GENRE INDUCTION FROM A LINGUISTIC APPROACH

Angvarrah Lieungnapar
Richard Watson Todd
Wannapa Trakulkasemsuk
King Mongkut's University of Technology Thonburi, Thailand
angvarrahl@gmail.com, irictodd@kmutt.ac.th, wannapa.tra@kmutt.ac.th

First received: 1 September 2016 Final proof received: 11 January 2017

Abstract

In most current work on genre, a set of genre categories needs to be predetermined. However, there are some cases where such predetermined genres cannot be clearly identified. Popular science, for instance, is a broad register carrying several specific purposes within it, suggesting that there are several genres of popular science, but it is unclear what these genres are. This paper introduces a linguistic approach to reveal hidden genres. For 600 written popular science texts from a variety of sources and disciplines, linguistic features were analysed using a range of computer programs and a cluster analysis conducted. The analysis produced four clusters with shared linguistic features, representing text types. The association of these text types with key features, functional relations, dominant sources, and prototypical members of each cluster helps us to induce genres on the basis of communicative purposes, a traditional criterion in identifying genres. Whether the produced text types are equivalent to genres was evaluated with a test set of data. The proposed approach achieves more than 70 % accuracy. The approach appears applicable for identifying genres of popular science and has pedagogical implications.

Keywords: genres; text types; popular science; linguistic features; cluster analysis

Text classification is a major focus in natural language processing (NLP) and computational linguistics. Text classification can be a confusing discipline since there are three terms commonly used to describe functional classifications of texts (genre, register, and text type) and these terms are used with different meanings by different authors. To start, then, we need to be clear about how these perspectives on texts are different and what we mean by these terms.

Genre is probably the most familiar of these three terms and in this paper we will follow Swales (e.g. 1990) in using genre to refer to a fairly specific set of texts which share a common communicative purpose. Most Swalesian genre analyses aim to characterize texts according to their conventional structure, such as the typical discourse moves and steps which are structural patterns representing a stretch of text defined by communicative functions (Kanoksilapatham, 2012). In investigating genres in this sense, however, each genre category firstly needs to be identified on the basis of communicative purposes. This means texts having the same communicative purposes are usually categorized into the same genre category, such as research article abstracts (e.g. Martin, 2003).

Register is perhaps more problematic in that different subdisciplines of applied linguistics identify registers at very different levels of specificity. In some work, registers are akin to specific occupational genres (e.g. Wardhaugh, 2006); in other work, register refers to a very general variation in style, such as whether the language is casual or formal (e.g. Bax, 2011). In this paper, we use this term to refer to a variety associated with a particular situation of use (Biber & Conrad, 2009). This means we take register to be more general than genre. A further difference between genre and register is that register analyses usually focus on lexical and grammatical features of texts, rather than the discourse-level features of genre analyses.

Text type is also used with two conflicting meanings. On the one hand, text type can refer to texts with a certain generic discourse structure, such as problem-solution (e.g. Paltridge, 1996). On the other, text type refers to groupings of texts which share linguistic features irrespective of their social contexts of use (Biber, 1989). It is this second meaning that we will use in this paper.

To summarize the meanings that we will use for these three terms, Table 1 presents our interpretations of the specificity, identifying characteristics and focuses of analysis of genre, register and text type.

In much of the work in text classification, texts are classified based on their topics, but text classification can also attempt to separate texts into different genres (Stede, 2012). Before classifying texts into the genres, genres should be initially identified on the basis of communicative purposes, the identifying characteristic of genres. Automated genre classification is a common goal in NLP. A key problem in automated genre classification is

doi: dx.doi.org/10.17509/ijal.v6i2.4917

that, in some cases, genre identification cannot be based on communicative purposes. This is because communicative purposes are largely intangible and intuitive and thus cannot be straightforwardly automated. For this reason, most existing NLP genre classification applications use a set of predetermined genres instead. Still, there are contexts where such predetermined genres cannot be clearly identified. Such a situation may be because the texts manifest hybridity (Bax, 2011) or it may simply be that the purposes are difficult to identify. That is, we may have a collection of texts from the same register which should be separable into different genres, but it is unclear what these genres should be.

Table 1. Meanings of genre, register, and text type

	Level of specificity	**Identifying characteristics**	**Focus of analysis**
Genre	Specific	Communicative purposes	Discourse structure
Register	Broad	Variation in style	Lexicogrammar
Text type	-	Linguistic features	Linguistic features

In this paper, we intend to address such a context where genres cannot be initially identified on the basis of communicative purposes nor a set of predetermined genres. In such a context, we need to take a different approach, and one possible alternative is a ‘text-first or linguistic approach’ (Askehave & Swales 2001, p. 207). In effect, such an approach means that we will start by conducting a text type analysis to categorize the texts into sets which share linguistic features. We will then use previous work, especially within register analysis, to identify the broad functions associated with the shared linguistic features. Finally, we will attempt to induce communicative purposes from these broad functions to characterize our sets of texts as genres. In other words, we aim to identify communicative purposes and genres post hoc. To check the validity of these induced genres, we will compare the automated classification against the human-generated classification.

The goal of this study, then, is to see if it is possible to classify a set of texts into genres and to identify the communicative purposes of these genres by conducting a linguistic feature-oriented text type analysis. More specifically, we intend to answer the following research questions: (1) How do the texts investigated cluster together based on linguistic features? (2) What are the linguistic characteristics of each cluster of texts? (3) How do the clusters manifest communicative purposes and represent genres?, and (4) How valid are the identified genres?

METHOD

Data

Popular science has been defined as the reporting of scientific facts that are written for audiences without a professional background in science (Hyland, 2009). It is fairly clear that the overall communicative purpose of popular science is to report scientific information to a general audience. Since its purpose is fairly clear, it is relatively straightforward to identify text samples which belong to popular science writing. Calsamiglia (2003), for instance, identifies scientific news reports in newspapers, popular scientific magazines such as *Scientific American* and *New Scientist*, and some television documentaries as popular science. The examples given by Hyland (2010) are popular science books written by scientists for an elite educated audience and specialized science sections in the press.

These exemplars seem to suggest that this type of writing is varied and manifests itself through a wider range of genres. This suggests further that popular science can be considered as a broad register with a variation in style, purposes and topics. In this paper, therefore, we will refer to popular science writing as an exemplar of a broad register comprising several more specific genres. The reason why previous work has focused on giving examples, rather than identifying genres, is that it is unknown what the genre categories should be. Since it is unclear in the literature what the genres of popular science include, popular science is selected as the data for investigation in the present study.

Data collection

The main goal of this study is to use linguistic features to induce genres when genre identification cannot start with communicative purposes and predetermined genres cannot be clearly identified. To ensure that our analysis is likely to cover several genres, we need a range of texts. However, for practical purposes, the popular science texts collected for this study were limited to fairly short written texts that are generally comparable in length and fall within the same period of publication. Even restricting the data collection to short written texts, there are far too many possible texts to be manageable. We therefore used three further criteria to select texts in such a way that we believe our final data set will provide a wide coverage of the range of the short written popular science texts that exist.

Given the notion that popularization is a matter of degree and operates along a continuum with practitioners positioned somewhere in the middle between researchers and the educated public (Giannoni, 2008), the first criterion is the concept of upstream/ downstream. That is, texts selected for the analysis must represent a continuum ranging from ‘upstream’ texts close to the site of production of the science to ‘downstream’ texts addressed to wider audiences especially the non-scientist or non-specialist (Hilgartner, 1990, p. 528). Since genres vary across sources and disciplines (Nesi & Gardner, 2012), texts were selected from a wide range of sources and disciplines to ensure balance and representativeness. For practicality, however, this study focuses on six sources, as shown in Table 2, and five disciplines (biology, earth, medicine, space, and technology). The six sources are likely to range from upstream to downstream although it is still unclear if the sources and disciplines fit with genres or not. The data in this study then is a collection of 600 texts comprising six sources (one hundred texts per source) and five disciplines (20 texts from each discipline in each source).

Table 2. Data selection from six sources ranging from upstream to downstream

	Sources	**Description**
Upstream	1. *Science* abstracts	1. Scientific texts not necessarily addressed to specialist scientists
↕	2. *Nature* research highlights	2. Scientific articles in journals for general scientists
	3. *Wikipedia* featured articles	3. Science encyclopedia
	4. *Science* news of the week	4. Science news in scientific journal
	5. *New Scientist* news upfront	5. Science news in popular science magazine
Downstream	6. *Wikinews* news stories	6. Science news in news reports

Data analysis

To achieve the purpose of the study, the data analysis proceeded in two stages. First, a linguistic analysis of each text was conducted. To ensure that as wide a range as possible of the potential linguistic features that are capable of manifesting purposes was selected, features ranging from discourse features to specific linguistic features were selected under two criteria: a capability of distinguishing texts and relevance to popular science texts. According to Myers (2003), some features such as metaphors and hedging are the main characteristics of popular science. However, these features are hidden, meaning that investigating them is subjective and time-consuming. This study, therefore, focuses on overt features that are easier to automate. A total of 63 linguistic features were identified and counted in each text. The features include some discourse features (e.g. genre moves), text features (e.g. readability), grammatical features (e.g. proportions of nouns) and specific linguistic features (e.g. phrasal verbs).

Some of these features are likely to overlap to a large extent with other features. Therefore, to avoid a duplicate influence of potential linguistic features on the cluster analysis, the correlated features were grouped together by using cluster analysis (see e.g. Leonard & Droege, 2008), although factor analysis is more conventional. An example of a set of features with high inter-correlations is verbs which include verb density, infinitive density and verb phrase density. Of those, verb density has the highest statistic F value and so is selected to be representative of this set. From this process, 19 features were selected for the analysis as representative of all features without overlap. Given the fact that these features are on different scales, each feature was normalized by converting the raw frequency counts to ratio scores, to proportions of total number of words, or to a relative frequency per 1000 words, as determined by the analysis software that was used. The operationalizations of the analysis and functions associated with all features are shown in Table 3.

Next, to classify the popular science texts on the basis of linguistic features into groups representing text types, cluster analysis was applied. Cluster analysis is an exploratory data analysis tool which aims at automatically sorting a substantial number of data objects (e.g. texts) into a much smaller number of coherent groups (called clusters) on the basis of similar variables (e.g. linguistic features). The analysis is not the same as the better known multi-dimensional analysis. The latter investigates register variation to find the quantitative distribution of linguistic features across text varieties (in effect, text varieties are the independent variable and linguistic features the dependent variable) whereas the ultimate goal of cluster analysis is to classify groups of texts (in effect, linguistic features are the independent variable and the text clusters the dependent variables). To identify groups of texts that are similar to each other and dissimilar to the texts belonging to other clusters in terms of linguistic features, IBM SPSS version 20.0 was used. The steps in doing a cluster analysis are as follows. Firstly, the analysis starts by selecting a clustering method. Due to its suitability for clustering relatively large data sets (600 cases) and since the appropriate number of clusters is unknown, the K-means approach was chosen. Based on the set of cluster centers, this technique assigns all cases observed into K number of clusters having minimal variability within the cluster and maximum variability between clusters. The next step is that outliers need to be eliminated as K-means clustering is sensitive to outliers. Outliers are texts having more maximum and minimum values compared to other texts. They need to be eliminated because outliers will be selected as initial cluster centers and thus they form clusters with small numbers of cases.

Table 3. Linguistic features selected for cluster analysis

No.	Linguistic features	Operationalizations	Tools	Reasons for inclusion
1.	Average sentence length	Average number of words per sentence within the text	Microsoft Word	Longer sentences are commonly used to mark complex and elaborated structure.
2.	Average paragraph length	Average number of sentences per paragraph within the text	Microsoft Word	Longer paragraphs are frequently used to mark high information density.
3.	Discipline-specific word density	Number of specialized vocabulary items in content-specific areas as a proportion of total number of words	RANGE	Discipline-specific words are frequently used to express referential information in specific subject areas.
4.	Phrasal verb density	Number of phrasal verbs as a proportion of total number of verbs	CLAWS, AntConc	Since phrasal verbs manifest a degree of informality and textual spokenness, a high frequency of this feature suggests a narrative purpose.
5.	Compound noun density	Number of open compound nouns as a proportion of total number of nouns	CLAWS, AntConc	A high frequency of compound nouns indicates greater density of information.
6.	Modal verb density	Number of modal verbs as a proportion of total number of words	CLAWS, AntConc	Modality is used to mark explicit persuasion.
7.	Verb density	Number of verbs as a proportion of total number of words	CLAWS, AntConc	Verbs indicate a verbal style that can be considered interactive or involved and are used for the overt expression of attitudes, thoughts, and emotions.
8.	Adjective density	Number of adjectives as a proportion of total number of words	CLAWS, AntConc	A high frequency of adjectives can be associated with a high informative focus and careful integration of information in a text.
9.	Adverb density	Number of adverbs as a proportion of total number of words	CLAWS, AntConc	Adverbs are used more frequently to indicate situation-dependant reference for narrating a story.
10.	Lexical repetition	Yule's characteristic K (the variance of the mean number of occurrences per word)	SCP	The larger Yule's K, the more the lexical repetition. Greater use of repetition results from the purposes of explicitly marking cohesion in a text and informative focus.
11.	Coordinating conjunction density	Number of coordinating conjunctions as a proportion of total number of sentences	CLAWS, AntConc	Coordinating conjunctions are commonly used to show formality in referentially explicit discourse.
12.	Content word density	Number of content words as a proportion of total number of words	CLAWS, AntConc	Content words mark precise lexical choice resulting in a presentation of informative content.
13.	Evaluation move density	Numbers of evaluation moves as a proportion of total number of sentences	AntMover	Evaluative language is normally used to express emotions and attitudes.
14.	Vocabulary diversity	Sums of probabilities of encountering each word type in 35-50 tokens	Coh-Metrix	A high diversity of vocabulary results from the use of many different vocabulary items. Narrative texts often have high vocabulary diversity.
15.	Logical connective density	Number of logical connectives per 1000 words	Coh-Metrix	A high frequency of logical connectives indicates an informative relation in a text.
16.	Prepositional phrase density	Number of prepositional phrases per 1000 words	Coh-Metrix	Prepositional phrases indicate a greater density of information.
17.	Negation density	Number of negation markers per 1000 words	Coh-Metrix	Negation is preferred in literary narrative.
18.	Pronoun density	Number of pronouns per 1000 words	Coh-Metrix	Pronouns refer directly to the addressor and addressee and thus are used frequently in highly interactive discourse.
19.	Flesch Reading Ease	Flesh Reading Ease formula	Coh-Metrix	Higher Flesch reading scores are easier to read.

After screening, 27 outlier texts were deleted, meaning that there are only 573 texts used for cluster analysis. Before conducting K-means analysis, the optimal number of clusters needs to be determined. A possible range of appropriate cluster solutions can be identified by applying a hierarchical technique, another approach which provides a visual representation of a hierarchical cluster structure based on a dendrogram and the agglomeration schedule output. Based on this technique, the tenability of a range from two to five cluster solutions was specified. Finally, K-means cluster analysis was conducted based on the standardized scores of 19 linguistic features in 573 texts for all possible ranges of cluster solutions. To reveal an optimal number of clusters, the significant ANOVAs produced by cluster analysis were considered. Based on the results produced from the cluster analysis and researchers' interpretation, four clusters seem to be optimal for this dataset.

FINDINGS

Clusters of Texts

The numbers of texts in each cluster and distances between the clusters can be seen in Table 4. Although Cluster 2 is a little larger than the other clusters, this does not make the clustering imbalanced as can be seen that the numbers of texts in each cluster are roughly equal. According to the distances between each cluster (as illustrated in the extended tree diagram in Figure 1), clusters differ in the closeness of their relationships. For example, Cluster 4 is the furthest from Cluster 1, suggesting that Cluster 4 is the most different from Cluster 1.

Linguistic characteristics of the four text types

Since cluster analysis was performed on the basis of linguistic variables, this approach also allows the identification of the distinctive linguistic characteristics for each cluster representing a text type. On the basis of the z-scores of the 19 linguistic features, Table 5 presents the linguistic characteristics associated with the four text types. Z-scores of linguistic features that are greater than 0.35 (representing standard deviation greater than the mean) were designated as cut-off points to represent noteworthy departures from central tendency and to shed light on the key linguistic features of each text type. These key features are frequently associated with a particular text type (except for Flesch Reading Ease that is a key linguistic characteristic of both Text type 1 and Text type 2), suggesting that the text types are distinctive. The key linguistic features (values higher than 0.35) are abstracted in Table 6. We can see, for example, that Text type 1 is comprised of texts with a high use of eight linguistic features while there are fewer identifying linguistic features in Text type 2.

Functional relations among text types

The analyses of linguistic features and cluster analysis only provide the foundations for the initial identification of clusters representing text types (a group of texts having shared linguistic features). To interpret genres, we need to associate linguistic features with functions which, in turn, can be linked to purposes. Under the assumption that one particular linguistic feature can be associated with more than one function, the interpretation is based on the previous literature discussing the association of linguistic features with functional relations, especially the work of Biber (1988). For instance, for Text type 1, many features (e.g. pronoun density, verb density, and logical connective density) are not only highly associated with the interpersonal function, but also associated with the narrative, persuasive and informative functions whereas many features in Text type 4 are highly associated with the impersonal function and somewhat associated with the informative and elaborated functions. The associations between key features and functional relations are shown in Table 6.

Table 4. Number of texts and distances between the clusters

Cluster	No .of texts	Distances			
		Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	133	0	2.59	4.32	4.31
2	187	2.59	0	2.93	3.70
3	121	4.32	2.93	0	3.09
4	132	4.31	3.70	3.09	0

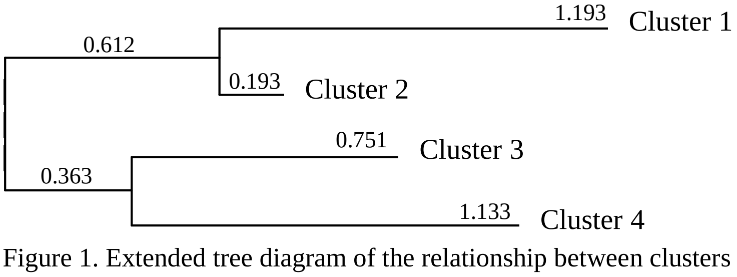

Figure 1. Extended tree diagram of the relationship between clusters

Table 5. Z-scores of linguistic characteristics of each text type

No.	Linguistic features	Text type 1	Text type 2	Text type 3	Text type 4
1.	Average sentence length	-0.49	0.07	-0.28	**0.68**
2.	Average paragraph length	-0.43	-0.62	**1.05**	0.34
3.	Discipline-specific word density	-0.21	-0.32	0.08	**0.59**
4.	Phrasal verb density	**0.47**	0.17	-0.25	-0.50
5.	Compound noun density	-0.11	-0.31	0.20	**0.36**
6.	Modal verb density	0.31	**0.39**	-0.57	-0.40
7.	Verb density	**0.67**	0.29	-0.86	-0.29
8.	Adjective density	-0.42	-0.62	0.24	**1.08**
9.	Adverb density	**0.45**	-0.19	-0.27	0.08
10.	Lexical repetition	-0.85	0.16	**0.97**	-0.29
11.	Coordinating conjunction density	-0.36	-0.29	-0.08	**0.83**
12.	Content word density	-0.24	-0.50	-0.04	**0.99**
13.	Evaluation move density	0.31	-0.19	**0.39**	-0.40
14.	Vocabulary diversity	**1.02**	-0.20	-0.82	0.06
15.	Logical connective density	**0.49**	-0.10	-0.26	-0.10
16.	Prepositional phrase density	-0.82	0.17	**0.53**	0.09
17.	Negation density	**0.59**	-0.23	-0.26	-0.07
18.	Pronoun density	**0.85**	-0.03	-0.34	-0.47
19.	Flesch Reading Ease	**0.55**	**0.45**	0.00	-1.20

Table 6. Functional relations associated with key features in text types

Text Type	Key features	Functions					
		Interpersonal	Narrative	Persuasive	Informative	Elaborated	Impersonal
1	phrasal verb density, verb density, adverb density, vocabulary diversity, logical connective density, negation density, pronoun density, Flesch reading ease	✓✓	✓	✓	✓		
2	modal verb density, Flesch reading ease	✓		✓✓			
3	average paragraph length, lexical repetition, evaluation move density, prepositional phrase density				✓✓		
4	average sentence length, discipline-specific word density, compound noun density, adjective density, coordinating conjunction density, content word density				✓	✓	✓✓

The associations of text types and functions manifest non-linguistic characteristics of each text type. In Table 6, we can see that most text types are associated with the informative function. This is the primary function of popular science texts and is manifested in most of the text types identified. This table also shows that the relational functions vary among texts types, meaning that these distinguishing functions may link to different communicative purposes.

In addition, we can see that these functional relations are likely to be associated with the upstream-downstream continuum. That is, key features (groups of linguistic features having z-scores above average of each text type) in Text types 1 and 2 share interpersonal and persuasive functions. These functions, which are likely to be used to show interpersonal interaction, seem more associated with downstream texts addressed to general readers. On the other hand, most key features in Text types 3 and 4 are associated with the informative, elaborated and impersonal functions. These functions, which illustrate a careful integration of information in a text with an impersonal style, seem more associated with upstream texts that are close to scientific articles. This allows us to say that the four text types of popular science texts can be represented on a continuum ranging from downstream to upstream.

Source and discipline relations among text types

To induce genres, text types found need to be linked to other external criteria, namely source and discipline. Tables 7 and 8 summarize the relationships between the text types, on the one hand, and disciplines and sources, on the other. We found that, in every discipline, texts are fairly equally distributed across text types, suggesting that the texts are not clustered in terms of discipline; in other words, there is no match between discipline and text types. On the other hand, there is a loose relation between sources and clusters. Although there is only one case of a zero match between a text type and a source (*Science* abstract source and Text type 2), for Text types 3 and 4, it is fairly clear that there is a relationship between sources and text types since these two text types are dominated by one particular source. That is, most of Text type 3 is *Wikipedia* articles (56.20%) while most of Text type 4 is *Science* abstracts (56.82%). Similarly, most *Wikipedia* articles (73.12%) are in Text type 3 whereas most *Science* abstracts (79.79%) fall into Text type 4. The compositions of sources in Text types 1 and 2, however, suggest a looser relation as each text type is comprised of texts from more than one source. The texts from these sources can be divided according to whether they are highly interpersonal (Text type 1) or highly persuasive (Text type 2). This relationship suggests that source as a non-linguistic criterion is likely to better help identify genres than discipline.

Table 7. The relationship between clusters and disciplines

Clusters	Biology	Earth	Medicine	Space	Technology	Total
1	27	19	29	26	32	133
	20.30	14.29	21.80	19.55	24.06	100%
2	38	43	35	38	33	187
	20.32	22.99	18.72	20.32	17.65	100%
3	27	19	29	26	32	133
	20.30	14.29	21.80	19.55	24.06	100%
4	23	25	31	23	30	132
	17.42	18.94	23.48	17.42	22.73	100%

Table 8. The relationship between clusters and sources

Clusters	***Science*** abstracts	***Nature*** research highlights	***Wikipedia*** featured articles	***Science*** news of the week	***New Scientist*** news upfront	***Wikinews*** news stories	Total
1	2.50	19.50	1.50	43.50	47.50	21.50	133.0
	1.50	14.29	0.75	32.33	35.34	15.79	100%
2	0.50	50.50	1.50	37.50	49.50	50.50	187.0
	0.00	26.74	0.53	19.79	26.20	26.74	100%
3	17.50	8.50	68.50	9.50	2.50	17.50	121.0
	14.05	6.61	56.20	7.44	1.65	14.05	100%
4	75.50	17.50	23.50	9.50	1.50	7.50	132.0
	56.82	12.88	17.42	6.82	0.76	5.30	100%

Communicative purposes and genre induction

Since communicative purpose is the main identifying characteristic of genres, we need to induce communicative purposes before inducing genres. Communicative purposes, however, cannot be directly derived from linguistic features. Consequently, we need to consider the overall findings (see Table 9), including the results of the prototypical members from cluster analysis. The prototypical members of each cluster representing text type are the texts that have the smallest distances from the cluster center, suggesting that these prototypical texts can be considered as the best examples representing each text type and can be used to induce communicative purposes, along with other results.

Although the informative function appears to be the characteristic in almost all text types, when considering all of the results, including linguistic and non-linguistic basis, we can say that popular science can be sub-divided into four genres that have distinctive communicative purposes. The communicative purposes of all clusters were induced and are presented in Table 10. The interpretation and induction of communicative purposes and genres are briefly illustrated as follows.

In Text type 1, for instance, the prototypical texts are *Science* news stories. These texts have a high vocabulary diversity, a high logical connective density, and a greater frequency of time adverbials (e.g. ‘100 years ago’, ‘now’, and ‘in November’) and narrative verbs (e.g. ‘start’, ‘have occurred’, and ‘had dropped’). These features can be interpreted as markers of narratives used to recount a story, express an opinion, or give information chronologically. The high frequencies of personal pronouns and verbs involve readers and make the narration interactive. Also, the reading ease score is high, suggesting the purpose of making the science easy for a general audience. These can reflect the underlying communicative purpose of Text type 1 as to narrate a scientific story to involve and entertain readers. Considering this communicative purpose, the interpretative label ‘Scientific narratives’ is proposed for the genre.

In Text type 2, on the other hand, the prototypical texts are news from *Science* news, *New*

Scientist and *Wikinews*. Although they are short texts, they use a high frequency of modal verbs in order to highlight certain future events, and to persuasively justify and evaluate current research on scientific discoveries related to public concerns (e.g. 'has discovered', 'the discovery of', and 'the results indicate that'). The focus on the importance and value of findings allows us to interpret the underlying communicative purpose of this text type as to discuss current or recent findings in science. Reflecting this communicative purpose, the interpretative label of the genre is 'Persuasive reports of scientific news'.

The interpretation of communicative purposes of Text types 3 and 4 are fairly straightforward. This is because they have clear prototypical texts: *Wikipedia* for Text type 3 and *Science* abstracts for Text type 4. Since the texts in Text type 3 have a high prepositional phrase density, lexical repetition, and average paragraph length, we can interpret the communicative purpose of Text type 3 as to describe and explain scientific information, which suggests a label of genres as 'Scientific descriptions'. Although the texts in Text type 4 also use some linguistic features to mark an informative focus (e.g. compound noun density), they also use some features to indicate the elaborated function to express condensed ideas with relevant details (e.g. coordinating conjunction density). While this function represents referentially explicit discourse marked by the explicit, elaborated identification of referents in a text, the impersonal style is marked by features such as discipline-specific word density. Based on this functional association, the communicative purpose of texts in Text type 4 is to summarize technical information. This allows us to identify this genre as 'Technical summaries'.

Table 9. Summary of the findings

Cluster	Key cluster features	Functional relations	Upstream/ downstream	Dominant sources	Prototypical Members
1	phrasal verb, verb, adverb, vocabulary diversity, logical connective, negation, pronoun, Flesch reading ease	interpersonal narrative persuasive informative	downstream	*New Scientist,* *Science* news	*Science* news
2	modal verb, Flesch reading ease	interpersonal persuasive	downstream	*Nature,* *New Scientist,* *Wikinews*	*Science* news *New Scientist* *Wikinews*
3	average paragraph length, lexical repetition, evaluation move, prepositional phrase	Informative	upstream	*Wikipedia*	*Wikipedia*
4	average sentence length, discipline-specific word, compound noun , adjective, coordinating conjunction, content word	informative elaborated impersonal	upstream	*Science* abstracts	*Science* abstracts

Table 10. Four genres of popular science and their communicative purposes

Text type	Communicative purposes	Genres of popular science
1	To narrate a scientific story to involve and entertain readers	Scientific narratives
2	To discuss current or recent findings in science	Persuasive reports of scientific news
3	To describe and explain scientific information	Scientific descriptions
4	To summarize technical information	Technical summaries

Evaluation of genre induction

To evaluate the validity of the linguistic approach with cluster analysis for genre induction, we need to compare the results of the automated analysis against human expert judgments, the 'gold standard' for validating automated approaches (Stokes, 2004, p. 28). To this end, a new set of 30 texts representing the same disciplines and sources was collected. The automated analysis using 19 linguistic features was conducted to identify the text type of each text. The 30 texts were also given to two experts who were asked to classify each text into one of the four text types. The results of the two classifications were compared using common Information Retrieval performance features: accuracy, precision, and recall. Accuracy approximates how effective the approach is by showing the probability of the true value of the classification. Recall approximates the probability of a positive classification being true while precision estimates the predictive value of a classification. The accuracy rate is quite high (77.5%), with precision and recall both being over 55%. While these results are not very high, they are almost exactly the same as the comparable figures when the two experts' classifications are compared against each other, suggesting that it may not be possible to obtain higher figures.

DISCUSSION

There are several key findings of this study. First, it was shown that it is possible to create fairly evenly balanced clusters, or text types, on the basis of linguistic approach. Second, our results indicated that these text types highlight different linguistic features and these different key features can be linked to different functional relations. The varied functional relations suggest downstream and upstream differences. Third, whereas disciplines are evenly distributed across clusters showing little association between text types and disciplines, sources are more clearly related to the text types. Fourth, it is possible to identify prototypical members of each cluster. Finally, it is possible to induce communicative purposes for text types. Overall, the findings indicate that it is possible to apply the linguistic approach to induce genres of popular science.

A key test of a new methodology is whether it provides useful insights into the problem addressed. One problem of genre classification from the traditional perspective is that genres might not be clearly separated from each other since their characteristics cannot be clearly distinguished (Bax, 2011). It seems unlikely that, without conducting this study, popular science articles in magazines and journals would be seen as falling into the two categories of interpersonal narrative (Text type 1) and persuasive (Text type 2) reports, and thus we would argue that this research has provided useful insights into the genre classification of popular science.

Since communicative purposes are recognized by the expert members in the field (Swales, 1990), the present study attempts to validate the proposed approach with experts. Even though the accuracy is not extremely high, this appears to be because, in the case of popular science, even expert informants have difficulties in classifying texts into the genres. The relatively modest level of accuracy, then, can be viewed as sufficient to make the analysis worthwhile and highlights the need for approaches in those cases where there are no clear intuitively identifiable genres based on communicative purposes.

From the perspective of automation, in this paper we used a combination of several different existing programs in the analysis with choosing the number of clusters and inducing communicative purposes conducted manually. Although the machine-generated solutions would increase efficiency and productivity in genre identification, the approach is still complex. It would be relatively straightforward to integrate all of the automated stages into a single program for automated identification of genres that could be used for text classification without the need for predetermining categories for texts to be classified into. If a fully automated analysis were constructed, it would allow registers other than popular science to be investigated fairly easily which may provide interesting insights into the genre organization of these registers. However, even if a more practical fully automated approach were created, the generalizability of using linguistic features to identify genres is unclear; further work in other fields within this framework is still needed.

The induction of genre categories of popular science from the alternative approach is noteworthy for pedagogical implications in teaching genres. First, the approach provides linguistic characteristics of each genre that are functional and pervasive (frequent and common across texts) whereas features of genres from a traditional approach are normally conventional and might occur only one time in a text (Biber & Conrad, 2009). The linguistic characteristics associated with particular genres can provide pedagogical objectives in teaching genres enabling learners to differentiate the linguistic characteristics of different genres. For example, learners could learn to use more frequent verbs, adverbs, phrasal verbs, and pronouns to differentiate their writing of scientific narratives from persuasive reports of scientific news.

Second, most previous teaching which focuses on particular genres has used the Swalesian approach to teach moves and steps (Nguyen & Pramoolsook, 2015). While such an approach is useful, the linguistic approach used in this study allows the teachers and material developers to also identify specific linguistic features that could be taught. For instance, to teach the specific genre of technical summaries, this paper suggests that the teachers should focus on the following linguistic features: discipline-specific words, compound nouns, adjectives, and coordinating conjunctions. This approach allows lexical and grammatical objectives to be assigned to genre specific courses. Also, to teach a course focusing on a particular genre, teachers and material developers need to provide students with a range of texts as teaching materials from that particular genre. However, it is unclear on what basis texts should be chosen as models for teaching. Applying a cluster analysis helps us identify a prototypical text in each cluster to serve as a representative of the particular genre. The extent to which a prototypical text in a cluster serves as an appropriate pedagogical model is worthy of further investigation.

Genre-specific approaches have become more common in teaching English in recent years, including in Southeast Asia. For example, in Malaysia, genre-specific courses related to English for Science and Technology have been introduced as additional subjects alongside the existing English language courses (Chan & Tan, 2006). It is also worth noting that there has been a recent increase in interest in teaching popular science in Southeast Asia, such as a dedicated course on popular science now offered in Singapore (National University of Singapore, not dated). If genre learning should be based on explicit awareness of language (Hyland, 2003), then the findings of this study have applications in Southeast Asian education both for genre-specific teaching and for teaching popular science.

CONCLUSION

Genre analysis usually uses non-linguistic criteria as a basis for identification and classification. Genre categories are typically identified and classified on the basis of communicative purposes. However, in some cases, such as popular science writing, genre categories cannot be clearly predetermined on this basis. The goal of this study is to induce unknown genres of popular science writing on the basis of linguistic features as an alternative approach. The approach was completed by distributional analysis of a wide range of linguistic features, the use of various computer programs to automatically identify linguistic features in texts, and the use of cluster analysis to identify clusters of texts with typical linguistic features and prototypical examples of each text type. Each text type manifests a distinctive set of linguistic features that are associated with a unique set of functional relations. Based on these associations, the linguistic approach with cluster analysis can predict the genre categories within popular science writing. This suggests that there is a direct relationship between linguistic features and genres. Although it is unclear whether this alternative approach has generalisability to other contexts, these findings have potential implications for other research into genre identification.

ACKNOWLEDGEMENT

This work was supported by a doctoral dissertation grant program, Higher Education Research Promotion, from the Office of the Higher Education Commission, Thailand.

REFERENCES

Askehave, I. and Swales, J. M. (2001). Genre identification and communicative purpose: A problem and a possible solution. *Applied Linguistics, 22*(2), pp. 195-212. doi:1093/applin/22.2.195.

Bax, S. (2011). *Discourse and genre: Analysing language in context*. Hamshire: Palgrave MacMillan.

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

Biber, D. (1989). A typology of English texts. *Linguistics, 27*(1), pp. 3-44.

Biber, D. and Conrad, S. (2009). *Register, genre, and style.* New York: Cambridge University Press.

Calsamiglia, H. (2003). Popularization discourse. *Discourse Studies, 5*(2), pp. 139-146. doi:10.1177/1461445603005002307

Chan, S. H. & Tan, H. (2006). English for mathematics and science: Current Malaysian language-in-education policies and practices. *Language and Education, 20*(4), pp. 306-321.

Giannoni, D. S. (2008). Medical writing at the periphery: The case of Italian journal editorials. *Journal of English for Academic Purpose, 7*(2), pp. 97-107. doi:10.1016/j.jeap.2008.03.003

Hilgartner, S. (1990). The dominant view of popularization: Conceptual problems, political uses. *Social Studies of Science, 20*(3), pp. 519-539. doi:10.1177/030631290020003006

Hyland, K. (2003). *Second language writing*. Cambridge: Cambridge University Press.

Hyland, K. (2009). *Academic discourse: English in a global context.* New York: Bloomsbury.

Hyland, K. (2010). Constructing proximity: Relating to readers in popular and professional science. *Journal of English for Academic Purposes, 9*(2), pp. 116-127. doi:10.1016/j.jeap.2010.02.003

Kanoksilapatham, B. (2012). Facilitating scholarly publication: Genre characteristics of English research article introductions and methods. 3L: *The Southeast Asian Journal of English language Studies, 18*(4), pp. 5-19.

Leonard, S. T. & Droege, M. (2008). The uses and benefits of cluster analysis in pharmacy research. *Research in Social and Administrative Pharmacy, 4*(1), pp. 1-11.

Martin, P. M. (2003). A genre analysis of English and Spanish research paper abstracts in experimental social sciences. *English for Specific Purposes, 22*(1), pp. 25-43. doi:10.1016/s0889-4906(01)00033-3

Myers, G. (2003). Discourse studies of scientific popularization: Questioning the boundaries. *Discourse Studies, 5*(2), pp. 265-279. doi:10.1177/1461445603005002006

Nesi, H. & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education.* New York: Cambridge University Press.

Nguyen, L. T. T., & Pramoolsook, I. (2015). Move analysis of results-discussion chapters in TESOL master's theses written by Vietnamese students. *3L: Language, Linguistics, Literature,* 21(2), pp. 1-15.

Paltridge, B. (1996). Genre, text type, and the language learning classroom. *ELT Journal,* 50(3), pp. 237-243.

Stede, M. (2012). *Discourse processing*. San Rafael, CA: Morgan & Claypool Publishers.

Stokes, N. (2004). *Applications of lexical cohesion analysis in the topic detection and tracking* *domain*. Doctoral dissertation, National University of Ireland, Dublin.

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

Wardhaugh, R. (2006). *An introduction to sociolinguistics*. Oxford: Blackwell.