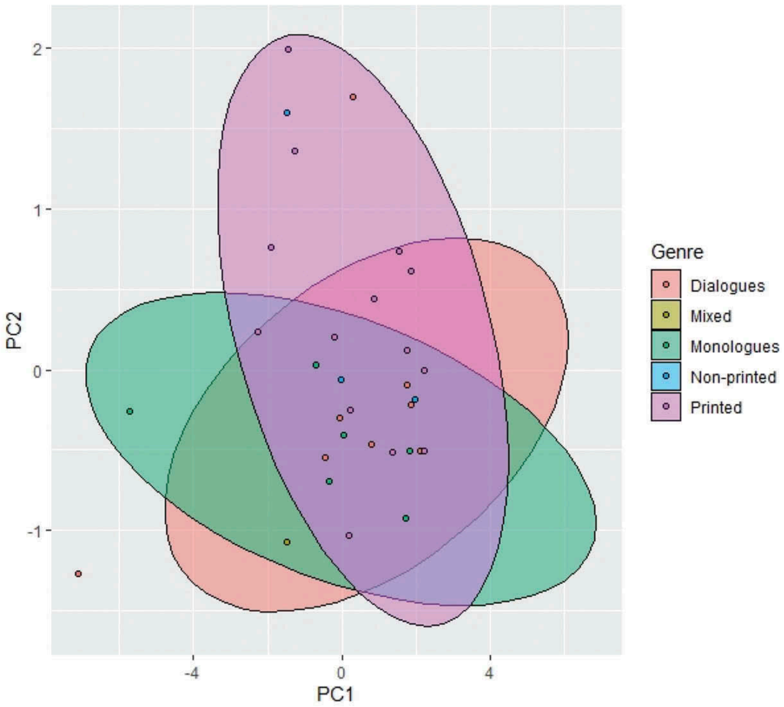


Figure 8. The plot of correlation between the five genres and the primary two components.

clauses) in future, as well as its 22 finer subtypes (Wan, 2017). The finding seems to link the discriminativeness of noun phrases in AGC to text formality. In order to look further into the salient internal structures within noun phrases, we provide the following analysis specific to the NP type in terms of internal structure types and frequency.

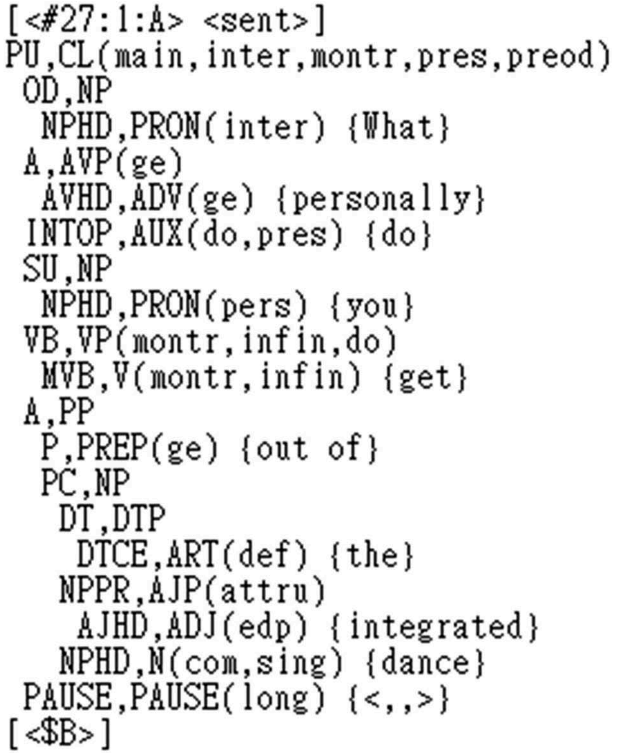
5.2.3.1. NP (Noun Phrase) Structures. Noun phrases are prevalent in the corpus. They serve as important syntactic functions with different syntactic positions. See the following example:

Example 4. '*What personally do you get out of the integrated dance.*'

S1A001<#27:1:A>

The parse tree of the sentence in example 4 is shown in Figure 9 below.

In Figure 9, there are three noun phrases: 1) '*What*' functions as the direct object of the verb 'get' 2) '*you*' functions as the subject of the main clause 3) '*the integrated dance*' functions as the prepositional complement.



```
[<#27:1:A> <sent>]
PU,CL(main,inter,montr,pres,preod)
 OD,NP
  NPHD,PRON(inter) {What}
 A,AVP(ge)
  AVHD,ADV(ge) {personally}
 INTOP,AUX(do,pres) {do}
 SU,NP
  NPHD,PRON(pers) {you}
 VB,VP(montr,infin,do)
  MVB,V(montr,infin) {get}
 A,PP
  P,PREP(ge) {out of}
  PC,NP
   DT,DTP
    DTCE,ART(def) {the}
   NPPR,AJP(attru)
    AJHD,ADJ(edp) {integrated}
   NPHD,N(com,sing) {dance}
 PAUSE,PAUSE(long) {<,,>}
[<$B>]
```

Figure 9. The tree diagram of example 4.

Their respective structural representations are: 1) ‘OD,NP:NPHD,PRON’ 2) ‘SU,NP:NPHD,PRON’ 3) ‘PC,NP:DT,DTP-NPPR,AJP-NPHD,N’.

In the corpus, there are 4,601 types of noun phrase structures out of the 317,556 tokens of noun phrases, and it demonstrates the greatest structural variance. This might be one of the reasons to account for NP’s discriminativeness of genre classification. In order to find out the most prominent internal structures of NPs, we extracted all the linear representations and ranked them according to term frequency. Table 5 below lists the top 10 structural representations of NP in descending order of frequencies with examples for illustration. The highlighted terms correspond to its structural representation of NP.

Out of the top 10 structural forms of NPs, there is a large overlap of the forms which points to the high occurrence of pronouns functioning as a subject, and prepositional complements with various inner structures. As 60% of the corpus of is composed by speech documents, it might explain why pronouns and prepositional complements are prevalent in the corpus and become salient features for genre classification.

Table 5. Top 10 inner structures of noun phrases.

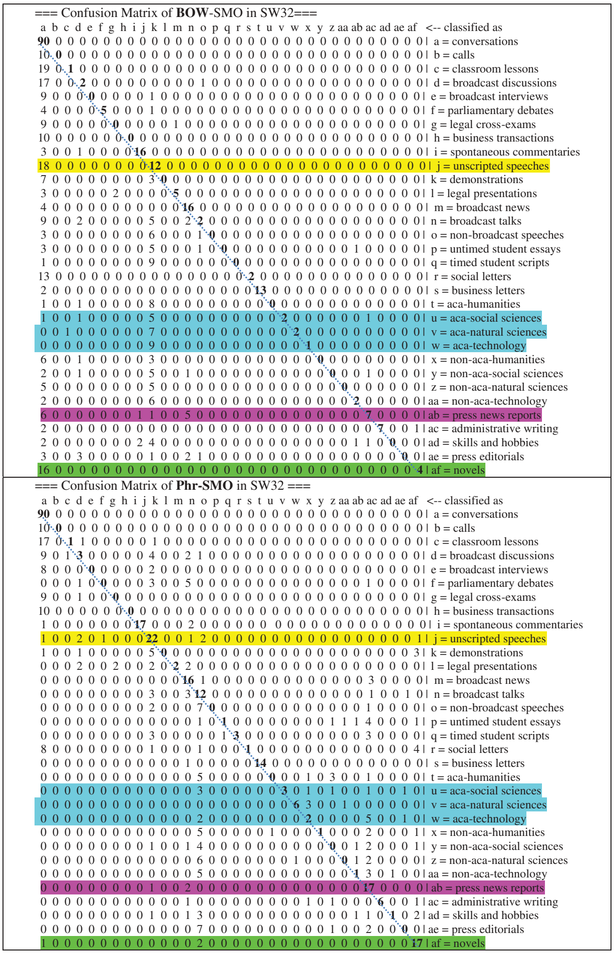
Rank	Structural Representations (NP)	Examples	Frequency
1	SU,NP:NPHD,PRON	**I** did it for about half a term.	68,592
2	PC,NP:DT,DTP-NPHD,N	I did it for about **half a term**.	25,534
3	PC,NP:NPHD,N	I was doing a unit of **English**.	20,608
4	SU,NP:DT,DTP-NPHD,N	And **the second one** is excellent.	12,707
5	CJ,NP:NPHD,N	Be leaving about half five or **something** I think.	10,426
6	OD,NP:NPHD,PRON	We didn't come and pick **her** up.	10,412
7	OD,NP:DT,DTP-NPHD,N	I don't see it makes **a difference**,	9,701
8	SU,NP:NPHD,N	**Part** of the reason will be we don't have to have a picnic.	9,265
9	PC,NP:NPHD,PRON	You are stuck with **it**, aren't you?	7,362
10	PC,NP:DT,DTP-NPHD,N-NPPO,PP	We could come round with **a bottle of something.**	7,260

5.2.4. *Error Analysis*

The above statistical analysis to noun phrases and its subtypes implies the interesting role of nominal structures for genre classification in general. However, it is still unclear how phrasal features outperform BOWs in relation to the specific genre types, and what genre types are most confusing for the respective feature types. This might be useful for genre-specific document retrieval tasks or stylometry studies. Table 6 below displays the confusion matrices of BOW and the Phr set in SW32 with the SMO classifier. As the NB classifier shows similar pattern to SMO, we finally chose the following results for simplicity.

In the confusion matrices, the numbers in the diagonal line are the correct predictions and the others are false predictions. Significantly outstanding performances of Phr in comparison to BOW was highlighted with different colours which are used to group different properties of the genre classes. For instance, class j (unscripted speeches), marked in yellow, was correctly predicted 22 times out of the 30 instances by Phr, which is 10 more times than BOW, indicating phrasal features' great discriminativeness for identifying informal (unscripted) spoken genre (speeches). Classes u, v, w, which are marked in blue, were also predicted more accurately by Phr. Interestingly, the three classes are all academic writings related to hard science that are highly formal and complex. Class ab (press news reports), marked in purple, was also predicted more accurately by Phr (17 vs. 7), which shows a style of factual report. Finally, class af (novels), marked in green, also shows significantly better predictions in the case of Phr (17 vs. 4), which demonstrates a narrative style. Notably, the four groups of genre classes which can be more accurately identified by the phrasal features presents an incremental degree of formality (Wan & Fang, 2018) in the continuum of language complexity and shows distinct structural characteristics (cf. adverbial clauses in Fang, 2006). In addition, the current finding about the role of phrasal features for genre

Table 6. Confusion matrices of BOW-SMO and Phr-SMO in SW32.



```
=== Confusion Matrix of BOW-SMO in SW32===
 a b c d e f g h i j k l m n o p q r s t u v w x y z aa ab ac ad ae af   <-- classified as
90 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | a = conversations
10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | b = calls
19 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | c = classroom lessons
17 0 0 2 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | d = broadcast discussions
 9 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | e = broadcast interviews
 4 0 0 0 0 5 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | f = parliamentary debates
 9 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | g = legal cross-exams
10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | h = business transactions
 3 0 0 1 0 0 0 0 16 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | i = spontaneous commentaries
18 0 0 0 0 0 0 0 0 12 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | j = unscripted speeches
 7 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | k = demonstrations
 3 0 0 0 0 0 2 0 0 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | l = legal presentations
 4 0 0 0 0 0 0 0 0 0 0 0 16 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | m = broadcast news
 9 0 0 2 0 0 0 0 0 5 0 0 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | n = broadcast talks
 3 0 0 0 0 0 0 0 0 6 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | o = non-broadcast speeches
 3 0 0 0 0 0 0 0 0 5 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 | p = untimed student essays
 1 0 0 0 0 0 0 0 0 9 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | q = timed student scripts
13 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | r = social letters
 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 13 0 0 0 0 0 0 0 0 0 0 0 0 0 | s = business letters
 1 0 0 1 0 0 0 0 0 8 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | t = aca-humanities
 1 0 0 1 0 0 0 0 0 5 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 1 0 0 0 0 | u = aca-social sciences
 0 0 1 0 0 0 0 0 0 7 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 | v = aca-natural sciences
 0 0 0 0 0 0 0 0 0 9 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 | w = aca-technology
 6 0 0 1 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | x = non-aca-humanities
 2 0 0 1 0 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 | y = non-aca-social sciences
 5 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | z = non-aca-natural sciences
 2 0 0 0 0 0 0 0 0 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 | aa = non-aca-technology
 6 0 0 0 0 0 0 0 1 1 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0 0 0 | ab = press news reports
 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0 1 | ac = administrative writing
 2 0 0 0 0 0 0 0 2 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 | ad = skills and hobbies
 3 0 0 3 0 0 0 0 0 1 0 0 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | ae = press editorials
16 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 | af = novels
=== Confusion Matrix of Phr-SMO in SW32 ===
 a b c d e f g h i j k l m n o p q r s t u v w x y z aa ab ac ad ae af   <-- classified as
90 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | a = conversations
10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | b = calls
17 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | c = classroom lessons
 9 0 1 3 0 0 0 0 0 4 0 0 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | d = broadcast discussions
 8 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | e = broadcast interviews
 0 0 0 1 0 0 0 0 0 3 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 | f = parliamentary debates
 9 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | g = legal cross-exams
10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | h = business transactions
 1 0 0 0 0 0 0 0 17 0 0 0 2 0 0 0 0 0 0  0 0 0 0 0 0 0 0 0 0 0 0 0 | i = spontaneous commentaries
 1 0 0 2 0 1 0 0 0 22 0 0 1 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 | j = unscripted speeches
 1 0 0 1 0 0 0 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 | k = demonstrations
 0 0 0 2 0 0 2 0 0 2 0 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 | l = legal presentations
 0 0 0 0 0 0 0 0 0 0 0 0 16 1 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 | m = broadcast news
 0 0 0 0 0 0 0 0 0 3 0 0 3 12 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 | n = broadcast talks
 0 0 0 0 0 0 0 0 0 2 0 0 0 7 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 | o = non-broadcast speeches
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 1 1 4 0 0 0 1 | p = untimed student essays
 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 1 3 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 | q = timed student scripts
 8 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 4 | r = social letters
 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 14 0 0 0 0 0 0 0 0 0 0 0 0 0 | s = business letters
 0 0 0 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 0 0 1 0 3 0 0 1 0 0 0 0 | t = aca-humanities
 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 3 0 1 0 1 0 0 1 0 0 1 0 | u = aca-social sciences
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 6 3 0 0 1 0 0 0 0 0 0 | v = aca-natural sciences
 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 2 0 0 0 0 5 0 0 1 0 | w = aca-technology
 0 0 0 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 1 0 0 0 1 0 0 0 2 0 0 0 1 | x = non-aca-humanities
 0 0 0 0 0 0 0 0 0 1 0 0 1 4 0 0 0 0 0 0 0 0 0 0 0 0 1 2 0 0 0 1 | y = non-aca-social sciences
 0 0 0 0 0 0 0 0 0 0 0 0 0 6 0 0 0 0 0 0 0 1 0 0 0 0 1 2 0 0 0 0 | z = non-aca-natural sciences
 0 0 0 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 1 3 0 1 0 0 | aa = non-aca-technology
 0 0 0 0 0 0 0 0 0 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 17 0 0 0 0 | ab = press news reports
 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 6 0 0 1 | ac = administrative writing
 0 0 0 0 0 0 0 0 0 1 0 0 1 3 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 2 | ad = skills and hobbies
 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0 0 0 0 0 0 0 0 0 1 0 0 2 0 0 0 0 | ae = press editorials
 1 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 17 | af = novels
```

classification does not only support Stamatatos et al. (2000)'s work which found NP structures' usefulness for AGC and also provides deeper analysis and interpretations to the interaction between various internal structures of phrases and genres of finer granularity, as demonstrated in the previous sections.

6. Conclusion

The present study has innovatively modelled 14 fine-grained syntactic internal features of triple layers for genre classification through machine learning techniques. FSM-derived model was used for feature construction. An incremental analysis to the classification results in terms of feature discriminativeness, prediction errors, and principle components was conducted, with the purpose of discovering the most useful syntactic features for different genre types, as well as providing sound explanations to the multiple–dimensional interactions between syntactic features and genre types.

Performance wise, we found internal syntactic features are not as discriminative as BOW individually but a combination of structural and lexical features would significantly help the classification performance. It suggests the effectiveness of conjoining complex features with simple features in genre classification. Although most sub-types of the internal structures show equal or inferior performance to BOW, the phrasal features, *esp.* noun phrases demonstrate great discriminativeness for differentiating most genres classes. The PCA analysis to the seven sub-types of phrasal features indicates the salient roles of NP and PP for AGC, with NP being the most principle component for identifying printed texts of high formality, while prepositional phrases are useful for identifying speeches of low formality.

Model-fitting wise, syntactic features tend to fit more to the SMO model, while lexical features to the NB model. This could probably be accounted for by the differences of lexical and syntactic features in terms of independence degree, dimensionality and sparseness. In addition, the superior performance of the fused set over the baseline set is greater for the SMO model which echoes Joachims (1998)'s claim about the capability of support vector machines in dealing with high-dimensionality features.

Finally, analysis to the confusion matrices found that phrasal features demonstrate great usefulness for identifying four groups of genre classes, *i. e.* unscripted speeches, fiction, news reports and academic writing, all distributed with distinct structural characteristics and they demonstrate an incremental degree of formality in the continuum of language complexity. The current work comprehensively studied the internal syntactic features for genre detection and laid a foundation for future studies in syntactic stylometry and complexity.

Notes

1. The sentence source code in ICE-GB.
2. The diagram is adapted from (Bird et al., 2009).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Hong Kong Polytechnic University [4-ZZKE].

ORCID

Mingyu Wan http://orcid.org/0000-0003-0083-5895
Chu-Ren Huang http://orcid.org/0000-0002-8526-5520

References

Bekkerman, R., & Allan, J. (2004). *Using bigrams in text categorization*. CIIR Technical Report IR-408 Center of Intelligent Information Retrieval. USA: University of Massachusetts Amherst.

Biber, D. (1988). *Variations across speech and writing*. Cambridge, UK: Cambridge University Press.

Biber, D. (1992). The multidimensional approach to linguistic analyses of genre variation: An overview of methodology and finding. *Computers in the Humanities*, *26*(5–6), 331–347.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge, UK: Cambridge University Press.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. Sebastopol, USA: O'Reilly Media.

Chen, K. J., Luo, C. C., Gao, Z. M., Chang, M. C., Chen, F. Y., Chen, C. R., & Huang, C. R. (1999). The CKIP Chinese Treebank. In *Journêes ATALA sur les Corpus annotes pour la syntaxe*. Talana, Paris VII..

Fábregas, A. (2007). The internal syntactic structure of relational adjectives. *Probus*, *19*(1), 1–36.

Fabrizio, S. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, *34*(1), 1–47.

Fang, A. C. (1996). The survey parser: Design and development. In Sidney Greenbaum (Ed.), *Comparing English world wide: The international corpus of English* (pp. 142–160). Oxford, UK: Clarendon.

Fang, A. C. (2006). A corpus-based empirical account of adverbial clauses across speech and writing in contemporary british english. In Salakoski, T. (ed.), *Advances in natural language processing* (pp. 32–43). Heidelberg, Berlin: Springer.

Fang, C. A., & Cao, J. (2015). *Text genres and registers: The computation of linguistic features*. New York: Springer, Heidelberg.

Fürnkranz, J. (1998). A study using n-gram features for text categorization. *Austrian Research Institute for Artificial Intelligence*, *3*(1998), 1–10.

Halliday, M., Matthiessen, C. M., & Matthiessen, C. (2014). *An introduction to functional grammar*. London: Edward Arnold.

Hou, R., & Huang, C. R. (in press). Classification of regional and genre varieties of Chinese: A correspondence analysis approach based on comparable balanced corpora. *Journal of Natural Language Engineering*.

Hou, R., Huang, C. R., Ahrens, K., & Lee, Y. M. (2019). Linguistic characteristics of chinese register based on the Menzerath – Altmann law and text clustering. *Digital Scholarship in the Humanities*, fqz005. doi:10.1093/llc/fqz005

Hou, R., Huang, C. R., Do, H. S., & Liu, H. (2017a). A study on correlation between Chinese sentence and constituting clauses based on the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, *24*(4), 350–366.

Hou, R., Huang, C. R., & Liu, H. (2017b). A study on Chinese register characteristics based on regression analysis and text clustering. *Corpus Linguistics and Linguistic Theory*. AOP. doi:10.1515/cllt-2016-0062

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features.* In *Proceedings of European Conference of Machine Learning* (pp. 137–142). Berlin Heidelberg: Springer-Verlag.

Karlgren, J. (2000). *Stylistic experiments for information retrieval* (Doctoral thesis). Stockholm University.

Karlgren, J. (2004). The whys and wherefores for studying textual genre computationally. In *Proceedings of AAAI fall symposium on style and meaning in language, art and music*, Arlington.

Karlgren, J., & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th international conference on computational linguistics (COLING 94)* (pp. 1071–1075). Kyoto, Japan.

Kessler, B., Nunberg, G., & Sch¨utze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th annual meeting of the association for computational linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 32–38). Madrid, Spain.

Kishima, S., & Ito, K. (1998). *U.S. Information processing apparatus using finite state machine. Patent No. 5,790,898.* Washington, DC: U.S. Patent and Trademark Office.

Lee, Y. B., & Myaeng, S. H. (2002). Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 145–150). New York, USA: ACM.

Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 37–50). Copenhagen, Denmark: ACM.

Li, Y. H., & Jain, A. K. (1998). Classification of text documents. *The Computer Journal*, *41*(8), 537–546.

Lidy, T., & Schindler, A. (2016). Parallel convolutional neural networks for music genre and mood classification. In *Proceedings of MIREX2016* (pp. 1–4). *New York, USA.*.

Lim, C. S., Lee, K. J., & Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management*, *41* (5), 1263–1276.

Lindemann, C., & Littig, L. (2006). Coarse-grained classification of web sites by their structural properties. In *Proceedings of the 8th annual ACM international workshop on Web information and data management* (pp. 35–42). Arlington, Virginia: ACM.

Liu, H., & Huang, C. R. (2016). EVALution-MAN 2.0: Expand the evaluation dataset for vector space models. In *Workshop on Chinese Lexical Semantics, LNCS* 10085 (pp. 261–268). New York, NY: Springer International Publishing.

Liu, M. C., & Wan, M. Y. 刘美君, 万明瑜. (2019). 中文动词及分类研究: 中文动词词汇语义网的构建及应用 [Mandarin verbs and its classification: The construction of Mandarin VerbNet and its NLP application]. *辞书研究* [Lexicographical Studies], *2*, 42–60.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics* (pp. 171–189). Berlin, Heidelberg: Springer.

Martin, J. R. (1984). *Language, register and genre in children's writing*. Geelong, Australia: Deaking UP.

Mehler, A., Geibel, P., & Pustylnikov, O. (2007). Structural classifiers of text types: Towards a novel model of text representation. *LDV Forum*, *22*(2), 51–66.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. ICLR 2013 (vol. 1301.3781), Scottsdale, Arizona, USA.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

Nanni, L., Costa, Y. M., Lumini, A., Kim, M. Y., & Baek, S. R. (2016). Combining visual and acoustic features for music genre classification. *Expert Systems with Applications*, *45*, 108–117.

Neergaard, K. D., & Huang, C. R. (2019). Constructing the mandarin phonological network: Novel syllable inventory used to identify schematic segmentation. *Complexity*, *2019*, 21. Article ID 6979830.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *MSRTR: Microsoft Research*, *3*(1), 88–95.

Pustylnikov, O. (2006). *How much information is provided by text structure? Automatic text classification using structural features* (Doctoral dissertation, Master thesis). University of Bielefeld, Germany.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Pearson Longman.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22), 41–46.

Scaringella, N., Zoia, G., & Mlynek, D. (2006). Automatic genre classification of music content: A survey. *IEEE Signal Processing Magazine*, *23*(2), 133–141.

Selic, B., Gullekson, G., & Ward, P. T. (1994). *Real-time object-oriented modeling* (Vol. 2). New York: John Wiley & Sons.

Sigtia, S., & Dixon, S. (2014). Improved music feature learning with deep neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6959–6963). Florence, Tuscany.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26*(4), 471–495.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.

Tan, C. M., Wang, Y. F., & Lee, C. D. (2002). The use of bigrams to enhance text categorization. *Information Processing & Management*, *38*(4), 529–546.

Wan, M. Y. (2017). *關於精細句法結構特徵在自動語體分類中的應用性研究* [The Application of Fine-grained Syntactic Features to Automatic Genre Classification] (PhD thesis). Cityu University of Hong Kong.

Wan, M. Y., & Fang, A. C. (2018). A re-examination of syntactic complexity by investigating the internal structure variations of adverbial clauses across speech and writing. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong.

Wan, M. Y., & Liu, M. C. (2018). Supervised word sense disambiguation with frame-based constructional features: A pilot study of *fán*煩 “to annoy/be annoying/be annoyed”. *International Journal of Knowledge and Language Processing*, *9* (2), 33–46.

Wan, M. Y., Xiang, R., Chersoni, E., Klyueva, K., Ahrens, K., Miao, B., … Huang, C. R. (2019). Sentence boundary detection of financial data with domain knowledge enhancement and cross-lingual training. In *Proceedings of the first workshop on financial technology and natural language processing* (pp. 122–129). Macao, China.

Wang, S., Huang, C. R., Yao, Y., & Chan, W. S. (2019). The effect of morphological structure on semantic transparency ratings. *Language and Linguistics*, *20*(2), 225–255.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Massachusetts, USA: Morgan Kaufmann.

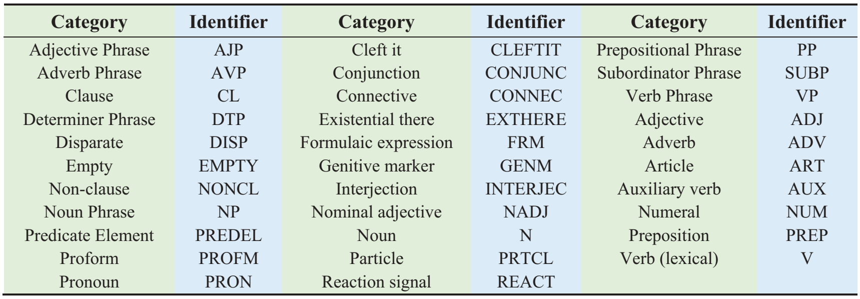
Wolters, M., & Kirsten, M. (1999). Exploring the use of linguistic features in domain and genre classification. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics* (pp. 142–149). Bergen, Norway: Association for Computational Linguistics.

Appendices

Appendix A. The Genre Breakdown of the ICE-GB Corpus.

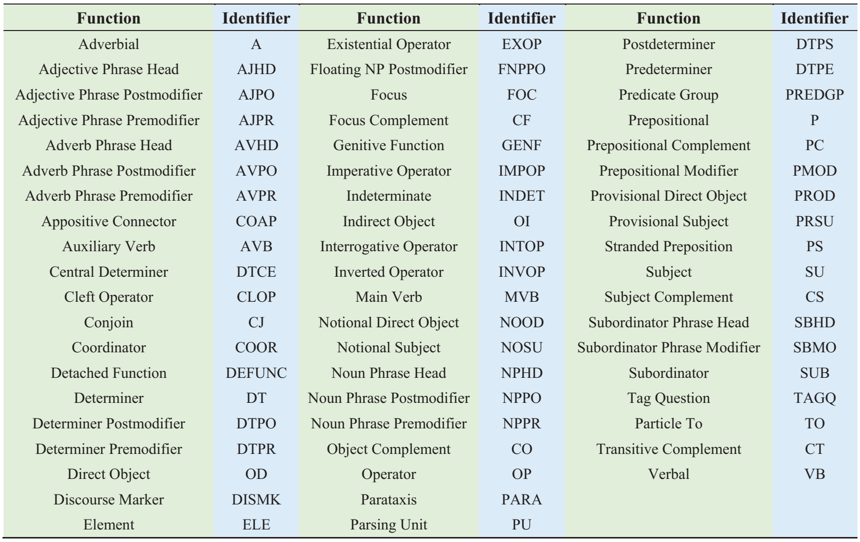
SW2	SW5	SW11	SW32	Text Codes
Spoken (300)	Dialogues (180)	Private (100)	Direct conversations (90)	s1a001-090
			Phone calls (10)	s1a091-100
		Public (80)	Classroom lessons (20)	s1b001-020
			Broadcast discussions (20)	s1b021-040
			Broadcast interviews (10)	s1b041-050
			Parliamentary debates (10)	s1b051-060
			Legal cross-examinations (10)	s1b061-070
			Business transactions (10)	s1b071-080
	Monologues (100)	Unscripted (70)	Spontaneous commentaries (20)	s2a001-020
			Unscripted speeches (30)	s2a021-050
			Demonstrations (10)	s2a051-060
			Legal presentations (10)	s2a061-070
		Scripted (30)	Broadcast talks (20)	s2b021-040
			Non-broadcast speeches (10)	s2b041-050
	Mixed (20)	Mixed (20)	Broadcast news (20)	s2b001-020
Written (200)	Non-printed (50)	Non-professional Writing (20)	Untimed student essays (10)	w1a001-010
			Student examination scripts (10)	w1a011-020
		Correspondence (30)	Social letters (15)	w1b001-015
			Business letters (15)	w1b016-030
	Printed (150)	Informational (100)	Academic humanities (10)	w2a001-010
			Academic social sciences (10)	w2a011-020
			Academic natural sciences (10)	w2a021-030
			Academic technology (10)	w2a031-040
			Non-academic humanities (10)	w2b001-010
			Non-academic social sciences (10)	w2b011-020
			Non-academic natural sciences (10)	w2b021-030
			Non-academic technology (10)	w2b031-040
			Press news reports (20)	w2c001-020
		Instructional (20)	Administrative writing (10)	w2d001-010
			Skills and hobbies (10)	w2d011-020
		Persuasive (10)	Press editorials (10)	w2e001-010
		Creative (20)	Novels/stories/fictions (20)	w2f001-020

Appendix B. The 32 Syntactic Categories in the ICE-GB Corpus.



Category	Identifier	Category	Identifier	Category	Identifier
Adjective Phrase	AJP	Cleft it	CLEFTIT	Prepositional Phrase	PP
Adverb Phrase	AVP	Conjunction	CONJUNC	Subordinator Phrase	SUBP
Clause	CL	Connective	CONNEC	Verb Phrase	VP
Determiner Phrase	DTP	Existential there	EXTHERE	Adjective	ADJ
Disparate	DISP	Formulaic expression	FRM	Adverb	ADV
Empty	EMPTY	Genitive marker	GENM	Article	ART
Non-clause	NONCL	Interjection	INTERJEC	Auxiliary verb	AUX
Noun Phrase	NP	Nominal adjective	NADJ	Numeral	NUM
Predicate Element	PREDEL	Noun	N	Preposition	PREP
Proform	PROFM	Particle	PRTCL	Verb (lexical)	V
Pronoun	PRON	Reaction signal	REACT		

Appendix C. The 58 Syntactic Functions in the ICE-GB Corpus.



Function	Identifier	Function	Identifier	Function	Identifier
Adverbial	A	Existential Operator	EXOP	Postdeterminer	DTPS
Adjective Phrase Head	AJHD	Floating NP Postmodifier	FNPPO	Predeterminer	DTPE
Adjective Phrase Postmodifier	AJPO	Focus	FOC	Predicate Group	PREDGP
Adjective Phrase Premodifier	AJPR	Focus Complement	CF	Prepositional	P
Adverb Phrase Head	AVHD	Genitive Function	GENF	Prepositional Complement	PC
Adverb Phrase Postmodifier	AVPO	Imperative Operator	IMPOP	Prepositional Modifier	PMOD
Adverb Phrase Premodifier	AVPR	Indeterminate	INDET	Provisional Direct Object	PROD
Appositive Connector	COAP	Indirect Object	OI	Provisional Subject	PRSU
Auxiliary Verb	AVB	Interrogative Operator	INTOP	Stranded Preposition	PS
Central Determiner	DTCE	Inverted Operator	INVOP	Subject	SU
Cleft Operator	CLOP	Main Verb	MVB	Subject Complement	CS
Conjoin	CJ	Notional Direct Object	NOOD	Subordinator Phrase Head	SBHD
Coordinator	COOR	Notional Subject	NOSU	Subordinator Phrase Modifier	SBMO
Detached Function	DEFUNC	Noun Phrase Head	NPHD	Subordinator	SUB
Determiner	DT	Noun Phrase Postmodifier	NPPO	Tag Question	TAGQ
Determiner Postmodifier	DTPO	Noun Phrase Premodifier	NPPR	Particle To	TO
Determiner Premodifier	DTPR	Object Complement	CO	Transitive Complement	CT
Direct Object	OD	Operator	OP	Verbal	VB
Discourse Marker	DISMK	Parataxis	PARA		
Element	ELE	Parsing Unit	PU		

Appendix D. Performances (F) of the seven phrasal structures in SW32 (grouped with five macro classes).

NP	PP	DTP	AJP	VP	AVP	SUBP	Genre
71.6	70.6	63.2	60	68	66.3	36.7	Dialogues
15.4	0	0	6.7	8.7	10.5	0	Dialogues
25.5	41.9	12.8	22.2	6.9	11.8	6.7	Dialogues
34.6	27.3	14	10.8	11.4	13	22.2	Dialogues
0	15.4	0	11.1	0	0	0	Dialogues
57.1	0	15.4	9.5	0	42.1	0	Dialogues
30.8	33.3	0	8	13.3	10.5	0	Dialogues
15.4	15.4	0	0	0	16.7	0	Dialogues
76.2	78	34.3	55.8	32.8	73.7	29.8	Monologues
34.7	45.9	6.9	14.8	20.6	19	3.4	Monologues
0	10.5	0	0	13.3	21.1	0	Monologues
52.6	26.7	26.7	15.4	33.3	14.3	0	Monologues
32.4	31.3	10.3	13.6	13.6	12.8	8.7	Monologues
15.4	0	0	11.1	15.4	0	0	Monologues
11.1	12.5	0	11.1	0	0	0	Non-printed
35.3	15.4	12.5	44.4	23.5	0	0	Non-printed
54.5	26.7	37	48.5	7.7	30.8	0	Non-printed
84.8	22.2	37.5	17.1	4.7	24.4	11.5	Printed
28.6	0	12.5	0	0	19	0	Printed
40	14.3	20	7.4	12.5	0	0	Printed
40	22.2	25	21.1	22.2	12.5	0	Printed
21.1	11.8	11.1	9.1	37.5	28.6	0	Printed
11.1	0	0	11.8	0	0	0	Printed
35.3	0	0	10.5	0	0	0	Printed
12.5	0	0	0	11.8	0	0	Printed
0	23.5	46.2	10.5	17.4	16.7	0	Printed
52	44.9	42.3	12.2	11.8	44.4	10.3	Printed
37.5	21.1	50	42.9	10.5	28.6	0	Printed
22.2	12.5	0	13.3	0	0	0	Printed
0	12.5	18.2	9.5	11.8	9.1	0	Printed
52.4	41.9	33.3	43.9	33.3	35.9	0	Printed
36.4	36.4	14.6	28.6	40	31.1	6.7	Mixed