Glottometrics 48, 2020, 52-66

Identification of English Styles on the Basis of Parts of Speech: A Case of Principal Component Analysis and Factor Analysis

Anastasia Gnatciuc[1]
Hanna Gnatchuk[2]

Abstract: The present study is concerned with the identification of English styles in terms of D. Biber's proposed dimensions (1988). In this study this identification is relied upon the data about the parts of speech. Moreover, we are interested in detecting the relationships between the styles (belles-lettres, official, news scientific and miscellaneous styles), on the one hand, and their subsections (editorial, romance, adventure, learned, reviews, etc.) in the Brown Corpus in terms of three dimensions: "involved (interactive) versus informational production", "description versus reporting" and "narrative versus non-narrative". The Brown Corpus texts were accessed in Python 3 (environment Anaconda) by means of the tagged algorithm. The results were statistically processed in R-Studio Program.

Key words*: English genres/styles, parts of speech, stylistics, corpus linguistics, computational linguistics, statistics, dimensions.*

1. Introduction: preliminary remarks

In general, the distinction between texts has been made by researchers considering different situations and functions of a text: interactive/non-interactive, literary/colloquial, formal/informal, etc. Douglas Biber (1988) in his book "Variation across speech and writing" has made a significant contribution to text linguistics. In particular, he made an attempt to determine the types of texts by means of empirical and statistical methods considering the normalized frequencies of a set of linguistic features (i.e. parts of speech, etc). The researcher referred text types to different genres (registers) in view of their interrelations.

The researcher is of the opinion that the functions of a text can be referred as dimensions "because they define continuums of variation" (1988:9). He declares that texts can be identified as informal or formal, but it is more effective and accurate to characterize it according to the degree of its formality (i.e. more or less formal). In such a way, "formal/informal can be considered a continuous dimension of variation" (1988:9). Considering the dimensions from a linguistic point of view, the researcher states that "each dimension comprises an independent group of co-occurring linguistic features, and each co-occurrence pattern can be interpreted in functional

[1] Institute of Applied informatics of Alpen-Adria University, email: anastasiagnatchuk@gmail.com
[2] Institute of Applied informatics of Alpen-Adria University, email: agnatchuk@gmail.com

terms"(14). Moreover, a researcher defines the dimensions as a set of linguistic features (linguistic variables) with their frequencies in texts.

As proposed by Biber (1988) the dimensions of "involved (interactive) versus informational production", "description versus reporting" and "narrative versus non-narrative", etc. were aimed to determine which text types are close to each other with regard to certain linguistic features. Biber intuitively distinguished interviews, personal letters and spontaneous speeches as the texts belonging to 3 different styles. When he considered the dimension of "involved (interactive) versus informational production", these texts turned out to be close to each other: they were informative in comparison with face-to-face conversation, more interactive than general fiction, official documents and press reportage. In addition, Biber groups certain linguistic features, which occur in texts. In such a way, he distinguished 7 factors (groups or textual dimension). Biber states that "Factor 1 represents a dimension marking high informational density and exact information content versus affective, interactional and generalized content" (1988:107). As this dimension encompasses a considerable portion of linguistic features, the researcher considers it a dominant linguistic dimension.

In such a way, each dimension classifies a certain set (sample) of texts differently. In particular, Biber (1988) found that personal telephone conversations have a high interaction charge of information. In contrast, academic prose and financial press reportage are full of information material. The computer program yields the dimension by grouping the features considering their co-occurrence patterns, but a linguist must interpret these dimensions.

As far as statistical studies of styles are concerned, the majority of researchers considered the distribution of certain linguistic features in texts. In particular, the measurement of structural complexity has been undertaken by considering a linguistic feature of a subordinate construction. As far as the frequencies of parts of speech are concerned, Stubbs (1980) revealed a great portion of adverbs in writing. But Blankenship (1974) did not detect any differences in the frequencies of adverbs in different styles. Poole and Field (1976) found a considerable number of adverbs in oral speech. Chafe (1982) noticed that the highest occurrences of adjectives are to be found in writing. In view of the above-mentioned contradictory findings Beaman (1984) "notes that the failure to control for differences in register, purpose, degree of formality, and planning contributes to the confusing picture emerging from previous quantitative studies" (Biber, 1988: 51). According to Biber, there is the necessity to conduct additional study in terms of other linguistic features. His approach to the variation of texts must be further studied by means of different linguistic features in order to deepen and clarify this model of variation.

2. The methodological fundamentals of research and the discussion of findings

This research is devoted to the study of English functional styles based upon the distribution of 11 parts of speech (linguistic variables) in the Brown Corpus. The data consist of 15 subsections of the Brown Corpus (*news, editorial, reviews, religion, hobbies, lore, belles-lettres, government, learned, fiction, mystery, science-fiction, adventure, romance and humor*) (Bird, 2009: 43). We are going to conduct a multivariate analysis of functional styles. Under multivariate we understand the analysis of 11 parts of speech which are analyzed in 15 subsections of the Brown Corpus. "Each text can be given a precise quantitative characterization with respect to each dimension, in terms of the frequencies of the co-occurring features that constitute the dimension" (Bib-

er, 1988: 20). These 15 subsections are divided into 5 styles, defined by Galperin (1981): *brief news style (news, editorial and reviews), scientific style (lore, learned), belles-lettres styles (belles-lettres, fiction, mystery, science-fiction, adventure, romance, humor), official (government) and miscellaneous styles (hobbies).* The data contain the normalized values for each part of speech (Formula 1.1) in each subsection, illustrated in Table 1.

$$n = \frac{k}{N} \qquad (1)$$

n = a normalized value;
k = absolute frequency in a subsection;
N = the total number of words in each subsection

Table 1
The normalized values of parts of speech for 15 subsections from the Brown Corpus

	styles	**Noun**	**Verb**	**ADP**	**DET**	**ADJ**	**ADV**	**CONJ**	**Pron**	**PRT**	**Num**	**Others**
News	*News*	0.34	0.16	0.13	0.12	0.07	0.03	0.03	0.02	0.02	0.02	0.00
Editorial	*News*	0.27	0.18	0.13	0.13	0.09	0.05	0.03	0.04	0.02	0.01	0.00
Reviews	*News*	0.29	0.15	0.13	0.13	0.10	0.05	0.04	0.03	0.02	0.01	0.00
Religion	*Misc*	0.24	0.17	0.15	0.14	0.08	0.06	0.03	0.05	0.02	0.01	0.00
Hobbies	*Misc*	0.29	0.17	0.13	0.13	0.09	0.05	0.04	0.03	0.02	0.01	0.00
Lore	*Scient*	0.27	0.17	0.14	0.13	0.08	0.05	0.03	0.04	0.02	0.01	0.00
Belles lettres	*Belles Lettres*	0.25	0.17	0.15	0.14	0.08	0.05	0.03	0.05	0.02	0.01	0.00
Government	*Official*	0.31	0.15	0.16	0.12	0.09	0.03	0.04	0.02	0.02	0.02	0.00
Learned	*Scient*	0.28	0.16	0.15	0.13	0.09	0.05	0.03	0.02	0.02	0.01	0.00
Fiction	*Belles Lettres*	0.23	0.21	0.12	0.13	0.06	0.06	0.03	0.08	0.03	0.00	0.00
Mystery	*Belles Lettres*	0.22	0.21	0.11	0.12	0.05	0.07	0.03	0.09	0.05	0.00	0.00
Science_ Fiction	*Belles Lettres*	0.22	0.21	0.12	0.13	0.07	0.06	0.03	0.07	0.04	0.00	0.00
Adventure	*Belles Lettres*	0.22	0.21	0.12	0.13	0.05	0.06	0.03	0.08	0.04	0.00	0.00
Romance	*Belles Lettres*	0.21	0.21	0.11	0.12	0.06	0.06	0.04	0.09	0.04	0.00	0.00
Humor	*Belles_ Lettres*	0.24	0.19	0.13	0.13	0.07	0.06	0.03	0.07	0.03	0.00	0.00

From Table 1 one can see 11 parts of speech, which are identified as tagged Brown Corpus in the program of Python. In other words, we used the tagging algorithm. According to Bird et al. (2009), ADJ stands for adjectives (*beautiful, amazing, fantastic*), ADP – adpositions (*on, of, at, with, by, into, under*), ADV – adverbs (*happily, really, still*), CONJ – conjunctions (*and, or, while, although*), DET – determiner, article (*a, an, the, some, most, no, every, which*), Noun – nouns (*sister, friend*), Numeral – numeral (*twenty-eight, 15:15, fifth, 1988*), PRT – particles (*at, on, out, per*), PRON – pronouns (*he, she, we, you*), Verb – verbs (*to speak, do, cook*), Others –

other things such as borrowings (*ersatz*) or something unclear. Under 'styles' we shall understand 5 styles, which are distinguished by Galperin (1981) on the basis of lexical features (a certain word stock, specific lexical stylistic devices: *repetition, metaphors, epithets, irony, zeugma, simile, periphrasis, euphemism, cliché, etc.*) as well as syntactic features (*parallel constructions, enumerations, elliptical sentences, repetitions, anaphora, epiphora, suspense, antithesis, inversion, etc.*).

It is worth mentioning that we have found the counts for each part of speech in Python by writing the appropriate program code. In such a way, we have received the absolute frequencies of each part of speech for each subsection. In Table 1 we have given the normalized values (relative frequencies) for each word category according to Formula 1.

3. The conduction of Principal Component Analysis

The first step of this analysis presupposes checking one requirement. In particular, our variables (word categories) must be intercorrelated. A high positive correlation will be the indicator that the two parts of speech systematically occur together. A negative correlation shows that the occurrence of one part of speech is highly connected with the absence of another part of speech. For this purpose, we can use the Pearson correlation analysis according to Formula 2. In our research we computed the Pearson correlation in R-Studio Program.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad (2)$$

r is the Pearson correlation coefficient;
x and y are two vectors (or sets of numbers) for certain variables;
x and y (with strokes) are the sample means of the two arrays of values.

Table 2
Correlations between word categories

	Noun	**Verb**	**Adpos**	**Det**	**Adj**	**Adv**	**Conj**	**Pron**	**PRT**	**Num**	**Others**
Noun	1.00	-0.9	0.65	-0.10	0.66	-0.93	-0.23	-0.92	-0.84	0.90	0.14
Verb	-0.90	1.00	-0.83	-0.17	-0.84	0.82	-0.03	0.96	0.95	-0.82	-0.30
Adpos	0.65	-0.83	1.00	0.38	0.77	-0.71	0.09	-0.85	-0.89	0.73	0.09
Det	-0.10	-0.17	0.38	1.00	0.23	0.11	0.02	-0.12	-0.31	-0.14	0.10
Adj	0.66	-0.84	0.77	0.23	1.00	-0.56	0.18	-0.85	-0.90	0.55	0.38
Adv	-0.93	0.82	-0.71	0.11	-0.56	1.00	0.16	0.86	0.80	-0.92	-0.02
Conj	-0.23	-0.03	0.09	0.02	0.18	0.16	1.00	0.10	-0.03	-0.21	0.37
Pron	-0.92	0.96	-0.85	-0.12	-0.85	0.86	0.10	1.00	0.95	-0.88	-0.15
PRT	-0.84	0.95	-0.89	-0.31	-0.90	0.80	-0.03	0.95	1.00	-0.76	-0.27
Num	0.90	-0.82	0.73	-0.14	0.55	-0.92	-0.21	-0.88	-0.76	1.00	-0.12
Others	0.14	-0.30	0.09	0.10	0.38	-0.02	0.37	-0.15	-0.12	-0.12	1.00

As one can see there are the cases of both positive and negative correlations.

The next problem according to Levshina (2015:354) is to determine *the number of dimensions or components*. In this case we shall deal with the concept of eigenvalue. The eigenvalue "shows how much of the total variance is explained by each component. The higher the correlations between a component and the variables, the greater the component's eigenvalue" (Levshina, 2015: 354–355).

Table 3
Eigenvalue percentage of variance and cumulative percentage of variance

	Eigenvalue	Percentage of variance	Cumulative percentage of variance
Component 1	6.82	62.03	62.03
Component 2	1.79	16.28	78.32
Component 3	1.13	10.27	88.59
Component 4	0.69	6.30	94.8
Component 5	0.28	2.63	97.5
Component 6	0.14	1.28	98.8

From Table 3 one can see that the first component explains the highest portion of variance (62.03 %). Nevertheless, the 3rd column displays that the percentage of each component increases. Aiming to choose the optimal number of components, it is necessary to consider the eigenvalue (the first column of Table 3) higher than 1. In this case we have the first 3 components. We can also visualize it by means of a scree plot (Figure 1).

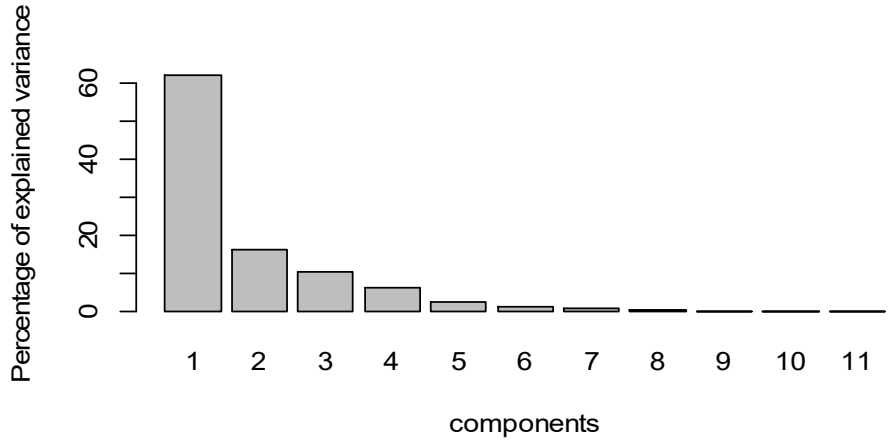

Figure 1. Contributions of PCA components to the variance of eigenvalue

As one can see, the x-axis is represented by 8 components (100 % of explained variance) created by the algorithm in R-Studio. It is possible to notice that there is no substantial decrease after the third dimension. In general, our scree plot with three components explains 88.59 % of variance and they have the eigenvalue higher than 1 (Figure 1).

The decrease can be captured by means of the exponential function y = a*exp(-b*x) where a = 241.6801 and b = 1.3200. The determination coefficient is $R^2 = 0.9843$.

The next task will be to interpret the dimensions by means of Variables Factor Map (Figure 2):

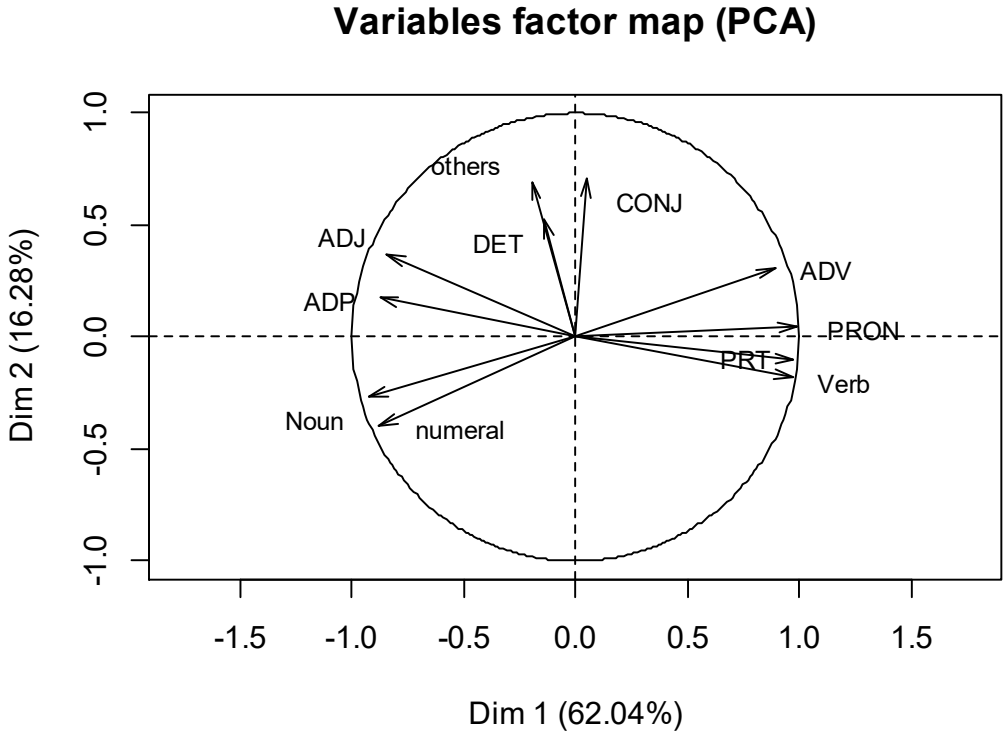

Figure 2. Principal Component Analysis with the 1st and 2nd components

Here one can see two components on two axes (78.3% of explained variance). The further interpretation of figures and the techniques, used in this research, are given in Book "How to do Linguistics with R" by Levshina (2015, 355:361). The arrows serve here as variables (ADJ, ADP, Verb, PRON, NOUN, etc), pointing from the center. The angles between two arrows signal to the strength of correlation: if the angle is small, then the correlation is strong and if the arrows move in the same direction, then the correlation is high. Our attention must be paid to the four circle sectors in Figure 2. The first dimension of Figure 2 (a horizontal axis x) deals with interactive or involved information. The first sector on the right comprises CONJ, ADV and PRON. The second sector on the right includes Particle and Verb. These sectors contain positive values on the horizontal axe (from 0.0 to the further positive values). These parts of speech are the features that indicate interaction character of information (positive values on x-axis are associated with interaction whereas negative values – with high information density). The third sector on the left contains Noun and Numeral. The fourth sector on the left includes parts of speech ADP, ADJ, DET and others. The 3rd and the 4th sectors (-) on the left display a high information charge of communication (texts). These sectors are on the horizontal axe (from -1.5 to 0.0). Nevertheless, it is necessary in this case to consider the direction or orientation of these arrows (variables). Let us consider the correlation coefficients in Table 4 for word categories and interpret them.

Table 4
Correlation coefficients for parts of speech for the 2nd dimension

	Correlation	**p-value**
Pronoun	0.98	$5.65e^{-12}$
Verb	0.97	$7.30e^{-10}$
Particle	0.97	$1.74e^{-09}$
Adverb	0.89	$8.13e^{-06}$
Adjective	-0.84	$8.16e^{-05}$
Adposition	-0.87	$1.56e^{-05}$
Numeral	-0.87	$1.56e^{-05}$
Noun	-0.92	$9.88e^{07}$

We shall start with the consideration of positively correlated coefficients and compare them with Figure 2. The positive correlations are to be found in Pronoun, Verb, Particle and Adverb which have quite high correlation coefficients. These findings are supported by the horizontal directions (or to the right direction) of the arrows in Figure 2 for our variables. The negative correlations are in ADJ, ADP, Numeral and Noun, which are to be found on the left part of the plot. In addition, these correlations are statistically significant by considering the level of 0.05. Nevertheless, we shall have a look at the estimates of regression coefficient for a qualitative response variable: style.

Table 5
The estimate of regression coefficients for qualitative variable of style

	Estimate	**p-value**
Belles lettres	3.78	$1.46e^{-05}$

As one can see, there is only one positive estimate (3.78) for belles-lettres style. Positive result (the 1st and 2nd sectors on the right) is connected with interactive communication. But there is no surprise as we consider (compare) the belles-lettres with scientific, official, news and miscellaneous styles. The belles-lettres style has a considerable number of dialogues. This shows us an interactive orientation of the belles-lettres style in comparison with the other styles (colloquial speeches have not been included in our corpus).

The next step is to visualize the subsections of the Brown Corpus in the plot (onto the space) and to see where our 5 styles *(brief news items, scientific items, belles-lettres items, official and miscellaneous*) are located. Figure 3 displays the results considering only the first and the second component. As one can see, the belles-lettres style (designated as a square) is to be found mostly on the right space of Figure 3. The right orientation (the first and the second rectangles on the right) is associated with the interaction or involvement of the information. The subsections of *humor, adventure, science-fiction, fiction, romance, adventure and mystery* (marked as gray labels) are placed on the right (they are united by belles-lettres functional style). The scientific style is to be found on the left space (the third and the fourth rectangles on the left), which accounts for informational charge of communication (negative values on the x-axis). Parallel to the scientific items, it is also possible to find on the left space the styles of official, brief news

and miscellaneous styles. They share informational density of the communication. It would be recommended in our corpus to include a selection of oral (colloquial) texts, as colloquial speech is supposed to be more available in our corpus as a part of the belles-lettres items in comparison with scientific, official, news and miscellaneous texts. The presence of oral texts may have an influence on the position/place of belles-lettres style on the map.

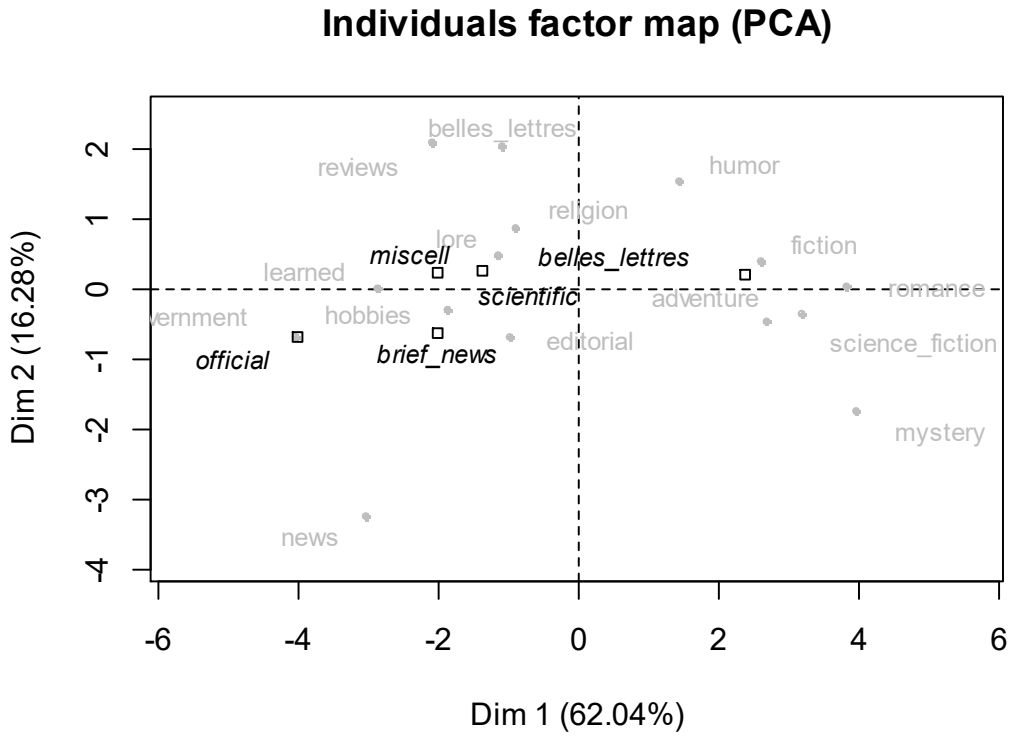

Figure 3. Direction or orientation of functional styles in terms of the 1st and 2nd component

The next task is to consider the variables factor map (Figure 4) with the 2nd and 3rd component (the second dimension). According to Levshina (2015) the 2nd dimension can be described as "description vs. reporting of past events" (358). In Figure 4 the first sector (*determiner, adposition, adverb*) and the second sector (*adjective, conjunctions, others*) on the right (+) represent *reporting of events* (positive values on x-axis). The remaining two sectors (-) account for the *description of events* (negative values on y-axis). When considering the third dimension, there is the distinction between *narrative-non-narrative texts*. In particular, the 4th sector on the left and the 1st sector on the right (Verb, Determiner, Adverb) show narrative texts (positive values on y-axis). The 3rd sector on the left (Noun, Pronoun and Numeral) and the 2nd sector on the right (Adjective, Conjunctions and Others) account for non-narrative texts (y-axis with negative values).

In view of the 2nd dimension, positive correlation estimates are for *conjunction, others and determiners*, shown in Table 6 at the level of significance 0.05:

Table 6
Correlation coefficients for parts of speech for the 2st dimension

	correlation	**p-value**
Conjunction	0.70	0.003
Others	0.69	0.004
Determiner	0.52	0.044

As far as the second dimension is concerned, the belles-lettres texts have significant correlations with the positive values in the 1st and 2nd sectors of Figure 4. Looking at Figure 5 one can find the 2nd rectangle areas on the right, occupied by the belles-lettres style. This makes a distinction of belles-lettres style from other ones.

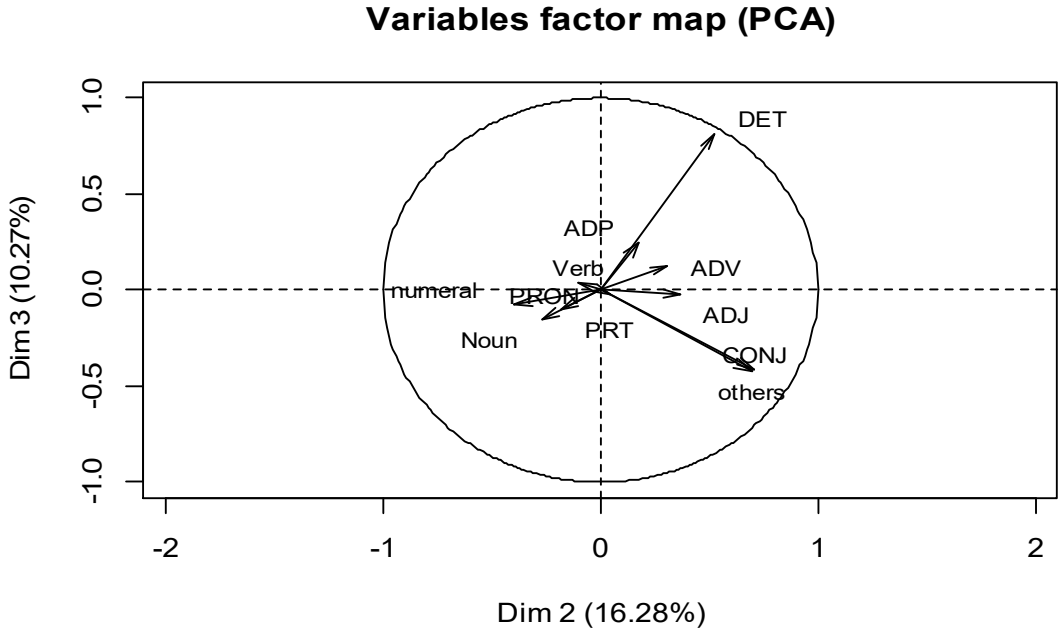

Figure 4. The Principal Component Analysis with the second and the third components

Table 7
Correlation coefficients for parts of speech for the 3rd dimension

	correlation	**p-value**
Determiner	0.81	0.000

We also plot the obtained results in Figure 5 and interpret them on the basis of Figure 2 and Figure 4. The results show us that

- ✓ the belles-lettres style is highly interactive according to the 1st component *(interactive (right) vs. information (left);*
- ✓ the belles-lettres style tends to report an event (the 2nd rectangle area in Figure 3) according to the 2nd dimension (*description versus reporting events*);
- ✓ the belles-lettres texts are slightly non narrative (the 2. rectangle area on the right) than narrative (the 1nd rectangle area on the right) (*narrative versus non-narrative*).

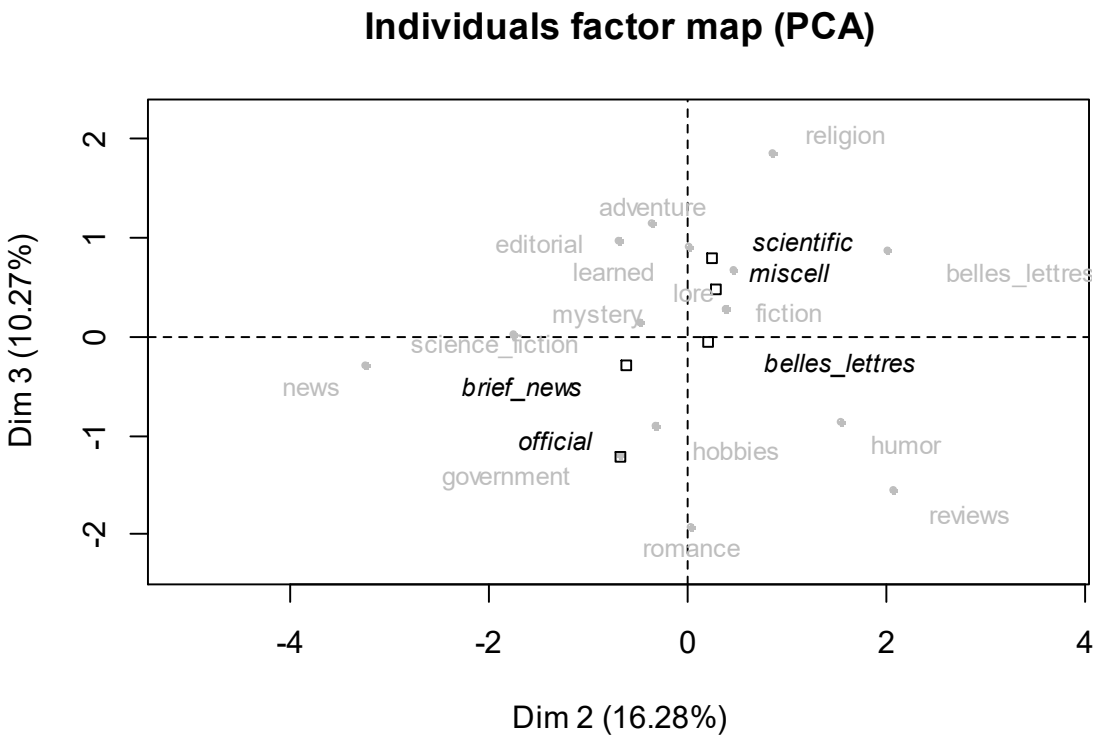

Figure 5. The replacement of 15 Brown subsections in accordance with Principal Components 2 and 3.

Levshina proposes in her book "How to do Linguistics with R" to "plot the confidence ellipses around the centroids to estimate the amount of overlap of the prototypes of the registers (dimensions 1 and 2)" (2015:359). In other words, we want to see which styles overlap considering the distribution of parts of speech. In accordance with it, we have received Figure 6.

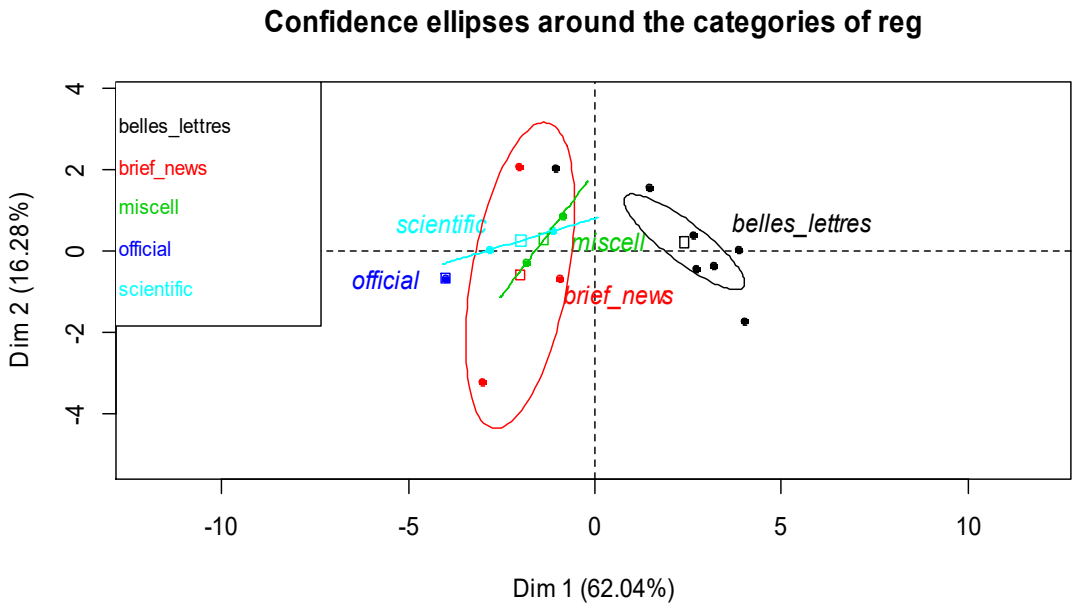

Figure 6. Confidence ellipses around the categories of styles

As one can see, the ellipsis of brief news items overlaps with scientific and miscellaneous within the 1st dimension of "interaction versus information". Only belles-lettres style has no coincidence with other styles in terms of the 1st dimension.

3.1. Factor analysis

According to Levshina (2015), "the main purpose of PCA is to find as few orthogonal (uncorrelated) components as possible while maximizing the total explained variance. It is used mainly to reduce dimensionality. In contrast, Factor Analysis (FA) is more widely used for exploring theoretical constructs, or latent variables, which are called factors" (361). In addition to it, FA rotates these variables or factors in order to enlarge the charge of variables on some similar factors. In other words, the factor analysis is intended to lessen (minimize, simplify) a set of analysed variables by rotating the factors. In such a way, there is a technical difference that differentiates FA from PCA. According to Biber (1988), "factor analysis is the primary statistical tool of the multi-feature/multidimensional approach to textual variation" (79). The factors correspond to the variables in our study. If several parts of speech have a high correlation (i.e. two parts of speech occur together), then we can define 1 factor. In other words, a factor is a set of linguistic features (in our case, parts of speech) which have a high frequency in a text. Before performing the analysis, it is necessary to determine the number of factors. In this case, we consider the optimal number of components in the previous study, which equals to 3.

Table 8
Factor loadings

	Factor 1	**Factor 2**	**Factor 3**
Noun	-0.98	-0.15	
Verb	0.89		-0.20
Adposition	-0.74	0.28	0.14
Determiner		1.05	-0.11
Adjective	-0.61		0.69
Adverb	0.98	-0.10	0.16
Conjunction	0.21	-0.14	0.51
Pronoun	0.93		-0.20
Particle	0.85	-0.17	-0.25
Numerals	-0.97	-0.14	-0.15
Others			0.51

Correlation coefficients are similar to factor loadings, but the difference is that the numbers of factor loadings can be higher than 1 or lower than -1. In general, these loadings, higher than 0.3, are very significant.

As one can see from Table 8, our first factor bears much similarity to the first Principal Component (see Table 4). In particular, very strong negative correlations are to be found for noun, numeral, adposition and adjective (negative loading) as well as positive correlations (positive loading) for pronoun, verb, particle and adverb. In such a way, this first factor distinguishes **interactive versus informational communication (texts)**. And these parts of speech with positive factor loadings can be used for informational communication (*news, government, learned, hobbies, reviews, lore, editorial*), whereas the parts of speech with negative factor loadings for interaction (*adventure, fiction, humour, science fiction, mystery, romance*). Figure 7 illustrates the subsections and their scores concerning Factor 1 as well as Factor 2. From Figure 7 one can see (we consider the horizontal direction) negative values on axis x, which account for informational communication (government, news, hobbies, learned, reviews, lore, editorial). Religion and belles-lettres are neutral in this case. Interactive communication deals with positive values (adventure, fiction, humour, science-fiction, mystery and romance). These literary pieces are supposed to have a variety of dialogues between characters in comparison with information-oriented texts.

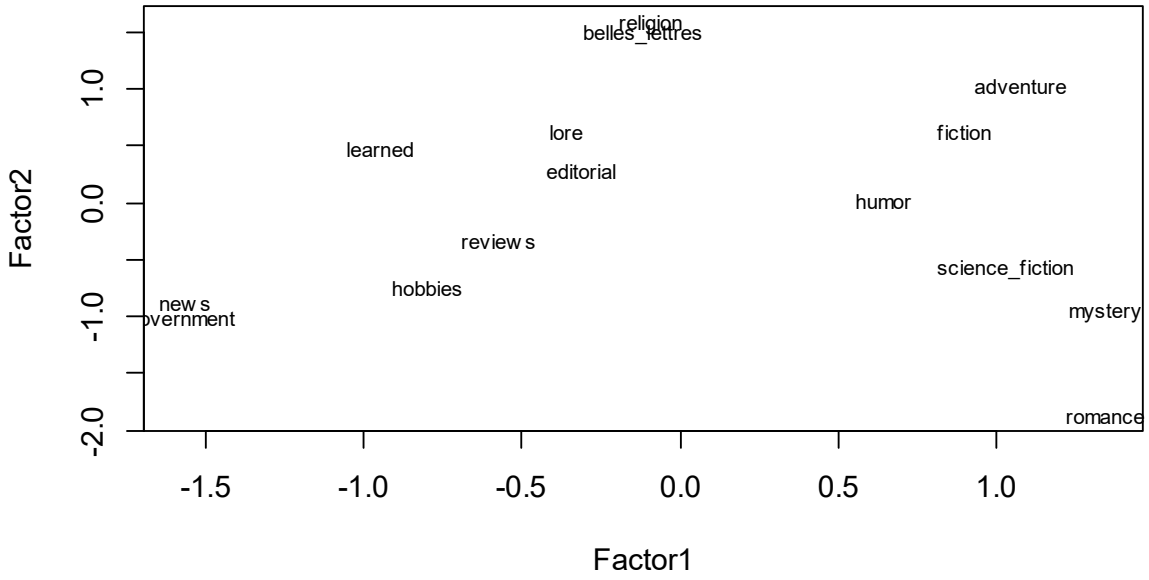

Figure 7. The scores of subsections from the Brown Corpus considering Factor 1 and 2

As far as Factors 2 and 3 are concerned, Factor 2 is concerned with **description versus reporting of some events**. Table 5 shows that positive correlations are for conjunctions, determiners and others. Only positive factor loading is to be found for determiner in Table 5. The second factor is not similar to the second Principal Component (only with a positive correlation for determiner). It would be better to have a look at the subsections and their scores: the description of the events is represented by negative values on the x-axis (romance, government, mystery, news, hobbies, science-fiction and reviews). Humour is neutral in terms of description versus reporting of some events. Reporting of events is characteristic of positive values (editorial, learned, lore, fiction, adventure, religion and belles-lettres).

As far as Factor 3 (narrative versus non-narrative) is concerned, the vertical axis of Figure 8 shows narrative (positive) and non-narrative (negative values) communication. To narrative belong romance, hobbies, science-fiction, reviews, editorial, learned and belles-lettres. In this case positive correlation is for determiner in Table 5 and Factor 3 does not contain any value for this part of speech. That's why this dimension is unclear in this case.

Nevertheless, it is necessary to check whether the number of factors is sufficient in our research. In this case one has to consider a p-value in Table 8. If the p-value is smaller than 0.05, than the number of factors are not sufficient. In our case, p-value is 0.00041 which shows that these 3 factors are not enough for the given research. It is recommended to increase the number of factors in our model.

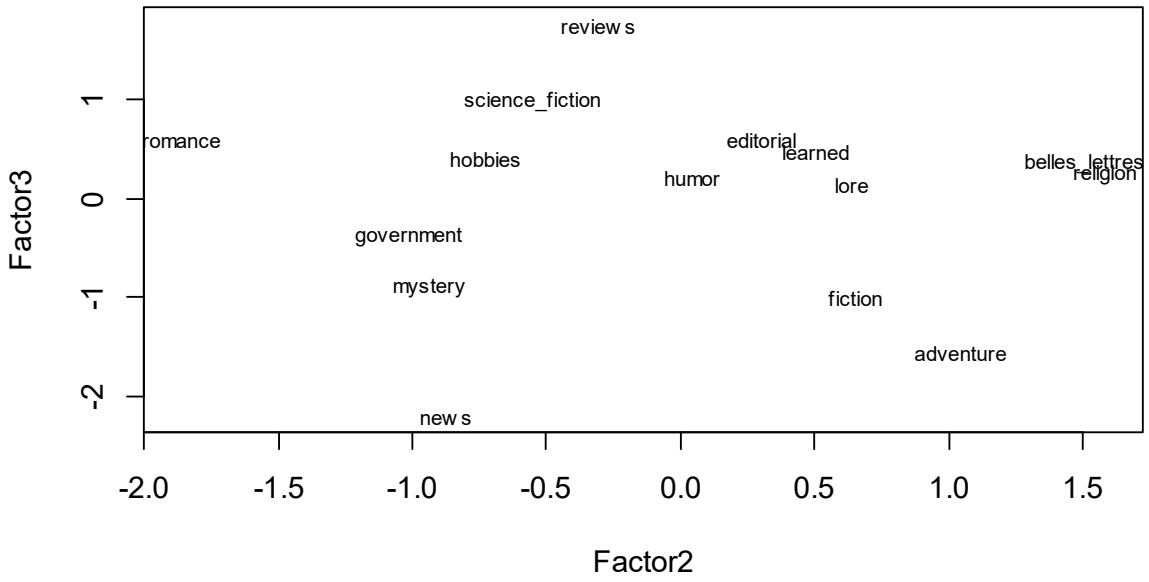

Figure 8. The scores of subsections from the Brown Corpus considering Factor 2 and 3

In such a way, we have conducted a multidimensional study of the identification of styles in the British corpus. Our study was devoted to extend and clarify the model of variation of texts by Biber (1988) as there was the necessity to consider other linguistic features which are parts of speech. PCA and FA have helped us find the dimensions (factors) which can be applied to the styles variations. It is quite obvious from the results that only Factor 1 has a distinguishing power in terms of interactive versus informational communications. Other factors (2 and 3) seem not to contain a distinction power as they contrast the results with PCA and look quite tricky. In addition, the number of dimensions (factors) must be extended as the p-value ($p = 0.00041$) has shown insufficient number of factors. It would also be recommended to consider other linguistic features in order to enrich the model of variation by Biber. In addition, it would be relevant to consider the other languages and compare the results.

References

Beaman, K. (1984). Coordination and subordination revisited: syntactic complexity in spoken and written narrative discourse. In: *Coherence in spoken and written discourse,* ed. By Deborah Tannen, pp. 45-80. Norwood, N.J.: Ablex.

Biber, D. (1988). *Variation Across Speech and Writing.* Cambridge University Press. Cambridge

Bird, S., Klein, E., Loper, E. (2009). *Natural Language Processing with Python*. O'REILLY.

Blankenship, J. (1974) The influence of mode, submode, and speaker predilection on style. *Speech Monographs* 41, 85-118.

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In: *Spoken and written language: exploring orality and literacy,* ed. By D. Tannen, pp. 35–54. Norwood, N.J.: Ablex.

Galperin, I. R. (1981). *Stylistics.* Moscow Vysshaja Shkola.

Levshina, Natalia. (2015). *How to do Linguistics with R: data exploration and statistical analysis.* John Benjamins Publishing Company. Amsterdam/Philadelphia.

Poole, M. E., Field T.W. (1976). A comparison of oral and written code elaboration. *Language and Speech* 19:305–311.

Stubbs, M. (1980). *Language and Literacy: the sociolinguistics of reading and writing.* London: Routledge and Kegan Paul.