Statistical exploration of situational parameters for stylistic variation in translation

Meng Ji

Humanities, Imperial College London
85 Sterling Place, South Ealing, London W5 4RB
m.ji@imperial.ac.uk

Abstract

The present paper sets out to offer an original quantitative investigation into the nature of style variation within the work of a single translator. Instead of focusing on what constitutes the basic features of one's translating style, the present paper aims to establish a tentative empirical line of argument regarding how to interpret the variation in stylistic patterns as revealed by corpus statistical techniques. It requires the quantification and statistical modelling of contextual factors that characterize or tend to co-occur with stylistic variation within the work of a translator. It is one of the few works done so far to enlarge the field of computational stylistics from a purely quantitative exploration of linguistics patterns in textual data to the integration of extra-linguistic data in qualitative textual analysis, with a view to establishing contextual factors that may hold the key to a deeper understanding of the rationale behind stylistic variation within a translator's work.

Keywords: stylistic variation, corpus stylistics, literary translations.

1. Introduction

The present paper is a progress report of the author's current research on a corpus-based study of the stylistic use of four-character expressions in two contemporary Chinese versions of Cervantes's *Don Quijote de La Mancha* (I, 1605) by Yang in 1978 and by Liu in 1995. It aims to provide an exploratory statistical account of the potential situational parameters or contextual features that may help explain the stylistic variation in the protagonist's archaic speeches detected within the first part of Liu's recent translation of the Spanish novel.

In a previous study, through the use of Pearson's moment-product correlation and multiple linear regression tests, three statistical models have been built up in an attempt to explain the stylistic use of Chinese archaic idioms by Liu in his work in relation to the original and Yang's version of the novel, which is generally held to be the first direct translation of *Don Quijote* into Chinese (all previous ones were via English). The three linear regression models are (1) Liu's use of archaic idioms as a result of the use of archaisms in both the original and Yang's translation; (2) due to the considerable linguistic contrastiveness between the two languages, there seems to be larger connection between the two sets of Chinese target texts, hence we propose to build a statistical model which treats the use of archaic idioms by Liu as a result of the use of archaic idioms by Yang in her earlier version of the novel; (3) and lastly, under the hypothesis that Liu has embarked upon the project with little previous knowledge of Yang's work, a third statistical model has been built up to examine the degrees of independence of Liu's use of Chinese archaisms from the relevant textual features of the original (see Diagram I & II).

Subdivision	Numbers of chaps. in which D.Q. speaks	Number of speeches	Number of archaic speeches (ES)	Number of archaic speeches (Yang)	Number of archaic speeches (Liu)
1	1	1	0	0	0
2	4	32	11	8	13
3	0	0	0	0	0
4	4	42	9	8	10
5	4	22	2	8	8
6	8	145	12	24	32
7	4	54	2	15	16
8	3	50	5	10	16
9	8	50	5	11	20
10	5	31	6	13	9
Total	41	427	52	97	124

Table I Distribution of Don Quijote's archaic speeches (Don Quijote, part I)

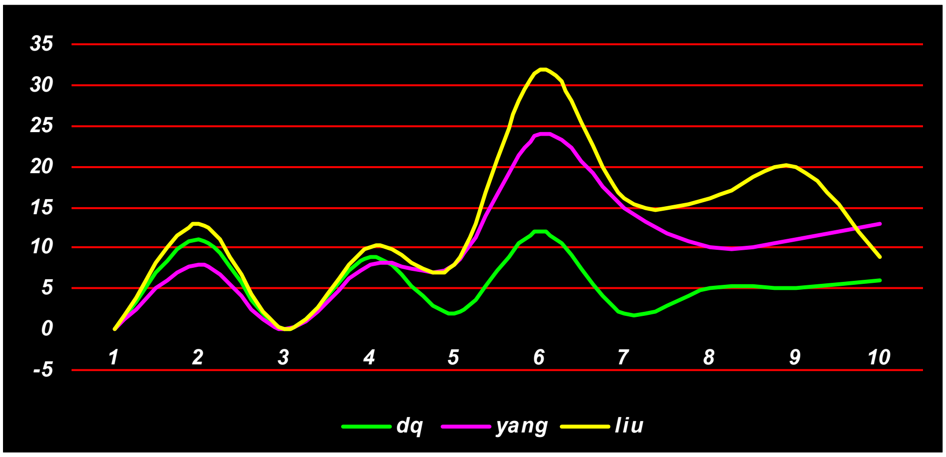


Diagram I Contrastive patterns of the distribution of Don Quijote's archaic speeches

In Diagram II, the first set of bars in black represents the actual frequency of Don Quijote's archaic speeches in Liu's translation of the novel; the following three sets of bars in gradually faded colours stand for the predicted frequency of Don Quijote's archaic speeches in Liu's work as calculated on the basis of the three statistical models thus proposed. A quick comparison of the actual account of the protagonist's archaic speeches and the three sets of hypothesized values leads to the conclusion that Model I, which has been built upon the hypothesis that Liu's use of archaic idioms has been a result of the textual features of both the original and Yang's earlier version of the novel, has proved to be relatively more successful in predicting the behaviour of language archaisms in Liu’s work, since the predicted frequency that Model I has yielded draws quite close to the real breakdown of frequency embodied in the first set of black bars.

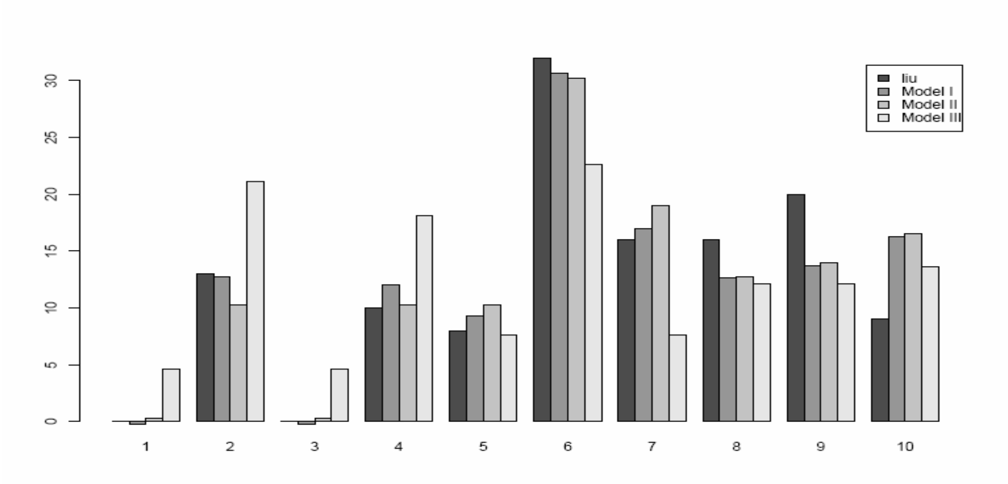


Diagram II Statistical modelling of the contrastive patterns of Don Quijote’s archaic speeches

However, if we were to have a closer look at Diagram II, we would also discover that despite the high rate of precision that has characterized the performance of Model I on Liu’s translation across the first seven subdivisions of the first part of the novel, major discrepancies between the actual frequency and the hypothesized occurrence of the protagonist’s archaic speeches as identified in Liu’s work begin to emerge as from Subdivision VIII. While in Subdivision VIII and IX, the rate of Don Quijote’s archaic speeches seems to overrun significantly its hypothesized counterpart; in the last subdivision of the first part of the novel, the frequency of the knight’s archaic speeches drops sharply when compared to the previous two; this is also the case when compared with the standard rate as suggested by the height of the two adjacent bars representing Model I and II.

This actually raises the issue of stylistic variation within the particular work of a given author, which has been discussed previously (Hoover, 2003: 341-60). However, the novel points presented in the current study consist in that in the place of devising a modified framework of variables that may increase the sharpness of statistical tests, e.g. cluster analysis, factor analysis, in distinguishing the stylistic variation among texts or chapters within the single work of a single author, it aims to furnish a qualitative analysis of the contextual factors that may help to explain the stylistic shift within the same work of a single translator. That is, rather than pursuing the issue of what constitutes the basic features of one’s literary style, which has been discussed exhaustively in the literature, the present paper aims to establish a tentative empirical line of argument regarding how to interpret the variation in stylistic patterns as revealed by corpus statistical techniques. It requires the quantification and statistical modelling of contextual factors that characterize the stylistic variation within the work of a single translator. It is one of the few works done so far to enlarging the field of computational stylistics from a purely quantitative exploration of linguistics patterns in textual data to the integration of extra-linguistic data in quantitative textual analysis, with a view to establishing contextual factors that may hold the key to a deeper understanding of the rationale behind stylistic variation within a translator’s work.

2. A new line of inquiry into stylistic variation in translation

In an effort to explain the idiosyncrasies in Liu's use of archaic idioms in the last three subdivisions of the first part of the novel, we first exclude the possibilities of the influence coming from the textual features of the original or Yang's first translation of the novel, for these factors are mainly accountable by the three statistical models introduced above, which have only proved to be efficient or statistically reliable in predicting the use of archaic idioms by Liu up to Subdivision VII. In other words, the unusual number of archaic idioms by Liu in the last three subdivisions of his translation is exactly in the textual proportion of Liu's work that cannot be easily seen as a consequence of the use of archaisms in the original or the use of archaic idioms in Yang's work. Given that it would turn out to be quite hard to pin down subjective factors such as the translator's peculiar writing habits, which have led to the style variation, we propose to investigate the nature of such style variation in two ways.

Firstly, we classify and annotate all the instances of the protagonist's archaic speeches by running the CATPCA to see whether the clustering results show any consistent patterns regarding the distribution of Don Quijote's archaic speeches in the last three subdivisions of Liu's translation. If the clustering result fails to distinguish the subjects of different sources of origins by offering a fairly mixed visualization of the distribution of Don Quijote's archaic speeches, we may probably assume that the stylistic variation under investigation may well have been due to an unconscious decision made on the part of the translator while working on the last three subdivisions of the novel; on the other hand, if after running the CATPCA on the dataset, revealing patterns begin to emerge which help to separate the subjects according to their appearance in each subdivision, we should then consider the style variation as a reflection of the deliberate strategy taken on the part of the translator in an attempt to foster a particular language style of his own.

Next, having confirmed the second presumption regarding the existence of a deliberate stylistic variation in Liu's work, we would also like to know what are the contextual factors that have favoured the stylistic variation in the texts, i.e. what are the situational features that tend to co-occur with the increase of archaic speeches in the protagonist's utterance in Subdivision VIII and IX; or what are the situational parameters seemingly associated with the reduction in the number of Don Quijote's archaic speeches in the last subdivision of the novel. Lastly, it should be pointed out that given the nature of CATPCA, it is hard to make any causal claims with regards to whether particular situational parameters have caused such stylistic shift. However, a closer look at the contextual factors that have been designated by the statistical test as especially relevant to the automatic grouping of some of the protagonist's archaic speeches, will have important implications for us to decide upon the effectiveness of the analytical framework thus devised for clustering purposes, as well as the need for a better analytical tool for style or register variation in textual analysis.

The analytical framework of situational parameters adopted in the present study has been built upon the original one developed by Biber in his famous sociolinguistic study of English register variation (Biber, 1994: 40-1). The framework provided in Appendix I is a modified version of the original one, and certain structural changes have been made to meet the requirements of the design of the current study, which is to identify contextual factors that help describe the apparently excessive use of archaic idioms by Liu in Subdivision VIII and IX, and the rather reduced use of archaic idioms in Subdivision X. Specifically speaking, the structural modifications made of the original framework include: firstly, due to the specific topic under discussion here, i.e. language archaism, which is closely related to the concept pair of formalism versus in-formalism, any item listed under the category of Feature in the original framework that seems to be of minor relevance to this specific concept pair is taken out to increase the explanatory power of this scheme, as well as to reduce the workload caused by applying a rather general and ample network to a domain-specific project of textual analysis. For example, due to their limited relevance to the present study, items falling under the category of Features in the original framework, e.g. mode or primary communication channel including written/spoken/signed/mixed (other) or medium of transmission, have been eliminated from the framework used in the current study. Secondly, as an essential part of the requirements of the statistical test to be used, i.e. categorical principal component analysis, also known as CATPCA in the Statistical Package of Social Sciences (SPSS), each of the selected category under Feature is given a rank of values which exhibits the nature of ordinal units of measurements, i.e. from the least to the most or from the farthest to the nearest, as required in this type of textual statistics. The only exception is Specific Subject, which will be identified later on as nominal measurement in using the CATPCA. Lastly, to facilitate the statistical analysis, each value of a certain textual feature or situational parameter in Biber's terms, is given a numerical code ranging from one to five depending on the inherent complexity of each textual feature.

3. CATPCA Performance Results

The idea underlying the categorical principal component analysis or CATPCA is to reduce the dimensions of analysis from a large number of variables into a smaller set of independent principal components, which have been built by following certain statistical procedures (Oakes, 1998: 96). The principal components (PC) thus created may be seen as a linear combination of a set of variables that helps explain as much as possible the variance shown among the subjects under investigation, in our case, the use of archaic idioms by Liu in the last three subdivisions of his work. Table II provides a summary of two-dimensional PC model built by running the CATPCA on the corpus data gathered in each subdivision, which have been obtained by following the scheme shown in Appendix I.

Dimension	Cronbach's Alpha	Variance Accounted For	
	Total (Eigenvalue)	% of Variance	Total (Eigenvalue)
1	0.829	4.336	30.969
2	0.769	3.501	25.009
Total	0.940(a)	7.837	55.978

Table II Model summary

As shown in Table II, the Eigenvalue attributed to each dimension, which is indicative of the extent of variance to which each principal component is able to contribute regarding the use of archaic idioms in different contextual settings as quantified by different variables or situational parameters specified in Appendix I. Dimension I (PC1) has an Eigenvalue of 30.969, which means that it can account for nearly one third of the variance in the entire dataset, while Dimension II (PC2) may help explain another one fourth of the information contained in the original set of variables. The relatively low total Eigenvalue, on the one hand, suggests the inherent disparity of archaic idioms across the three subdivisions, which can be well expected, and on the other hand, it reflects the genuineness of the corpus data thus gathered, since information retrieved from naturally-occurring language tends to show less uniformity when studied within the framework of purposely built-up framework of analysis.

Diagram III displays the distribution of archaic speeches marked with their source of origin within the two-dimensional space generated by the CATPCA. As can be seen from Diagram III, the original dataset has been grouped into four clusters with most data points marked by number 8 or 9 gathered in the lower-left quarter of the graph. If we were to draw an added X-axis through the zero point on the vertical scale while making it parallel to the horizontal scale at the bottom of graph, we can see that the main body of points marked by number 10 floats above the added X-axis. Also, although not very significant, the main cluster of points 10 shows certain visible distance from the main cluster of points 8 & 9. This is suggestive of the fact that despite the relative low total Eigenvalue in explaining the textual characteristics of the whole dataset, the two-dimensional PC graph has been quite successful in separating the protagonist's archaic speeches appeared in Subdivisions VIII & IX from those found in Subdivision X. The clustering result is suggestive of the fact that the style variation detected in the last three subdivisions of the novel has been due to a deliberate decision made on the part of the translator. This can be seen as the essential first step of achieving the goal of the present study, i.e. to separate archaic speeches from different subdivisions and then to identify the potential textual factors or situational parameters that characterize the previously detected variation in the use of archaic idioms within the highlighted parts of the novel.

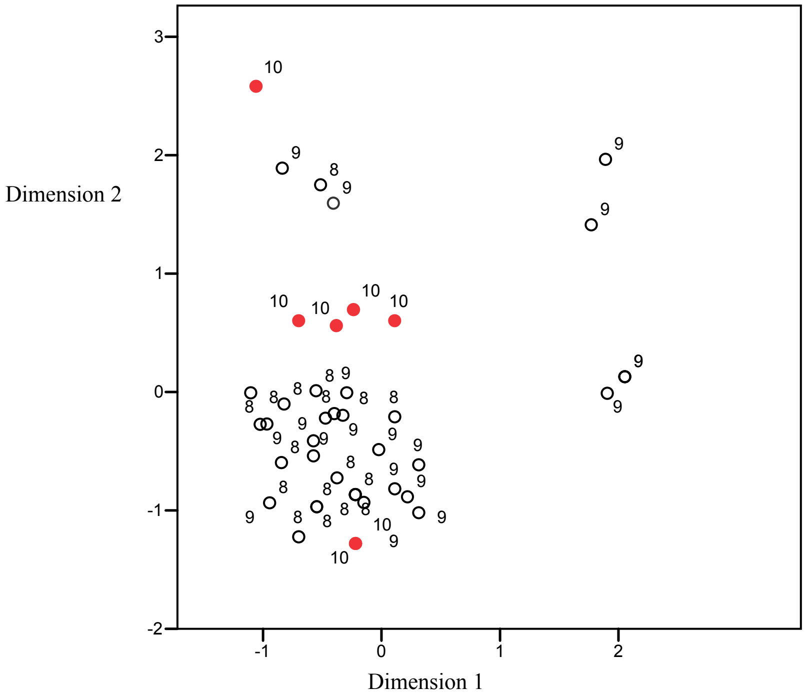
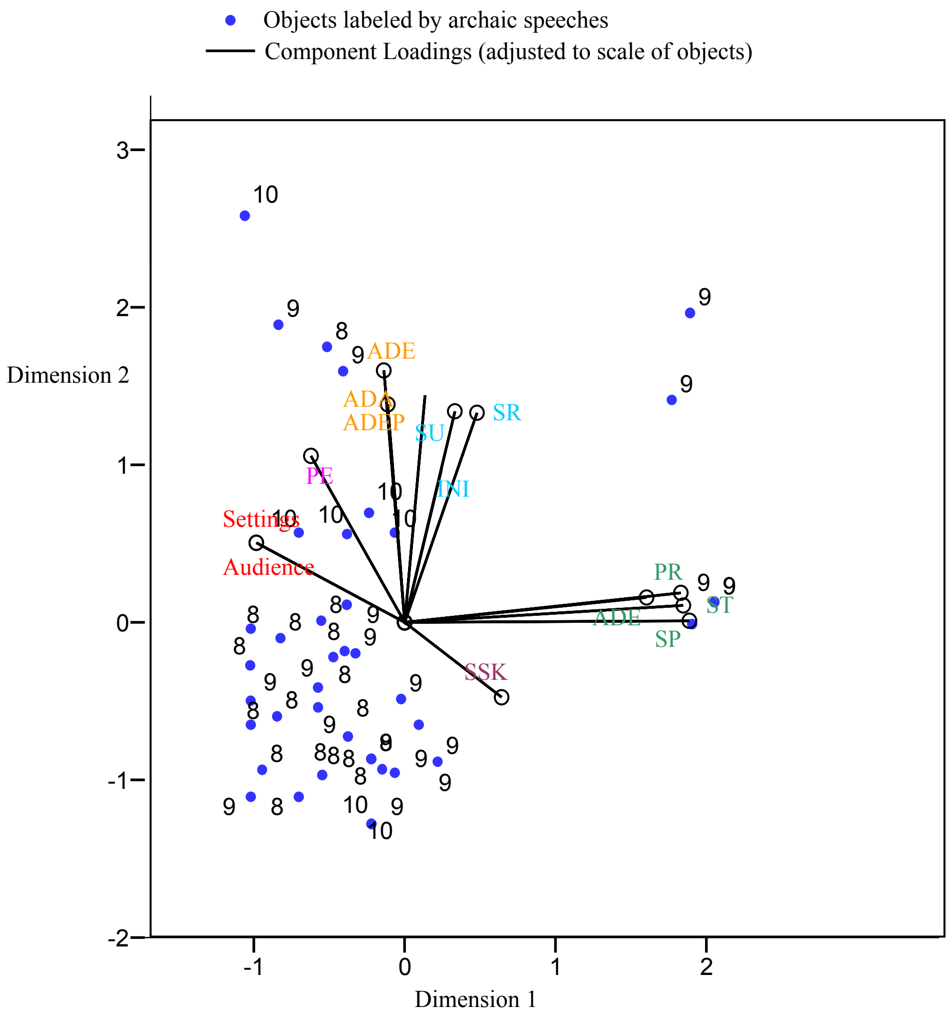


Diagram III Object points labelled by archaic speeches

Now, an interesting question worth asking here is what are the specific sets of variables or situational parameters that have separated most points marked by 8 and 9 from those marked by 10. To answer this question, we would need to examine the distribution of archaic speeches within the two-dimensional space as substantiated by the different sets of original variables or situational parameters. Diagram IV offers a combinatory projection of the subjects (or archaic speeches) and the variable sets against the two-dimensional graph built by the CATPCA.



N.B. ADE= Addressor's evaluation of speech content; ADA= Addressor's attitude to speech content; ADEP= Addressor's epistemological stance on speech content; AEE= Addressee's evaluation of speech content; PR= Personal relationship; SPK= Shared personal knowledge; SSKT= Shared specific knowledge of topic;

Diagram IV Biplot of component loadings and objects

In Diagram IV, the direction of the vector (line) points from the lowest to the highest score of the ordinal or nominal rank, i.e. from 1 to 3 or 4 or 5 depending upon the nature of each variable; and the length of its projection on each dimension indicates the degree of variance that it can explain among the subjects along that dimension. The longer the vector, the more effective the variable is in describing the contextual features that characterize the occurrence of the protagonist's archaic speech in different communication settings.

Closely related variables have been marked with the same colour to facilitate further qualitative analysis: for example, (1) the addressor's evaluation of textual events, the addressor's attitude towards textual events, and the addressor's epistemological stance to textual events form a variable cluster which are marked in orange; (2) with the same vector direction and rather similar vector length, subject topic, levels of self-revelation and instruction form the second variable cluster which are marked in light blue; (3) settings of communication and the presence of audience overlap with each other and thus forming the third variable cluster, which is marked in red; (4) lastly, the four strikingly correlated variables mapping the middle-right quarter of the graph, i.e. personal relationship, the number of addressee (s), relative status and shared personal knowledge between the addressor and the addressee (s) have been unquestionably grouped together as the fourth cluster variable. Their distinctiveness from the rest of variables is made salient by the highlighted green colour.

The grouping of different variables suggests that in explaining the distribution of archaic speeches within the two-dimensional space, closely-related variables tend to yield similar interpretations. For example, the two points marked by number 9 which locate to the middle-right of the graph have been grouped together, for as their relative position to the green variable set indicates, these two archaic speeches uttered by the protagonist appear in similar contextual settings that are characterized by (1) the addressees are plural; (2) the relative status between the addressor and the addressees is mixed; (3) the personal relationship between the addressor and the addressees is also mixed including intimate friends such as Don Quijote and Sancho Panza, familiars like Don Quijote and the priest and the barber, or acquaintance like Don Quijote and the Princess or the tortured boy Andrés, or people unknown to the protagonist like members of the Santa Hermandad; (4) the level of shared personal knowledge between the addressor (el Quijote) and his addressees is also mixed.

The recurrence and intensity (number of variables that form a variable cluster or variable set as marked by the same colour) of variable sets, which result from the similar explanatory power that related variables have, somehow indicates the limited efficiency or redundancy of the analytical framework of situational parameters introduced above. Also, as we can see from Diagram IV, it is in the lower-left quarter of the graph where most of blue points marked by 8 or 9 cluster, there is a noticeable lack of vectors charting this area, especially in the downward direction of the Y axis. This actually constitutes the main defect of this statistical experiment proposed with our dataset, for the absence of vectors covering this object-intensive area implies that although we have succeeded in grouping archaic idioms from Subdivision 8 or 9 together and somehow separated them from their counterparts in Subdivision 10, we do not seem to have found an eloquent argument or have discovered any situational parameter that can be used to explain why such a distributional pattern has been formed and suggested by the statistical technique.

Nevertheless, at this stage, one thing that we can make sure of is that as the statistical test clearly shows, certain similarities indeed exist between the use of archaic idioms in Subdivision 8 and in Subdivision 9 in terms of the contextual features within which they have been deployed, and all we need to do is to make up for the insufficiency of the current analytical framework through rereading carefully the texts and going through the iterative process of formulating and testing new hypotheses with the aid of statistical techniques. This may well turn out to be a laborious process; however, the important implication of the proposed research method consists in that by following such an approach to textual or discourse analysis, we are now in a much better position to make well-supported arguments and discard unwarranted presumptions bearing on the nature of style variation within the work of a single author or translator.

Appendix I A modified analytical framework for variation in Don Quijote's archaic speeches

Register **Situational Parameter**	**Feature**	**Value**	**Variable code in CATPCA**
1.1 Topic	Specific subject	Chivalric heroism	1
		Chivalric romance	2
		Non chivalric	3
		Chivalric general	4
2.1 Addressee	2.1.1 Number	Single	1
		Plural	2
	2.1.2 Audience	Present	1
		Absent	2
2.2 Purposes, intents and goals	2.2.1 Persuade	High	1
		Medium	2
		Low	3
	2.2.2 Reveal self	High	1
		Medium	2
		Low	3
	2.2.3 Instruct	High	1
		Medium	2
		Low	3
2.3 Relations between addressor and addressee	Relative status and power of AR & AE	AR > AE	1
		AR < AE	2
		Mixed	3
2.4 Extent of shared knowledge	2.4.1 Shared specific knowledge of topic	High	1
		Medium	2
		Low	3
	2.4.2 Shared personal knowledge	Intimate	1
		Familiar	2
		Acquainted	3
		Unknown	4
		Mixed	5

2.5 Personal relationship	n/a	Formal	1
		Informal	2
		Mixed	3
2.6 Addressor and the textual event	2.6.1 AR's personal evaluation of the textual event	Important	1
		Everyday	2
		Trivial	3
	2.6.2 AR's attitudinal stance towards the textual event	Emotionally involved	1
		Normal attitude	2
		Emotionally removed	3
	2.6.3 AR's epistemological stance towards the textual event	Belief/conviction	1
		Accept fact	2
		Dubious/uncertain	3
		Against	4
3.1 Settings of the place of communication	Coverage	Public	1
		Private	2

References

Oakes M. (1998). *Statistics for Corpus Linguistics*, Edinburgh: Edinburgh University Press.

Biber D. (1994). *Sociolinguistic perspectives on Register*, Oxford University Press.

Hoover D.L. (2003). Multivariate analysis and the study of style variation. *Literary and Linguistic computing*, Oxford University Press, 18: 341-60.

Cervantes M. (1605). *Don Quijote de La Mancha*, Francisco Rico (ed.) Barcelona: Instituto Cervantes.

Liu J. S. (1995). *Don Quijote de La Mancha*, Guang Xi: Li River Publisher.

Yang J. (1979). *Don Quijote de La Mancha*, Beijing: People's Literature Publisher.