

Phraseology in Corpus-Based Translation Studies

A Stylistic Study of Two Contemporary Chinese

Translations of Cervantes's *Don Quijote*

Meng Ji

Thesis Submitted for the Degree of Doctor of Philosophy

Department of Humanities

Imperial College London

2008

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Meng Ji

CID No.: 0045 4789

Date: 21st October, 2008

Table of contents	3
Acknowledgements	7
List of abbreviations	8
List of tables and graphs	9
Abstract	11
Introduction	12

Chapter I Construction of a Parallel Corpus of *Don Quijote* (Part I)

1.1 Corpus Construction in Translation Studies	16
1.2 Selection of ST/TT Corpus Texts	20
1.3 Segmentation of Chinese Texts (SCT).....	24
1.3.1 SCT in Chinese Corpus Linguistics	24
1.3.2 Selection of Chinese Segmentation Tool	27
1.3.3. Assessment of Segmentation Results	29
1.3.3.1 Segmentation Results of Four-Character Expressions	29
1.3.3.2 Segmentation Results of Two/Three-Character Expressions	30
1.3.3.3 Conclusion	37
1.4 Parallel Text Alignment.....	37
1.4.1 Initial Experiments with Automatic Aligner.....	37
1.4.2 The Need for Manual Alignment	40
1.4.3 Assessment of Manual Alignment	41
1.4.3.1 Technical Problems Caused by Different Word Orders	41
1.4.3.2 Technical Problems Caused by Different Syntaxes	42
1.5 Conclusion	45

Chapter II Corpus Data Retrieval and Annotation

2.1 Problem-Oriented Annotation	46
2.2 Automatic Extraction of Four-Character Expressions	49
2.3 A Pilot Classification of Chinese FCEs	53

2.4 Redefining Chinese Idioms.....	55
2.4.1 Idiomatic Variants	56
2.4.2 Idiomaticity and Figurativeness	61
2.5 A Classification of Figurative Idioms and Archaic Idioms	61
2.5.1 Figurative Idioms	61
2.5.2 Archaic Idioms	63
2.6 Conclusion	67

Chapter III Phraseological Patterns in Yang's Translation

3.1 Quantifying Style in Corpus-Based Translation Studies	70
3.2 Sampling	72
3.3 Time Factor in the Use of Idioms in Contemporary Chinese	74
3.4 Yang's Phraseological Preference for SBs, MPs and SSYs	79
3.4.1 Semantically Bipartite Phrases (SBs).....	80
3.4.2 Morphologically Patterned Phrases (MPs).....	82
3.4.3 Structurally Symmetrical Phrases (SSYs).....	85
3.5 Conclusion	87

Chapter IV Phraseological Patterns in Liu's Translation

4.1 Liu's Phraseological Preference for Figurative Idioms	91
4.2 Cognate Metaphors: Value, Obedience, Comparison.....	92
4.3 Non-Cognate Metaphors.....	93
4.4 Other Figurative Speeches in the ST.....	101
4.5 Liu's Phraseological Preference for Archaic Idioms	106
4.6 Conclusion.....	109

Chapter V Use of Figurative/Archaic Idioms in the Two Translations

5.1 Use of Figurative Idioms in the Two Chinese Translations	111
--	-----

5.1.1 Yang's and Liu's Translation of ST as Figurative	111
5.1.2 Yang's and Liu's Translation of ST as Non-Figurative	114
5.1.3 Pattern Comparison	115
5.2 Use of Archaic Idioms in the Two Translations	118
5.2.1 Quantification of Archaisms in Don Quijote's Speeches.....	119
5.2.2 Quantification of Don Quijote's Archaic Speeches in the Two Translations	122
5.2.2.1 Strength of Correlation	122
5.2.2.2 Correlation Statistical Modelling	126
5.3 Conclusion	132

Chapter VI Quantitative Exploration of Style Variation in Liu's Translation (I)

6.1 A Context-Motivated Theory for Style Variation in Literary Translation.....	133
6.2 Statistical Method Applied: CATPCA.....	136
6.3 Variable Reduction and Pattern Recognition (Sub VIII and IX)	139
6.3.1 Some Notes on Manual Data Encoding	139
6.3.2 Dimension Solution	142
6.3.2.1 Two Dimension Solution	143
a) Variance Accounted For	143
b) Reversal of Data Encoding	145
c) Conceptual Scale Construction	147
d) Testing of Scale Reliability	148
6.3.2.2 Three Dimension Solution	153
6.4 Conclusion	157

Chapter VII Testing of Statistical Model for Style Variation (Liu, I)

7.1 Model Testing and Logistic Regression	159
7.2 Two Dimension CATPCA Model for Hypothesis Testing	162
7.3 Data Preparation: Computation of Factor Scores	166
7.3.1 Creation of New Variables	166
7.3.2 Variable Selection Method	167
7.4 Interpretation of Statistical Results	168
7.5 Conclusion.....	171

Chapter VIII Conclusion

8.1 Summary of the Main Purpose of the Thesis	173
8.2 Balancing of Manual and Automatic Textual Analysis	174
8.3 Systematicity and Replicability of the Processing of Textual Information	176
8.4 Construction of Theoretical Models to Elicit the Nature of Literary Translation	180
8.5 How the Methods and Procedures Developed in My Thesis Might Help Future Research	181
References	183
Appendix I Random selection of test sample sets from Liu's work	191
Appendix II Random selection of test sample sets from Yang's work.....	194
Appendix III Translations by Yang and Liu of the source text as figurative	197
Appendix IV A complete situational framework for register studies.....	214
Appendix V Analysis of don Quijote's increased archaic speeches in Subdivision VIII & IX.....	216
Appendix VI Data encoding of Subdivision X (Liu, I)	218
Appendix VII Computation of factor-based scores.....	219

Acknowledgements

The completion of this thesis owes a great deal to the unfailing support and encouragement of many people. I would like to give my heartfelt thanks to my supervisor, Dr. Juan Lalaguna, who is a great expert in the field of translation history and Spanish literature and has provided a unique perspective and invaluable guidance throughout the preparation of this work. His wisdom and encouragement are very much appreciated. I am also grateful for the professional support from colleagues working in the Department of Humanities at Imperial College London, especially my second supervisor Mrs. Nicky Harman, Head of Program Mr. Mark Shuttleworth, Dr. Jorge Díaz-Cintas and Dr. Gabriela Saldanha (now based at the University of Birmingham). All of them have inspired me in one way or another through their interest in my work.

I extend my sincere thanks to the following people who have been extremely generous in sharing their linguistic and statistical expertise with me over the past few years: Dr. Stefan Evert, lecturer of computational and cognitive linguistics at the University of Osnabrück, Germany; Professor Dr. Ahmet Suerdem (quantitative sociology), Department of Business Administration, University of Bilgi, Turkey; Professor John Knapp, Department of English, University of Northern Illinois, US; Dr. Richard Xiao, lecturer of English linguistics, University of Central Lancashire, UK; Prof. Dr. Reinhard Koehler, Department of Computational Linguistics and Mathematics at the University of Trier, Germany; Victor H. Cervantes, Laboratorio de Psicometría, Universidad Nacional de Colombia, Bogotá.

Last but not least, I offer this thesis to my beloved parents Cheng Quan and Ze Rong. Without their unconditional support and unstinting encouragement throughout those years, all of this would not have been possible.

List of Abbreviations

FCEX: Chinese four-character expressions

CMT: Context-motivated theory

CCDQ: Parallel Castilian and Chinese Corpus of *Don Quijote* Part I

POA: Problem-oriented annotation scheme

MP: Morphologically patterned four-character expressions

SS: Syntactically schematic four-character expressions

SSY: Structurally symmetrical four-character expressions

SB: Semantically bipartite four-character expressions

AI: Archaic idioms

I-AI: Instantiated archaic idioms

FIG: Figurative idioms

I-FIG: Instantiated figurative idioms

PTC: parallel translation corpora

CTC: comparable translational corpora

SCT: segmentation of Chinese text

LCMC: Lancaster Corpus of Mandarin Chinese

CLAS: Chinese Lexical Analyzer System

LBS: Lexical segmenter

List of tables and graphs

Table I A brief review of CST's yearly compilation	28
Table II Global statistics of four-character segments in Yang's and Liu's translations	51
Table III Idioms and fixed expressions in the Babel English-Chinese Parallel Corpus	52
Table IV Distribution of FCEX types in Yang's and Liu's work	73
Table V Distribution of idioms in LCMC versus UCLA Chinese corpus	75
Table VI Summary of Pearson's correlation test	76
Table VII Comparison of normalized frequencies between LCMC and UCLA corpus	77
Table VIII Translation of figurative speech in <i>Don Quijote</i> by Yang and Liu	112
Table IX Three 100-sample set for comparative purposes (Liu/Yang)	114
Table X Use of FCEXs by Liu in translating source texts: figurative vs non-figurative	115
Table XI Use of FCEXs by Yang in translating source texts: figurative vs. non-figurative.....	115
Table XII Forms of archaism used in don Quijote's speeches Part I	120
Table XIII Selection of archaic expressions in <i>Don Quijote</i> Part I.....	120
Table XIV Ten subdivisions of <i>Don Quijote</i> Part I	121
Table XV Frequencies of don Quijote's archaic speeches.....	122
Table XVI Frequencies of don Quijote's archaic speeches in translations.....	122
Table XVII Pearson's correlation test for don Quijote's archaic speeches in CCDQ	123
Table XVIII Summary of Pearson's correlation test	125
Table XIX Summary of linear regression models	127
Table XX Summary of adjusted R ² in the linear regression models	128
Table XXI ANOVA of the linear regression models	130
Table XXII Adapted version of Biber's analytical framework for the current study	138
Table XXIII Model summary of two dimension solution (1/2)	143
Table XXIV Component loadings for two dimension solution (1/2)	144
Table XXV Model summary of two dimension solution (2/2)	146
Table XXVI Component loadings for two dimension solution (2/2)	147
Table XXVII Reliability statistics for dimension 1 (Sub VIII and IX).....	152
Table XXVIII Inter-Item correlation matrix for dimension 1 (Sub VIII and IX).....	149
Table XXIX Item-total statistics for dimension 1 (Sub VIII and IX).....	150
Table XXX Reliability statistics for dimension 2 (Sub VIII and IX).....	151
Table XXXI Inter-item correlation matrix for dimension 2 (Sub VIII and IX).....	152
Table XXXII Item-total statistics (Sub VIII and IX).....	152
Table XXXIII Model summary of three dimension solution	153
Table XXXIV Component loadings of the 3D CATPCA model	154
Table XXXV 3D model refinement (1/3)	154
Table XXXVI Component loadings of refined 3D model (1/3)	155
Table XXXVII Model refinement (2/3)	155

Table XXXVIII Component loadings of refined 3D model (2/3)	155
Table XXXIX 3D Model refinement (3/3)	156
Table XL Component loadings of refined 3D model (3/3)	156
Table XLI 2D CATPCA model for stylistic variation in speech archaisms	159
Table XLII Dependent variable encoding (DVE) for logistic regression	162
Table XLIII CATPCA model summary (Subdivisions VIII, IX and X) (Liu, I)	163
Table XLIV Component loadings	164
Table XLV Reliability statistics for dimension 1 (Sub VIII, IX and X)	164
Table XLVI Inter-item correlation matrix for dimension 1 (Sub VIII, IX and X).....	164
Table XLVII Item-total statistics for dimension 1 (Sub VIII, IX and X).....	165
Table XLVIII Reliability statistics for dimension 2 (Sub VIII, IX and X).....	166
Table XLIX Inter-item correlation matrix for dimension 2 (Sub VIII, IX and X).....	166
Table L Item-total statistics for dimension 2 (Sub VIII, IX and X).....	166
Table LI Variables in the equation	170
Diagram I Types of translational corpora	18
Diagram II A classification of Chinese four-character expressions retrieved from CCDQ.....	54
Diagram III Idiomatic variation – extension (1)	58
Diagram IV Idiomatic variation – extension (2)	58
Diagram V Comparison of normalized frequencies between LCMC and UCLA corpus	78
Diagram VI Identification of general contrastive patterns between Yang’s and Liu’s work	80
Diagram VII Translation of figurative speech in <i>Don Quijote</i> by Yang and Liu	112
Diagram VIII Distribution of FCEXs across three sample sets	114
Diagram IX Contrastive patterns of FCEXs in Yang’s translation	116
Diagram X Contrastive patterns of FCEXs in Liu’s translation	116
Diagram XI Contrastive patterns in don Quijote’s archaic speeches	123
Diagram XII Comparison of three hypothetical models	131
Diagram XIII Scree plot for CAPTCA model construction (Subdivision VIII and IX)	142
Diagram XIV Context motivated theory and style variation in literary translation	158
Diagram XV Statistical structural modification from GRV to VSA	160
Formula I Expected frequency calculation in χ^2 test	73
Formula II Chi-square test	73
Formula III Adjusted determination coefficient	128
Formula IV ANOVA of linear regression	128
Formula V Logistic regression equation	167

Abstract

The present work sets out to investigate the stylistic profiles of two modern Chinese versions of Cervantes's *Don Quijote* (I): by Yang Jiang (1978), the first direct translation from Castilian to Chinese, and by Liu Jingsheng (1995), which is one of the most commercially successful versions of the Castilian literary classic. This thesis focuses on a detailed linguistic analysis carried out with the help of the latest textual analytical tools, natural language processing applications and statistical packages. The type of linguistic phenomenon singled out for study is four-character expressions (FCEs), which are a very typical category of Chinese phraseology. The work opens with the creation of a descriptive framework for the annotation of linguistic data extracted from the parallel corpus of *Don Quijote*. Subsequently, the classified and extracted data are put through several statistical tests. The results of these tests prove to be very revealing regarding the different use of FCEs in the two Chinese translations. The computational modelling of the linguistic data would seem to indicate that among other findings, while Liu's use of archaic idioms has followed the general patterns of the original and also of Yang's work in the first half of *Don Quijote* I, noticeable variations begin to emerge in the second half of Liu's more recent version. Such an idiosyncratic use of archaisms by Liu, which may be defined as style shifting or style variation, is then analyzed in quantitative terms through the application of the proposed context-motivated theory (CMT). The results of applying the CMT-derived statistical models show that the detected stylistic variation may well point to the internal consistency of the translator in rendering the second half of Part I of the novel, which reflects his freer, more creative and experimental style of translation. Through the introduction and testing of quantitative research methods adapted from corpus linguistics and textual statistics, this thesis has made a major contribution to methodological innovation in the study of style within the context of corpus-based translation studies.

Introduction

My strong interest in Spanish literature and culture from the very beginning has drawn me naturally to the topic under consideration here, namely the translation of Cervantes's *Don Quijote* into Mandarin Chinese. *Don Quijote*, undoubtedly the best well-known literary work of Spain's Golden Age literature, occupies today a position of pride among translations of foreign literature into Chinese. Despite the contrasts between the two cultures and the length of time that has elapsed, the adventures and misadventures of the Castilian hero have always been well-received by Chinese readers, which is quite a miracle considering China's changing socio-political climate in the course of the last century.

In pursuing this particular topic, my general aim is to help explain *Don Quijote*'s popularity in China. To this effect I intend to approach the subject from a micro-structural perspective; that is to say, through the detailed comparison of two different versions of *Don Quijote* (Part I) separated in time by nearly twenty years: Yang Jiang's translation published in 1978 and Liu Jingsheng's, published in 1995. I should endeavour to underline the evolving nature of the language used in translating *Don Quijote* and the stylistic distinctiveness, if any, of the translators. The representativeness of the two Chinese versions under investigation may be said to relate to the translator's keenness to be seen to be writing within the general socio-cultural background at that moment in time, which in turn would help to explain the wide acceptance generally afforded to Yang's version and indeed, the massive commercial success enjoyed by Liu's translation in recent years.

For the purpose of outlining the stylistic profile of the authors, I concentrate on a particular feature of the language used extensively in the two target texts, i.e. Chinese four-character expressions. They constitute a very typical category of Chinese phraseology, which includes a wide variety of phrasal units, ranging from morpho-syntactically patterned words or phrases to figurative idioms, archaism and variants of idiomatic expressions.

The corpus-driven methodology of analysis, which is still in its early stages of development and promises much through the use of ever more sophisticated corpus techniques, could be used to renovate and improve traditional textual analyses. First of all, this particular approach prioritizes the generation of quantitative textual data over the unwarranted claims made by personal impression. Secondly, the textual data extracted from the parallel corpus will then be used as raw material in the process of ascertaining underlying linguistic patterns in the corpus texts. This has been made possible thanks to the availability of a number of standalone text mining applications, which may help yield essential statistical information from the raw data, such as keyword indexing, collocation patterns, token/type ratio, node word distribution spectrum, mutual information score, etc.

Textual data, after being properly annotated and encoded, may then be used for statistical modelling to explore the relationship between different sets of data, as well as the structure of possible

contextual or co-textual factors that may contribute to the linguistic patterns thus detected in the corpus texts. The exploratory role of statistical pattern modelling in corpus-based literary studies should not be underestimated. Given the fast developing nature of the discipline as a whole, the introduction of statistical methods will greatly ease the formulation and testing of theoretical hypotheses bearing on the nature of literary translation.

As I intend to show in detail, all this statistical information has greatly enhanced the capacity for detecting and identifying certain features of a language; extracting, organizing and classifying data, and analysing and comparing large quantities of translated texts. In this regard, this dissertation may be broadly divided into four main tasks: (1) construction of a parallel Castilian-Chinese corpus of *Don Quijote*; (2) extraction and annotation of raw corpus data; (3) detection of underlying linguistic patterns; (4) computational modelling of co-textual factors to explore the cognitive rationale behind the translator's particular use of language.

I begin in Chapter I with a discussion of various types of translational corpora developed so far and freely available. It has been noted that despite the availability of a number of statistically built parallel corpora, small-scale topic-specific corpora still dominate the mainstream of corpus-based translation studies. They are usually built to a large extent manually and entail a great deal of effort and dedication on the part of corpus builders. This has been the case in the construction of the parallel corpus of *Don Quijote*. To be more specific on the sort of technical challenges encountered in this project, I have focused on two basically related issues: segmentation of Chinese texts and parallel alignment of source and target texts. Inevitably the solutions to the problems encountered in the performance of these two tasks have significantly affected the subsequent design of the whole project.

In Chapter II, I outline a type of problem-oriented annotation scheme, which through an exchange of two major procedures in corpus data mining, i.e. the linguistic annotation of the corpus and the automatic extraction of corpus data, could help keep in balance the developing nature of many corpus annotation tools and the labour-intensive process of manual tagging. The problem-orientated scheme developed here is a corpus-based typological study of Chinese four-character expressions, which prepares the raw data and makes them more amenable to further quantitative exploration of the parallel texts.

Based on the large amount of linguistic data automatically retrieved from the parallel corpus, I proceed in Chapter III to a quantitative analysis of the parallel corpus, which should lead to the identification of three general phraseological patterns separating the two Chinese translations: Yang's relatively higher use of morpho-syntactically patterned four-character expressions (Phraseological Pattern 1, PP-1 hereafter); Liu's preferred use of Chinese figurative language (PP-2), and Liu's relatively higher use of Chinese archaic idioms (PP-3).

The remaining part of Chapter III is devoted to a detailed sample-based study of PP-1. It is

argued that PP-1 signals quite effectively how the translator manipulates language at the morpho-syntactical level, i.e. Yang's initial attempt at assimilating the source text content into the target linguistic system. Chapter IV goes on to examine PP-2 in Liu's work, pointing to a stylistic modification of language at an ideational or conceptual level. Chapter IV also offers a sample-based analysis of PP-3 as identified in Liu's work, suggesting that the use of archaism, as an important rhetorical device, may well also have an impact on the reception of the target texts. Through the comparative study, it becomes clear that in dealing with source text figurative expressions, Liu's approach has been more creative and target-text-oriented. He massively deploys Chinese figurative expressions, which sometimes may only match with source text expressions at a pragmatic or functional rather than semantic level.

I move in Chapter V from qualitative analysis to a rather exploratory computational modelling of textual patterns that have been identified so far. My purpose here is to bring statistical ways and means to further the study of literary translations. Quantitative analysis produces a number of very interesting findings regarding the triangular relationship between the source text and the two target texts. Firstly, while the frequency of use of figurative idioms in Liu's work is generally higher than that found in Yang's work, the enhanced figurative nature of Liu's work has proved to be more responsive to the figurative features in the source text than is the case in Yang's work. Secondly, with respect to archaisms, while Liu's use of archaic idioms may be reasonably predicted by the statistical models built up on the basis of the data retrieved from Yang's work and the original, noticeable differences come to the surface from the second half of Part I of the novel. This translational phenomenon is known as style variation, which brings me to the notion of a context-motivated theory of interpretation that would help in my search to integrate quantitative methods adapted from textual statistics and cognitive linguistics.

In Chapter VI I use a context-motivated theory with a view to exploring the nature of Liu's idiosyncratic use of archaisms in the second half of the novel (Part I). The corroboration of such an interpretative framework, which draws upon Biber's quantitative study of register variation (Biber, 1998), has been put into operation through the use of the statistical procedure known as Categorical Principle Component Analysis (CAPTCA). Despite its wide application in social/behavioural sciences and in quantitative sociolinguistics, this procedure has been rarely used in translation studies. The qualitative analysis performed by CMT aims to explore rather than to test the latent structure of co-textual factors that may help to reveal the cognitive nature of style-shifting in literary translations.

Chapter VII is a continuation of the statistical analysis carried out in Chapter VI. It applies a different type of statistical analysis, i.e. logistic regression, to test the hypothetical model built by using the exploratory CAPTCA procedure. The data used for hypothesis testing are retrieved from the last three subdivisions of Liu's translation. The test results show that the statistical model built in the previous chapter has efficiently pointed to the latent structure of source text contextual factors that may help to explain the stylistic variance detected in the second half of Liu's version of *Don Quijote* (I). The regression model also reveals that within CMT's explanatory framework, the style variation found in

Liu's work may be construed as the translator's freer, creative and more experimental style of writing, which may well reflect in turn the translator's internal self-consistency with himself in an attempt to render a version of *Don Quijote* following in the main his own impression and general understanding of the novel.

In the course of defining a firm line of empirical research in corpus-based translation studies as outlined above, I argue for an interdisciplinary approach to the study of style, which as will be shown in this thesis, will prove to be quite useful to future research. It is believed that an interdisciplinary study of translating style will benefit greatly from the cross-fertilization of research methodologies and insights obtained from different but related disciplines, such as computational stylometry, corpus phraseology, textual statistics, quantitative sociolinguistics and cognitive stylistics.

To be specific, there are two core issues in the thesis showing an overt predisposition towards interdisciplinary treatment. These are, firstly, the expansion of the scope of corpus-based analysis of translation from lexico-grammar to phraseology, which has been made possible thanks to the technical development in the automatic retrieval of multiword expressions; and secondly, the deployment of quantitative analysis. One of the major contributions the present thesis makes to the study of literary translations is the establishment of a three-level quantitative analytical framework, namely raw data generation and classification, pattern modelling and cognitive exploration of the rationale behind the observed linguistic patterns.

Chapter I Construction of a Parallel Corpus of *Don Quijote* (Part I)

For the purpose of analyzing the two selected Chinese versions of *Don Quijote* (Yang, 1978 & Liu, 1995), there were three methodological options open to use. I could have followed a method of interlinear textual analysis, guided by my intuition regarding the subject matter (Allen, 1979: 1-3; Power, 1967); or I could have set up a comparative model to examine linguistic equivalences that may occur between source and target texts (Leuven-Zwart, 1989: 151-181); or I could have adopted a corpus-based quantitative approach which would yield the textual data required to address a range of potential research questions.

I was always attracted by the last option. My own view is that many current approaches to the study of style in translation are somehow methodologically limited and not very productive. The range and scope of current textual analysis seem to have reached a limit, especially in terms of the methods that have been used so far. Is it worth therefore asking whether corpus tools may not be the way forward? These corpus tools do provide the kind of quantitative data that would enable researchers to engage in a convincing manner with soundly based empirical analysis. Such a new perspective is bound to elicit valuable insights into language behaviour in translation, as well as to facilitate a deeper understanding of the linguistic peculiarities and idiosyncrasies that may help define the particular style of a translator. In this regard, a corpus-based study of the two Chinese translations of *Don Quijote* has considerable potential to bring methodological benefit that would help advance and renovate textual studies.

The second reason why I have decided to follow this corpus-based approach to the study of the two Chinese translations is that, should this particular approach of mine prove to be a fruitful and efficient way of analysing translated texts, it would serve as inspiration to others and also be of practical help to any future corpus-oriented research in translation studies. It is my belief that the methodological framework used in corpus linguistic research is highly replicable and largely extensible. Any contribution made in the course of refining and advancing the current use of corpora in doing textual analysis would leave substantial space for the development of future work in this particular field of translation research.

1.1 Corpus construction in Translation Studies

Rapid advances in the construction of language corpora have provided us with a powerful technical platform for the study of literature and language never seen before. In particular, Translation Studies, as a growing academic discipline in its own right, situated at the crossroads of applied linguistics and comparative literature, has been called upon to turn its attention to the use of computerized parallel corpora as a highly efficient way of exploring translation-related activities and textual phenomena (Baker, 1995: 223-243).

Baker's proposal has been a milestone in the development of Translation Studies as an individual academic discipline through its introduction of modern computational techniques and quantitative linguistic material as furnished by language corpora, which will gradually revolutionize the way we observe, analyze and conceptualize human translation as a central cross-linguistic socio-cultural endeavour, as well as the subsequent translation products. To a certain extent, it may be said that the further expansion of the linguistic branch of translation studies could be explored effectively through the use of modern language resources and techniques, and the large amount of material evidence, real language in context, as accumulated in the form of translational corpora.

From initial efforts and discussions on the construction of parallel or comparable corpora of translation (McEnery & Wilson, 1993; Teubert, 1996: 238-264), the compilation and development of electronic resources for translation studies has shown a diverse trend, ranging from small-scale topic-specific corpora to massive statistically-built parallel corpora. Also, as a crucial feature of annotated corpora, the linguistic information added to raw corpus texts has become increasingly sophisticated. This in turn has prepared the ground for further expansion of the scope of research based on the generation and description of textual data from electronic translational corpora.

In this section, I shall first be looking retrospectively at the various types of translational corpora (see Diagram I) already developed or under development, with a view to assessing the current situation as regards corpus-based translation studies, as well as to pointing to possible directions for future research in the field based on ongoing parallel/multilingual corpus engineering and natural language processing.

Diagram I exhibits different types of translational corpora in the field. Some of them have been made available for some time, while others are still under development at this moment. As we can see, the scale and diversity achieved so far in translational corpus engineering has been quite conspicuous for a relatively short period of time; especially when we compare it to the construction of large-scale monolingual corpora, e.g. in English (e.g. the Brown Corpus of American English, 1961), which was initiated more than three decades earlier than the former.¹ The advances made in parallel corpus construction would seem to be even more prominent, if we take into account the much higher levels of difficulty implied in solving technical problems relating to parallel text matching and alignment, especially working with typologically different languages (Piao, 2002: 207-30).

At the highest rank of category shown in Diagram I, we have two major types of translational corpora: parallel translational corpora (PTC) and comparable translational corpora (CTC). There has been some confusion in the literature regarding the establishment of a consistent terminological framework for corpus type categorization (Baker, 1999: 281-98). In my view, the differences between the two may be better described and understood by looking at their underlying structural features: while

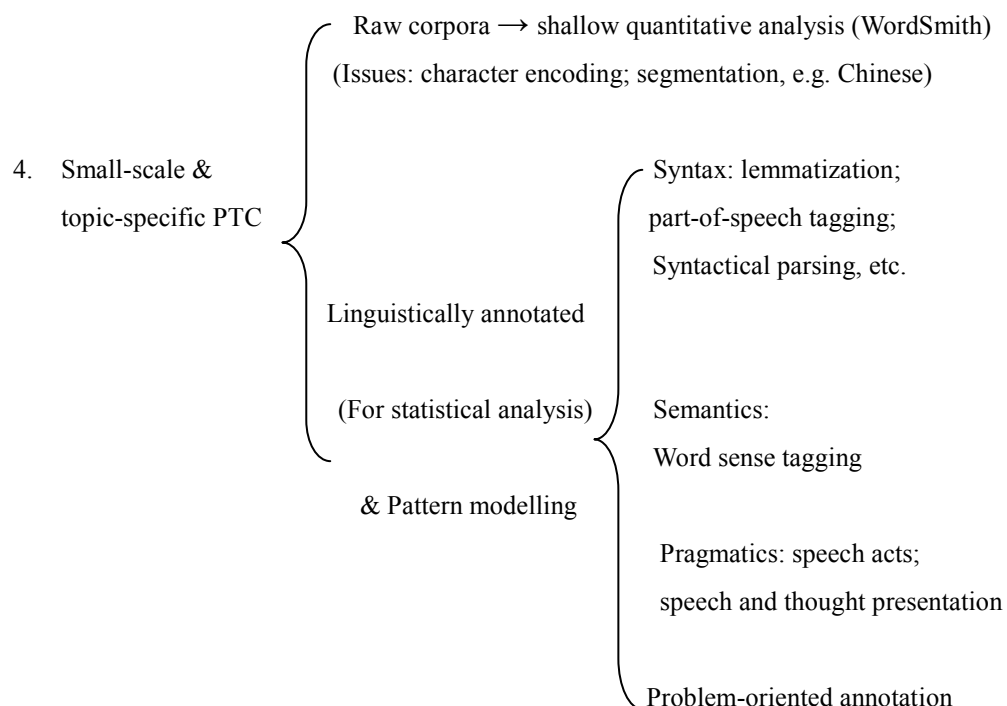
¹ One of the earliest and well-known parallel corpora was the English-Norwegian Parallel Corpus (ENPC), which was built in the early 1990s.

a PTC contains both the source and target texts, a CTC is a compilation of translated texts only, with a view to investigating the nature and regularities of translated language (Baker, 1995: 223-43).

Diagram I Types of translational corpora

A: Parallel (Multilingual) Translational Corpora (Both ST and TTs) (PTCs)

1. Large-scale balanced PTC
2. (Large-scale) genre-specific PTC, e.g. newspapers, legal documentation
3. Large-scale non-balanced and non-classified PTC



B: Comparable Translational Corpora (TTs only) (CTCs), e.g. TEC²

1. Large-scale balanced CTC, comparable to large-scale monolingual corpora, etc.
2. Genre-specific CTC, like TEC

Within each category, PTC or CTC can be further classified according to the text types that they may cover, whether they be large-scale balanced corpora or genre-specific corpora. The purpose of building large-scale balanced corpora is to investigate general linguistic features of the language in use; whereas the compilation of genre-specific translational corpora aims primarily to address research

² Translational English Corpora (TEC) is constructed at Centre of Translation and Intercultural Studies, University of Manchester. It may be accessed at <http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm>

questions regarding specific aspects of translated language or within certain text domains. In theoretical terms, both types of subcorpora may be equally explored in the construction of a translational corpus platform; however, experience shows that unlike the construction of monolingual corpora (one language only), large-scale national corpora like the British National Corpus, the Modern Chinese Language Corpus, the American National Corpus, etc., the development of large-scale balanced corpora, which are populated by translated texts only, is both size-limited and much less balanced.

It is not difficult to understand why this should happen: while native-speakers' use of language is in abundant supply in every aspect of social life, translated language may be found only in a rather limited range of communicative environments, such as language teaching and learning, conference interpretation, and mainly, published translations. The obvious shortage of translation material makes the establishment of a balanced framework of sampling notably difficult, which in turn has restricted the scope of contrastive studies, mainly between translated and non-translated languages, that current versions of PTC and CTC may be able to cover. In fact, most PTCs and CTCs developed so far are genre-specific corpora, which favour a restricted range of text genres that may be easily collected from published translations, e.g. legal texts, official documents, fictions, commercial fliers or technical manuals, etc.

Automatic construction of parallel/multilingual translational corpora usually approaches technical issues quite differently from the way it is done in the compilation of a comparable translational corpus. In building up a CTC, the establishment of a well balanced sampling framework may be crucial; without an adequate text selection scheme, the comparison between corpora in translated languages and original languages – in this case, translated language is seen as a legitimate language type in its own right, which is precisely the rationale behind the construction of TEC – will be seriously compromised. On the other hand, for the construction of a PTC, the first technical problem is a proper automatic alignment of source and target texts at word or sentential level, which represents one of the most difficult problems in natural language processing and machine translation (Véronis, 2000:25).

Automatic parallel corpus construction, which started with Gale and Church's famous sentence-length principle (Gale and Church, 1991: 177-84), has evolved rapidly into a rich theoretical paradigm, which attempts to integrate both empirical linguistic cues, such as similar sentence length or cognate words contained in the source text or its corresponding target text segment, and statistical alignment techniques, e.g. sentence-based dynamic programming (Gale and Church, 1991: 177-84; Brown et al. 1991: 169-76) and language-independent fuzzy chunk matching (Deng et al. 2006: 1-26). The continuous and collaborative efforts made by computational linguists have greatly improved the efficiency of automatic parallel text alignment systems, as well as facilitating the construction of a number of domain-specific parallel corpora, prominently in European languages and other well explored language pairs, such as English-Chinese or English-Japanese.

However, the enormity of costs and copyright issues involved in the development of such parallel corpora often prevents them from becoming freely accessible to translation scholars or research students. This in turn has given rise to the type of parallel corpora listed last in Diagram I, i.e. small-scale topic-specific PTC, which in fact constitute the mainstream corpora used today in most individual research projects in line with corpus-based translation studies.

As the most frequently used type of corpora in translation research, small-scale topic-specific parallel corpora are usually built to a large extent manually with specific questions born in the researcher's mind. To start with, they normally require a great deal of effort and dedication on the part of the researcher to prepare the textual material especially selected for the purpose, either inspired by certain methodological orientations (for normalization see Kenny, 2001; Munday, 1998: 542-556; for explication, see Pápai, 2004: 143-65; Puurtinen, 2004: 165-76; for simplification, see Laviosa, 1997b: 531-540), or by adopting an openly corpus-driven approach to the particular ST/TT or TT/TT pairs under investigation (Saldanha, 2005). Due to the limited size of this kind of corpora, the specific questions that the researcher may be seeking to answer are exploratory in nature.

Thus, the findings obtained as an end-result of each individual project are largely circumscribed and are usually domain- or author- specific. Nevertheless, the usefulness of this type of DIY corpus, when studied in conjunction with larger-scale comparable corpora, translational or non-translational, may be maximally extended; and the findings revealed in those tailor-made corpora may also be further contrasted with the data collected from more general referential corpora.

1.2 Selection of ST/TT Corpus Texts

China's interest in *Don Quijote* would appear to have been rekindled at the beginning of the twentieth-century as part of the introduction of modern literature from Russia and Japan (Chen, 2002: 123). Ever since then, there have been a number of Chinese translations of this seventeenth-century Castilian masterpiece: first, there was a translation into Classical Chinese in 1922, followed by five versions into vernacular Chinese in the thirties and five more in the fifties. An important fact about these early Chinese versions of *Don Quijote* is that they are invariably relay translations. That is, instead of translating directly from Castilian, the translators have resorted to their English or French versions first and then translated them into Chinese. This may have been due, it is suggested, to the lack of professional expertise in this specific language pair at the time.

All the same, the huge contrast between the two languages has never diminished China's enthusiasm and interest in this early seventeenth-century Castilian novel. In fact, despite the changing political climate in modern Chinese history, *Don Quijote* has managed to survive quite well, and the prestigious position that it now occupies in translated western literature in China has never been jeopardised either by earlier democratic movements or the later communist regime. This may well have been due to the chivalric theme and ideals of the novel which are somehow open to a variety of

interpretations, e.g. literary, socio-cultural or political. The literary critic Zhang observes that *Don Quijote* has never been studied in China as a comedy, a good story seeking mainly to entertain, but rather as a serious political tract intended to bring back a former and susceptible character of different interpretations under varying contextual circumstances in modern China (Zhang, 1987: 43-50).

The first direct translation from Spanish into Chinese did not come until 1978, a crucial year in the history of modern China when the country was in the middle of wide economic reforms shortly after the notorious Great Cultural Revolution (1966-1969). Yang was commissioned by the Ministry of Culture and Education of the newly established People's Republic to produce the first direct version of *Don Quijote* into Mandarin Chinese. This was supposed to be another attempt to re-interpret western literature within the socio-political agenda of the communist regime. Yang was entrusted with this task because of her high profile in Chinese literary and translation circles at the time. Now in her mid-nineties, Prof. Yang still holds a very high position in the world of letters. Prof. Yang was educated in one of the most prestigious universities in China. Her early days of overseas studies cover a short period of stay at Oxford University and scholarly visits to the University of Paris. She is very fluent in English and French and has translated a number of western literary works into Chinese, including *Gil Blas* by the French author Alain-René Lesage. She admitted that her knowledge of Spanish was quite limited when she first embarked upon the ambitious project of translating *Don Quijote* in the mid 1970s (Hu, 2004: 41-55; Zou, 2006: 17-25).

She was then in her early fifties and quite obviously determined to do her best because she started to learn Spanish at great speed. This determination is much to her credit, but it has also caused some concern as to the originality of her work. Whatever the case, Yang's translation of *Don Quijote* is still widely acknowledged to be, at least officially, the first direct version of *Don Quijote* into Mandarin Chinese. The great authority enjoyed by Yang's translation is reflected in the fact that it is the most published Chinese version of the novel today in China. It has also been collected into textbooks for Chinese secondary education as the standard version of the novel. There is therefore no doubt that Yang's translation has been a landmark in the story of *Don Quijote* in China, especially in the way later generations of Chinese translators, and the wider Chinese public, in general, look upon this literary figure with a great deal of interest and respect.

In more recent times, from the early 1990s onwards, the translation of Cervantes's masterpiece into Chinese has experienced another burst of popularity, far more conspicuous and fruitful than ever before. The seven Chinese versions of *Don Quijote* published during this period (1990, three in 1995, 1996, 2000 and 2001) have been largely characterized by the reputed scholarship of the translators and the competitive urge to produce the best quality language in the target texts. This may be partly explained by a better acquaintance with the source text on the part of recent Chinese translators, most of whom are professional translators, university or college professors, and partly by an implicit desire to foster to accommodate to the popular taste of a potentially vast Chinese readership.

The author of one of the three 1995 versions of *Don Quijote*, which I have selected for my study, is Liu Jingsheng (1956-). I have opted for his 1995 version because of the following considerations:³ (1) among his contemporaries Liu is probably one of the most prolific Hispanic-Chinese translators today in China; (2) he has followed a path typical of professional literary translators, in that he was educated with Spanish as his first foreign language ever since he was nine; and (3) his contacts with foreign literature have been primarily via Hispanic novels, while his knowledge of other major European language like English or French is minimal. Liu is a good example of the new generation of Chinese translators who are both highly professional and well aware of the demands from the publishing market of foreign literature nowadays in the country.

When he published his two-volume version of *Don Quijote* in 1995, he was barely forty, probably the youngest translator ever of *Don Quijote* into Mandarin Chinese; a fact, as he considers when looking back at his career, which not only shows the great confidence that he had in translating between the two languages, but also reflects the changing circumstances surrounding China's publishing market in the 1990s. More and more commercial publishers started to turn their attention to professional translators of younger generations, whose language styles would tend to be more idiomatic than those preferred by established academic scholars. An easy and fluent style was deemed to be crucial for a lengthy literary classic like *Don Quijote* to be made more accessible to a new generation of Chinese readers, whose literary taste and expectation somehow holds the key to the successful story of a literary translator today in China.

Liu seems to have been aware of the wind of change in China's publishing market. When working on *Don Quijote*, he has largely followed the same language style that has proved to be quite successful in the translation of other types of more popular genres. Liu's bold attempts at injecting more idiomatic elements into the target language have provoked heated debates in several nation-wide periodicals from the perspective of quality assessment. Despite the fierce criticism that Liu's version has received from older generations of Chinese literary translators, his work has obviously stood the test of time for his version has been reprinted several times in the last ten years or so.

With a gap of some twenty years in between, Yang's and Liu's translations are not totally unrelated. The high standing of Yang's version over the years would explain its influence upon any future Chinese version of the novel for years to come. This has been the case with Liu's more recent version of *Don Quijote*. In our interview, Liu admitted quite openly that before starting his own translation, he had read Yang's work throughout and was therefore quite familiar with her writing style. Liu praised Yang's work as faithful and elegant, but at the same time he believed it was lacking in what he saw as the most important factor for a literary translation to be successful in China's new circumstances. To be popular with the general reader, Liu argued, the new style of writing would have to be more idiomatic; that is to say, sound like more ordinary, generally accessible Chinese. He said he intended to put this right in his new version of the novel.

³ Data collected in an interview with Liu in April 2007.

The main purpose of this thesis is thus to compare the two Chinese translations discussed above, with a view to investigating their different styles of writing and, perhaps, to raising questions as to the extent to which the two different Chinese versions of the Castilian novel may represent different stages in the stylistic development of the language. It may be said that the style of translating, as the style of writing in general, can be largely subjective, personal. But all the same, it must surely relate in a very real sense to the particular circumstances of time and place. In this regard, I suggest that the special relevance of these two translators must lie precisely in their keenness to fit their work within the socio-cultural context at the time, which in turn must help to explain their success at different points in the history of *Don Quijote* in modern China.

The Spanish version of the source text used in this thesis is Francisco Rico's 1998 edition of the novel. It should be noted that Rico's edition differs from the editions used by Yang and Liu. The version of *Don Quijote* used by Yang was edited by Francisco Marín (6th edition) and published in the series of *Clásicos Castellanos* (Madrid, 1952). In the preface to her translation, Yang details the reasons why she finally decided to use Marín's edition. She found the edition by Juan Avallé-Arce (Madrid, 1977). She compared this new edition with Marín's and found that the latter, apart from being better documented, was more careful with an important detail in Chapter XXIII (Part I), i.e. the loss of Sancho's donkey. While the new edition by Avallé-Arce largely followed the first print of the novel in 1605, Marín's version was based on a consideration of both the first and second prints (in the same year), where in the latter the author made certain changes to Sancho's loss of donkey, which thus made sense of the author's self-criticism in the second volume of the novel. Marín added the change that appeared in the second print of 1605 in his own edition of the novel and pointed out that it was a mistake made by the author. Yang believed Marín's treatment was a more sensible way of dealing with this detail and hence decided to use Marín's edition for the purpose of her own translation.

The version of *Don Quijote* used by Liu was edited by Evaristo García (La Habana, 1974) (3rd edition). Liu recalled in our interview that it was the only version of *Don Quijote* available to him at the time when he was working on this novel. It was donated by the Cuban Embassy in Beijing and it is now kept in the National Library of China, Beijing. The selection of Rico's edition for inclusion in the parallel corpus was initially based on rather practical considerations. Apart from being an updated version of the novel officially promoted by the Spanish authorities, Rico's edition has been neatly digitalized and made freely-accessible on the official website of the Cervantes Institute.⁴ This has therefore saved me a great amount of labour in scanning printed books into the computer and other legal issues regarding copyright. Most importantly, as the officially promoted edition of the novel, Rico's editorial work has paid special attention to the differences between important previous versions of the novel which mainly focus on the treatment of the loss of Sancho's donkey, while other variations among the different copies of *Don Quijote* have been seen as trivial and minor by Rico. The ultimate purpose of Rico's edition is to provide a *clear* text which is exempt from any detail that may cause

⁴ The digitalized version edited by Rico may be accessed at <http://cvc.cervantes.es/obref/quijote/default.htm>

confusion and contradiction between his work and previous ones: “De ahí que antes de arriesgar una solución nuestra ... hayamos intentado compulsar en todas las impresiones del siglo XVII y en las más importantes de las posteriores la totalidad de los lugares que nos parecían dudosos”⁵ (Therefore, to avoid making any risky solution ... we have collated all doubtful places in those imprints produced in the seventeenth-century and in those most important imprints afterwards). In this sense, despite any trivial differences in typography between the two versions used by Yang and Liu, the Spanish text selected for inclusion in CCDQ has provided a standard version of the source text for the purpose of my research.⁶

1.3 Segmentation of Chinese texts

1.3.1 Segmentation in Chinese corpus linguistics

Having selected the source and target texts to be used in the parallel corpus construction, we now come to the first serious task in processing the corpus texts which has turned out to be the most challenging part in the whole procedure. I started the construction of the parallel Castilian-Chinese corpus of *Don Quijote* (hereafter CCDQ), with not much reference material which I could draw on to build a framework adequate for the compilation of corpora in this particular language pair. General guidelines do exist but very few of them have touched upon the particular technical problems that the construction of a parallel corpus in Chinese and Spanish may present. As a result, most of the linguistic and technical hitches that I have encountered were to a certain extent unexpected. The construction of CCDQ has been an arduous task, entailing the identification of problems and solving them as I went along.

In the construction of corpora in alphabetical languages, such as the branches of the large Indo-European family tree, the aligning process may be relatively straightforward. But this is not obviously the case in Mandarin Chinese, for the processing of Chinese natural texts has its own, unique problems and therefore requires appropriate techniques. The first major technical problem one encounters is the segmentation of Chinese natural texts. Due to the simple fact that Chinese texts do not carry word delimiters, the first task when processing a Chinese text is to tokenize or to segment text strings into Chinese words. As strange as it might seem to people familiar with the construction of corpora in alphabetical languages, this is not that simple. Chinese texts are far more complex than would appear to be the case at first (Wu and White, 1990: 681-92).

A Chinese sentence may often correspond to several alternative sequences of authentic words; therefore, a crucial problem in Chinese segmentation is how to choose the correct solution among all the possible ones. The following segmented text from Liu's translation of *Don Quijote* is a good

⁵ See http://cvc.cervantes.es/obref/quijote/introduccion/presente_edicion/txtcritico.htm

⁶ In the case of substantial differences between the two Chinese translations which might be caused by the different source texts, I would have gone back to the original, but until the completion of the thesis, this has not happened. On the contrary, major differences detected between the two target texts largely remain at the stylistic level.

example of the ambiguities that may occur when retrieving natural words from inputted Chinese text strings.

CS (I, I: 37-38) En un lugar de la Mancha¹, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero², adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos³ los sábados, lentejas los viernes, algún palomino⁴ de añadidura los domingos, consumían las tres partes de su hacienda. El resto della concluían sayo de velarte, calzas de velludo⁵ para las fiestas, con sus pantuflos de lo mismo, y los días de entresemana se honraba con su vellorio⁶ de lo más fino.

EN⁷ (I, I) In a village of La Mancha, the name of which I have no desire to call to mind, there lived not long since one of those gentlemen that keep a lance in the lance-rack, an old buckler, a lean hack, and a greyhound for coursing. An olla of rather more beef than mutton, a salad on most nights, scraps on Saturdays, lentils on Fridays, and a pigeon or so extra on Sundays, made away with three-quarters of his income. The rest of it went in a doublet of fine cloth and velvet breeches and shoes to match for holidays, while on week-days he made a brave figure in his best homespun.

CH (Liu, I, I: 13) 曼查¹ (La Mancha) 有个地方, 地名就不用提了, 不久前住着一位贵族他那类贵族, 矛架² (lance-rack)上有一支长矛. 还有一面皮盾、一匹瘦马和一只猎兔狗. 锅里牛肉比羊肉多, 晚餐常吃凉拌肉丁, 星期六吃脂肪³ (animal fat) / 煎鸡蛋, 星期五吃扁豆, 星期日加一只野雏鸽⁴ (pigeon), 这就用去了他四分之三的收入; 其余的钱买了节日穿的黑呢外套、长毛绒袜子和平底鞋⁵ (low-heeled shoes), 而平时, 他总是得意洋洋地穿着上好的棕色粗呢衣⁶ (wool coat).

The Chinese characters underlined are marked by the CST⁸, for they are recognized by the CST as dictionary entries. Words in italics are legitimate words in the context that have been mis-segmented by the CST and extracted manually. A forward slash is placed between the two characters that have been left out by the CST as a single character word. A number of segments are totally mismatched, which has consequently reduced the reliability of the tokens segmented as a whole. This is because when the text strings are processed by the CST, they are treated equally as dictionary entries that have been collected by the application so far.

However, this has obviously reduced flexibility in the real use and formation of Chinese words, whereas a single character is always looking for its most suitable “companion” to form a legitimate word in the context. For example, the text string 和平底鞋 can be either segmented as 和平 ‘peace’ / 底鞋 (meaningless) or as 和 ‘and’ / 平底鞋 ‘low-heeled shoes’. In any computational segmentation tool that is programmed on a maximum matching algorithm (Wong and Chan, 1996: 200-3), the former choice of segments tends to be singled out, just as with the results produced by CST; however, this is an obvious mismatch, since in the case of the former, the second segment 底鞋 is simply meaningless.

The reason for the co-existence of several combinational possibilities of characters to form

⁷ The English version used is by J. Omsby (1885) (the Project Gutenberg) <http://www.gutenberg.org/etext/996>.

Back translations have been added for some incorrectly segmented phrases as an illustration of what can go wrong.

⁸ An explanation of how the CST (Chinese Segmentation Tool) works will be supplied shortly.

words within a sentence would need to be explained in terms of the peculiar mechanism of Chinese morphology. A basic concept which needs to be clarified first is that a Chinese character is essentially different from an alphabetical word as a morpho-syntactic token. In fact, a Chinese orthographic symbol is much closer to a morpheme and there are several morphological patterns in which Chinese characters can be combined together to form legitimate words, e.g. juxtaposition of semantically analogous characters like 家室 ‘home-house’, 服从 ‘obey-follow’, 整齐 ‘ordered-trimmed’, 讴歌 ‘eulogize-praise’, 危险 ‘danger-risk’, 粮食 ‘cereal-food’, etc.; combination of a verb and its direct object like 司马 ‘administrate-horses’, 将军 ‘lead-troops’, 继室 ‘re-marry-wife’, 告老 ‘report-retirement’, 食言 ‘break-promises’, etc.; combination of verb and its modifier like 烧死 ‘burn-dead’, 杀害 ‘murder-dead’, 剪断 ‘cut-apart’, 获得 ‘obtain-gained’, 延长 ‘extend-longer’, 填满 ‘fill-full’, etc. Although these words can be easily identified when treated in isolation, it may turn out to be a tricky task to tell them apart in a non-segmented text string.

For example, in segmenting the text string 延长合同期限 ‘extend the period of contract’, in theory, there are possible four Chinese words that can be extracted from the string, which are 延长 ‘extend’, 合同 ‘contract’, 同期 ‘same period’, 期限 ‘time limit’. However, by relocating them in context, we will find that 同期 is not a legitimate word in the relevant context as the other three are; and its extraction from the text string would lead to an irreversible mis-segmentation of the phrase as a whole. In this sense, we can say that the reading of Chinese texts implies human segmentation which requires the reader to go back to the context looking for valid textual information. As can be expected, such a process can hardly be conducted at a similar level by any current segmentation application. In a sense, mis-segmentation is a constant factor in Chinese natural language processing, at least for the time being⁹.

Another key factor that may adversely affect the results of Chinese text segmentation is linguistic creativity. It is arguably true that when it comes to creative writings, such as translating, translators may wish to familiarize the target readership with the original text by normalizing the peculiar linguistic expressions of the source language. The very nature of translating as a creative activity, as well as the topics it touches upon, i.e. foreign names or socio-cultural issues unfamiliar to the target readership, implies substantial technical difficulty for an automatic Chinese segmentation application, since there may well be a good number of words created by the Chinese translator for lack of an appropriate expression in the target language.

In fact, this is what has happened with the segmentation of the two selected Chinese versions of *Don Quijote* by using the CST: the Chinese transliterated, literal or free translations of the many culture-loaded words in Cervantes’s masterpiece seem to have overwhelmed the CST, leading to a number of mismatched tokens. For instance, the word *aljamiado* is translated as 西班牙/文 ‘Spanish’ by

⁹ In some journal articles on Chinese segmentation technology (e.g. Feng, 2001) very high precision rates of above 95% have been reported, but it should be noted that the efficiency of those segmenters largely depends on the text genres selected for experimentation; the reported high precision rates may be seriously compromised when it comes to the segmentation of open textual sources with varying linguistic complexities.

Liu, and as 阿/尔/哈/米/亚/文 ‘Al-ja-miado’ by Yang, which is a phonetic translation. A dash is placed within each Chinese translation, where a missegment is produced by CST through not having recognized the word as a dictionary entry collected by the application so far. I am aware that Liu’s translation of this particular term may be rather contentious or problematic, but from a technical point of view, his translation has been better recognized and segmented by the online annotation tool, when compared with Yang’s translation, which has been treated as an unknown named entity and consequently wrongly segmented.

1.3.2 Selection of Chinese segmentation tool

Generally speaking, there are currently two main types of segmenting systems in use, specifically developed to deal with Chinese natural texts: the statistical Markov model and the lexical or dictionary-based model. First of all, before a critical evaluation of the results produced by CST is attempted, a brief comparison will be made between these two segmentation systems as to their respective advantages and shortcomings when dealing with different types of Chinese texts.

There is, first, the Markov model. This is based on the assumption that the probability of two or more random Chinese characters appearing together can be predicted statistically from previous appearances of each character in a large corpus (Fu, et al. 2005: 590-4). Statistical models may be easily set up and adapted to any domain by being provided with sufficient training data. But it is required that in order to set up an adequate language model, the design of a training corpus for the purpose of probabilistic analysis should take into consideration all factors that may affect the authenticity of data being used, i.e. the representativeness of different text genres as components of the training corpus. This is because any single language in real life differs considerably from one text type to another.

Furthermore, a statistical model is not completely stable in the sense that probabilities need to be revised constantly. Some technical problems have been reported by users of Chinese segmenters based on the Markov model. For instance, in segmenting and annotating the Lancaster Corpus of Mandarin Chinese (LCMC), researchers from the Linguistic Centre at Lancaster University have mainly employed the Chinese Lexical Analyzer System (CLAS), a powerful segmentation tool developed by the Institute of Computing Technology of China along the lines of the statistical Markov model. In spite of the high rate of precision and recall achieved by CLAS in annotating the bulk of LCMC, it has encountered substantial difficulties in distinguishing the different syntactic functions or semantic meanings of homographic words appearing in the traditional Chinese literary genre of martial arts (McEnery and Xiao, 2003: 361-78).

Secondly, there is the lexical or dictionary-based approach to Chinese text segmentation. Similar to the operation of a probabilistic tokenizer, which requires a large-scale training corpus to build an adequate language model, the successful performance of a lexical segmenter largely depends

on the availability of an electronic database with extensive and updated dictionary entries in Chinese. The idea behind the functioning of a lexical segmenter is quite straightforward: since the word-boundary-recognizing capacity of a lexical segmenter may be enhanced through the addition of new lexical items, the more words or fixed expressions the electronic database stores, the better the segmenting result will be.

In practice, though, to make automatic segmenters more efficient, program developers tend to integrate the two basic systems in order to optimize the performance of natural language processing applications. This is the case with the CST, which I have selected for the construction of a parallel corpus of *Don Quixote*¹⁰. CST features the two fundamental properties which a hybrid segmenter must have: statistics-driven and lexis-supplemented. Firstly, the CST has been developed on the basis of Paul Denisowski's Chinese-English Dictionary, which is an open-ending online dictionary continuously expanding.

Besides being freely available, CST is also very useful in that it encourages constructive interaction between its developers and its users by providing an automatic submission page for anyone interested in contributing new entries to the dictionary. This has been designed to broaden the scope for the inclusion of additional material by allowing a potentially huge number of users to make their own contributions to CST, which in turn will adjust the dynamic lexical structure of CST to meet the needs of its users and facilitate the process of segmentation. A short list is given below which shows the entries progressively added to CST (Table I).

Table I¹¹ A brief review of CST's yearly compilation of words¹²

DEC 1997	APRIL-JUNE 1998	FEB-MAY 2003	OCT-DEC 2005
5, 091	35, 045	77, 137	93, 746

Another characteristic of the functioning of CST as a working model is that it has been programmed to segment the Chinese text strings on a maximum matching basis, always looking for the longest combination of characters possible. Such a segmenting principle draws on the hypothesis that longer segments tend to be more specific and thus can possibly better fit into a larger sentential context. In many cases, it is rather an effective way of dealing with long sentences; but at the same time, it is achieved at the expense of smaller segments which represent higher recall rates.

For example, this maximum matching approach implies an inherent contradiction. Given the privileged position that a larger character combination enjoys for being separated from a text string, it is likely that it may “impose” itself over a legitimate but relatively shorter token and cause mis-segmentation, e.g. in 他日常吃的砂锅杂烩里 ‘in the earthenware cooking pot that he uses daily’, both 他

¹⁰ Chinese segmentation tool is available at <http://www-rohan.sdsu.edu/~chinese/annotate.html>.

¹¹ See <http://www.mandarin tools.com/cedict.html> for a more detailed description of the compilation of CST. In Table I, each column stands for the Chinese words collected into the CST during that specific period of time.

¹² The latest statistics provided in the relevant website date back to 2005.

‘he’, 他日 ‘another day’, 常吃 ‘eat frequently’ exist in the dictionary, therefore, based on the maximal matching strategy, the segmentation result produced by CST is

他日 / 常吃 / 的 / 砂锅 / 杂烩 / 里¹³.
 Another day / eat frequently / possessive adjective, of / earthenware cooking pot / stew / inside
 (N.B. the Chinese translation of “Una olla de algo más vaca que carnero, salpicón las más noches”)

While the correct segmentation should be

他 / 日常 / 吃 / 的 / 砂锅 / 杂烩 / 里
 He / daily / eat / possessive ad., of / earthenware cooking pot / stew / inside

Here, the mismatch of 他 ‘he’ with 日 ‘day’ creates a false token 他日 ‘the other day’ which actually does not exist in the text and prevents the legitimate word 日常 ‘daily’ from being extracted correctly from the string. This kind of mismatching should not be overlooked; the accumulation of these false tokens would undermine the validity of the data generated as a whole. However, the example given above just reflects one aspect of the application’s under-specificity and more types of mismatching caused by the current version of the CST came to the surface as the segmentation process unfolded.¹⁴

1.3.3 Assessment of segmentation results

In order to give a systematic and fully-illustrated description of the CST’s performance in segmenting the Chinese texts which I have put together for the construction of CCDQ, I shall first present the results when searching for segments made up of four characters by using Michael Barlow’s Parallel Corpus Concordancer (ParaConc). These four-character segments normally represent the highest ranking combination possible in the Chinese lexicon. Then, I shall look at two-character words, which constitute the bulk of the segmentation pool.

1.3.3.1 Segmentation results of four-character expressions

In view of the large quantity of four-character items obtained by using the regular expression search of Barlow’s parallel corpus concordancer, only fifty items will be selected at random to demonstrate CST’s performance in extracting four-character tokens (See Appendix I). The samples exhibited in Appendix I represent some one per cent of the total four-character expressions extracted¹⁵, but are already enough to demonstrate the high precision with which CST has segmented the Chinese texts, since not even a single mis-segmentation has been spotted here. The reason for the successful performance of CST in extracting four-character items may be due to the fact that these four-character

¹³ Liu, *Don Quijote*, Part I, p. 13

¹⁴ The period in which CST was used to segment Chinese corpus texts of CCDQ (I) was from March to May, 2006

¹⁵ The statistics of four-character segments extracted from the two Chinese translations are reported in Table II, Chapter II

words or phrases are in most cases idiomatic expressions long conventionalized in Chinese lexicon; their component characters, furthermore, show a much lower instability in combining with other characters to form a new word in the text string.

As can be observed from the examples listed in Appendix I, another syntactical feature that may also account for the satisfactory performance of CST in retrieving four-character items is, that most of these four-character expressions are delimited by a preceding or a following punctuation mark, or separated from other words by an adjective or, the adverbial auxiliary words 的 or 地. The four-character tokens identified and extracted by CST have proved to be quite reliable and may thus be used as valid corpus data in future textual analysis.

1.3.3.2 Segmentation results of two/ three-character expressions

Despite the relatively shorter length of Chinese two-character words, their modes of morphological formation tend to be much more diverse and flexible when compared with Chinese four-character items, which are mostly fixed or idiomatic expressions. Therefore, in an attempt to identify the main cases of missegmentation made by CST and examine the authenticity of the tokens generated as a whole, I will be paying more attention to those mismatched two-character words spotted so far.

The results of attempts to classify those falsely segmented tokens in a systematic way can hardly be exhaustive, especially when taking into account the large quantities of texts being processed by the CST; however, one thing for sure is that the summary statistics, though crude, will help indicate what kinds of mismatching tend to occur in the target texts. Persistent false tokens will be thus detected and eliminated, which in turn will contribute to the enhancement of the validity of segmented tokens that can therefore be preserved as potential data for further textual analysis. That is, the results of segment verification are important in that they would help provide guidelines for the direction which further textual analysis may take on the basis of the linguistic features of validated corpus data.

Due to the fact that mis-segmentation may occur with a whole range of Chinese words, e.g. noun, adjective, adverb, verb, etc., in order to establish a working classification of mismatched two-character segments, I shall proceed to the categorization of Chinese word formations from a morphological perspective (Zhou and Duan, 1994: 85-7). The advantage of classifying Chinese two-character tokens from a morphological perspective consists in that first, it will help simplify an otherwise very complicated classifying practice; and secondly, the analytical framework detailed below will allow us to form some idea of how Chinese natural language is processed by computational applications, as well as elucidating the technical under-specificities featured in these corpus tools at the moment. This, I suggest, will be of great practical help later on in my research, in terms of identifying valid and potentially interesting corpus data to be processed.

(1) Firstly, there are Chinese words whose computational linguistic properties are relatively stable in

the sense that they tend to be identified and segmented as dictionary entries compiled in computational applications, e.g. 祖国 ‘motherland’, 同胞 ‘compatriots’, 父母 ‘parents’, 友谊 ‘friendship’, 幸福 ‘happiness’, etc.

- (2) Secondly, there are Chinese morphologically derived words that can be generated from stem words via a productive morphological process. Within this category, four main pathways of morphological derivation have been established: affixation, repetition, splitting and merging.

Affixation refers to the process by which a stem word turns into a compound word after being modified by a prefix or suffix to meet certain Chinese grammatical requirements, such as plurality, tense or aspect, e.g. 我 ‘I’ → 我们 ‘we’; 看 ‘watch/see’ → 看过 ‘have watched/seen’; 做 ‘do’ → 做着 ‘doing’.

Similarly, repetition is another key strategy deployed in the derivational process of word formation in Chinese. It is widely used across Chinese words of different grammatical categories, e.g. noun, verb, adjective, adverb, etc., while functioning as an indicator of emphasis, quantifier of the quality described, or simply an attempt at more indirect speech. It is a rather interesting linguistic phenomenon in modern Chinese; however, since it tends to occur with multi-character words, it will not constitute a main concern in my analysis of the segmentation process as performed by CST at the level of two characters. Examples of repetitive words are 陈陈相因 ‘following old traditions year after year without any innovation’; 红彤彤 ‘red and bright’; 黑黝黝 ‘black and shining’.

The concept of split words is used to describe a set of expressions that are separate words at syntactic level but single words at a semantic level, such as in 出了名 ‘being famous’, the bi-character word 出名 ‘to be famous’ is split by the particle 了 denoting the perfect tense. In fact, split words constitute such a problematic issue that they account for a considerable number of mismatched tokens that appeared in the segmentation of the Chinese texts (see the following analysis of ten segmented sentences selected from Liu’s translation of *Don Quijote*).

Similar drawbacks have been experienced by CST when it comes to segmenting merged words, which refer to expressions that have been derived from compressing two or more single-character words, e.g. the word 不得已, which means ‘being obliged to do something’ is a compressed expression of two separated words 不得不 ‘have to do something’ and 自己 ‘oneself’.

- (3) Thirdly, there are the nominal entities, covering among others, personal and family names, food, clothes, architecture, social habits, subjects bearing on literary, religious or historical issues, etc.

This word category of named entities has proved to be an even more challenging task for CST when segmenting the Chinese texts in CCDQ, for it entails the recognition of unknown words in the modern Chinese translations of *Don Quijote*. In fact, in most cases CST has failed to retrieve the translated foreign nominal entities as listed above as integral and legitimate tokens; and the outcome of

segmentation has been notably fragmentary. This seems to suggest that any attempt at studying foreign-name-related subject matters in the Chinese texts of CCDQ would turn out to be particularly infertile.

- (4) Lastly, there are newly created words, which together with nominal entities constitute one of the most difficult tasks in Chinese word segmentation, namely unknown word recognition (Gao and Li, 2005: 531-74).

In CCDQ, the creation of new Chinese words has been necessary because of the lack of a similar social, cultural or political context in the target language; and despite their limited quantity, they play a key role in our understanding of the cross-cultural interpretation of the original text made by the Chinese translators. However, no matter how appealing this subject may appear, it cannot be addressed properly as a potential research topic at this stage, for my main concern here is the sort of technical hitch one encounters when trying to extract these new words correctly with linguistic tools such as the CST.

The next ten segmented sentences, selected at random from Liu's translation of *Don Quijote* Part I, will be used for an initial evaluation of the performance of CST in segmenting the Chinese corpus texts of CCDQ at the two-character level. The assessment of the segmentation results will draw on the four main Chinese word categories explained above, i.e. dictionary entries, morphologically derived words, named entities, and new words. Some of the manually corrected segments may exceed the two-character length, but the selection criterion is based on their original length as defined by CST. Meanwhile, preference for the selection of samples is given to long descriptive or narrative text strings in the target texts, where the extraction of segments within a sentence is supposed to be too sophisticated for an automatic segmenter.

Items in boxes indicate segments incorrectly identified by CST either by breaking their integrity as a single word or by incorporating irrelevant morphemes; a slash is placed between two characters that the segmentation tool has failed to separate; a line drawn under a Chinese word unit is the original segmentation mark left by CST; DWs for lexical items identified as dictionary entries; MDWs for morphologically derived words, NEs for named entities and NWs for newly created words. Only those mismatched entities are marked with their respective categories as DWn, MDWn, NEn or NWn for further analysis of the segmenting performance of CST. No manual resegmentation was carried out.

1. CH, (Liu) I, I: 13 锅里牛肉比羊肉多, 晚餐常吃 / 凉拌肉丁 (NE₁) 'cold chopped beacon', 星期六吃 / 脂油煎鸡蛋 (NE₂) 'fried eggs', 星期五吃扁豆, 星期日加一只 / 野雏鸽 (NE₃) 'wild pigeon', 这就用去了他四分之三的收入。

CS I, I: 38 Una olla de algo más vaca que carnero, salpicón (NE₁) las más noches, duelos y quebrantos (NE₂) los sábados, lantejas los viernes, algún palomino (NE₃) de añadidura los domingos, consumían las tres partes de su hacienda.

EN I, I: A midday stew with rather more shin of beef than leg of lamb, the leftovers for supper most nights, lardy eggs on Saturdays, lentil broth on Fridays and an occasional pigeon as a Sunday treat ate up three-quarters of his income.

2. CH (Liu) I, IV: 29 “即便如此，”唐吉珂德 (NE₄) ‘don Quijote’/ 说，“鞋钱和输血的钱也被你无缘无故地抽打他抵消了。就算他把你给他买的鞋穿破了，可是你也把他的皮打破了；就算他生病时理发师为他/输了血 (MDW₁) ‘blood transfusion’/, 他没病时你却把他打出了血。这样说来，他就不/欠你钱 (MDW₂) ‘owe you money’/了。”

CS, I, IV: 69 —Bien está todo eso —replicó don Quijote (NE₄)—, pero quédense los zapatos y las sangrías por los azotes que sin culpa le habéis dado, que, si él rompió el cuero de los zapatos que vos pagastes, vos le habéis rompido el de su cuerpo, y si le sacó el barbero sangre (MDW₁) estando enfermo, vos en sanidad se la habéis sacado; así que por esta parte no os debe (MDW₂) nada.

EN, I, IV: ‘That is all very well’, replied Don Quixote, ‘but the shoes and the blood-lettings will be set against the flogging you have given him without due cause: for if he has done some damage to the hide of the shoes that you bought him, you have damaged his own hide and if the barber bled him when he was ill, you have done the same to him in good health; so that on this account he owes you nothing’.

3. CH, (Liu) I, V: 33 看到自己动弹不得，唐吉珂德 (NE₅) ‘don Quijote’/ 想起了自己的老办法——回想小说中的某一情节。他又疯疯癫癫地想起巴尔迪维诺 (NE₆) ‘Valdovinos’/ 在山上被卡尔洛托 (NE₇) ‘Carloto’/ 打伤后遇到/曼图亚 (NE₈) ‘Mantua’/ 侯爵的故事。这个故事孩子们知道，青年人知道，老年人更是大加赞赏，深信不疑，就像笃信穆罕默德的故事一样

CS, I, V: 76-77 (Don Quijote) (NE₅) Viendo, pues, que, en efeto, no podía menearse, acordó de acogerse a su ordinario remedio, que era pensar en algún paso de sus libros, y trájole su locura a la memoria aquel de Valdovinos (NE₆) y del marqués de Mantua (NE₈), cuando Carloto (NE₇) le dejó herido en la montaña, historia sabida de los niños, no ignorada de los mozos, celebrada y aun creída de los viejos, y, con todo esto, no más verdadera que los milagros de Mahoma.

EN, I, V: Finding, then, that he could not move, it occurred to him to resort to his usual remedy, which was to think about some passage from his books; and his madness brought to his memory the episode from the story of Baldwin and the Marquis of Mantua in which Carloto leaves Baldwin wounded in the forest, a tale to every little boy, not unfamiliar to youths, celebrated and even believed by old men, yet with no more truth in it than the miracles of Muhammad.

4. CH, (Liu) I, IX: 57 首先/出击者 (DW₁) ‘the person who strikes a blow’/ 是倜傥的/比斯开人 (NE₉) ‘Biscayan’/. 这一剑有力凶猛，要不是劈偏了，完全可以把他/比斯开人 (NE₁₀) ‘Biscayan’/ 桀骜的对手干掉，我们的骑士及其/征险 (NW₁) ‘in search of adventures’/ 生涯也就结束了。然而幸运的是，还有更重要的事情有待这位骑士去完成，所以利剑劈偏，只是把他左半边的甲冑 (DW₂) ‘armour’、大半个/头盔 (DW₃) ‘helmet’/ 和半只耳朵由左肩劈下 (MDW₃) ‘divide straight through’/, 七零八落地散在地上，使骑士十分难堪。

CS, I, IX: 121 Y el primero que fue a descargar el golpe (DW₁) fue el colérico vizcaíno (NE₉); el cual fue dado con tanta fuerza y tanta furia, que, a no volvérselo la espada en el camino, aquel solo golpe fuera bastante para dar fin a su (vizcaíno) (NE₁₀) rigurosa contienda y a todas las aventuras) (NW₁) de nuestro caballero; mas la

buena suerte, que para mayores cosas le tenía guardado, torció la espada de su contrario, de modo que, aunque le acertó en el hombro izquierdo, no le hizo otro daño que desarmar (MDW₃) le todo aquel lado, llevándole de camino gran parte de la celada (DW₃), con la mitad de la oreja, que todo ello con espantosa ruina vino al suelo, dejándole muy maltrecho.

EN, I, IX: The fiery Biscayan was the first to strike a blow, which was delivered with such force and fury that had not the sword turned in its course, that single stroke would have sufficed to put an end to the bitter struggle and to all the adventures of our knight; but that good fortune which reserved him for greater things, turned aside the sword of his adversary, so that although it smote him upon the left shoulder, it did him no more harm than to strip all that side of its armour, carrying away a great part of his helmet with half of his ear, all which with fearful ruin fell to the ground, leaving him in a sorry plight.

5. CH, (Liu) I, XX: 133 桑乔 (NE₁₁) ‘Sancho’/ 仍习惯地拉着他的驴, 这是和他荣辱与共的伙伴 (DW₄) ‘friends’, 紧跟在唐吉珂德 (NE₁₂) ‘don Quijote’/ 后面。他们在那些遮云蔽日 (DW₃) ‘covering clouds and sky’/ 的栗树和其它树中间走了很长一段路, 发现在高高的岩石脚下有一块草地, 一股激流从岩石上飞泻 (MDW₄) ‘running straight down’/ 而下。

CS I, XX: 238 Seguía Sancho (NE₁₁) a pie, llevando, como tenía de costumbre, del cabestro a su jumento, perpetuo compañero (DW₄) de sus prósperas y adversas fortunas; y habiendo andado una buena pieza por entre aquellos castaños y árboles sombríos (DW₃), dieron en un pradecillo que al pie de unas altas peñas se hacía, de las cuales se precipitaba (MDW₄) un grandísimo golpe de agua.

EN, I, XX: Sancho followed him on foot, leading by the halter, as his custom was, his ass, his constant comrade in prosperity or adversity; and advancing some distance through the shady chestnut trees they came upon a little meadow at the foot of some high rocks, down which a mighty rush of water flung itself.

6. CH (Liu) I, XXXIX: 312 在去亚历山大里亚德拉帕利亚 (NE₁₂) ‘Alessandria della Paglia’/ 的路上, 我听说伟大的阿尔瓦 (NE₁₃) ‘Alva’/ 公爵正要去佛兰德 (NE₁₄) ‘Flanders’, 就又改变了主意, 投奔了他, 服侍他巡行, 处死埃格蒙 (NE₁₅) ‘Egmont’/ 和奥尔诺斯 (NE₁₆) ‘Horn’/ 伯爵的时候我也在场。后来, 我终于在瓜达拉哈拉 (NE₁₇) ‘Guadalajara’/ 一位名叫迭戈德乌尔维纳 (NE₁₈) ‘Diego de Urbina’/ 的著名军官手下当上了少尉。

CS, I, XXXIX: 496 y, estando ya de camino para Alejandría de la Palla (NE₂), tuve nuevas que el gran duque de Alba (NE₁₃) pasaba a Flandes (NE₁₄). Mudé propósito, fuime con él, servíle en las jornadas que hizo, halléme en la muerte de los condes de Eguemón (NE₁₅) y de Hornos (NE₁₆), alcancé a ser alférez de un famoso capitán de Guadalajara (NE₁₇), llamado Diego de Urbina; (NE₁₈)

EN, I XXXIX: as I was already on the road to Alessandria della Paglia, I learned that the great Duke of Alva was on his way to Flanders. I changed my plans, joined him, served under him in the campaigns he made, was present at the deaths of the Counts Egmont and Horn, and was promoted to be ensign under a famous captain of Guadalajara, Diego de Urbina by name.

7. CH (Liu) I, XL: 322 有的人要这种签名信 (DW₆) ‘letters with signature’/ 并没有歹意, 而有的人则别有用心, 以防万一。例如, 他们去基督教国家抢掠时被抓住了, 就拿出签名信 (DW₇) ‘letters with signature’, 说这信可以证明他来的目的, 是要留在基督教国家里, 而抢掠则是被土耳其人强迫所为。

CS I, XL: 510 Algunos hay que procuran estas fees (DW₆) con buena intención, otros se sirven dellas acaso y de

industria: que, viniendo a robar a tierra de cristianos, si a dicha se pierden o los cautivan, sacan sus firmas (DW₇) y dicen que por aquellos papeles se verá el propósito con que venían, el cual era de quedarse en tierra de cristianos, y que por eso venían en corso con los demás turcos.

EN, I, XL: Some obtain these testimonials with good intentions, others put them to a cunning use; for when they go to pillage on Christian territory, if they chance to be cast away, or taken prisoners, they produce their certificates and say that from these papers may be seen the object they came for, which was to remain on Christian ground, and that it was to this end they joined the Turks in their foray.

8. CH (Liu) I, I: 14 不过, 所有这些小说, 他都觉得不如 闻名遐迩 (DW₈) 'famous' / 的 费利西亚诺德席尔瓦 (NE₁₉) 'Feliciano de Silva' / 写得 好, 此人的 平铺直叙和繁冗 (MDW₅) 'elongated' / 陈述被他视为明珠, 特别在读到那些殷勤话和挑逗信时 'when' / 更 (DW₉) 'especially' / 是如此。许多地方这样写道: “以你无理对我有理之 / 道理 (DW₁₀) 'reason', 使我自觉理亏, 因此我埋怨你漂亮也有 / 道理 (DW₁₁) 'reason'。”还有: “高空以星星使你的神圣更加神圣, 使你 / 受之无愧 (DW₁₂) 'being worth something' / 地接受你 / 受之无愧 (DW₁₃) 'being worth something' / 的伟大称号而 受之无愧 (DW₁₄) 'being worth something' /。”

CS, I, I: 40 y, de todos, ningunos le parecían tan bien como los que compuso el famoso (DW₈) Feliciano de Silva, porque la claridad de su prosa y aquellas enricadas (MDW₅) razones suyas le parecían de perlas, y más / cuando (DW₉) llegaba a leer aquellos requiebros y cartas de desafíos, donde en muchas partes hallaba escrito: “La razón de (DW₁₀) la sinrazón que a mi razón se hace, de tal manera mi razón enflaquece (DW₁₁), que con razón me quejo de la vuestra fermosura”. Y también cuando leía: “Los altos cielos que de vuestra divinidad divinamente con las estrellas os fortifican y os hacen merecedora (DW₁₂) del merecimiento (DW₁₃) que merece (DW₁₄) la vuestra grandeza”.

EN, I, I: But of all there were none he liked so well as those of the famous Feliciano de Silva's composition, for their lucidity of style and complicated conceits were as pearls in his sight, particularly when in his reading he came upon courtships and cartels, where he often found passages like "the reason of the unreason with which my reason is afflicted so weakens my reason that with reason I murmur at your beauty;" or again, "the high heavens, that of your divinity divinely fortify you with the stars, render you deserving of the desert your greatness deserves."

9. CH (Liu) I, III: 23 他自己年轻的时候也曾投身于这项光荣事业, 周游各地, 到处 征险 (NW₂) 'in search of adventures', 连 马拉加 (NE₂₀) 'Malaga' / 的 佩切莱斯 (NE₂₁) 'Percheles', 里亚兰岛 (NE₂₂) 'Rairan', 塞维利亚 (NE₂₃) 'Seville' / 的 孔帕斯, 塞哥维亚 (NE₂₄) 'Segovia' / 的 阿索格拉 (NE₂₅) 'Azoguevo', 巴伦西亚 (NE₂₆) 'Valencia' / 的 奥利韦拉 (NE₂₇) 'Olivera', 格拉纳达 (NE₂₈) 'Granada' / 的 龙迪利亚 (NE₂₉) 'Rondilla', 圣卢卡尔 (NE₃₀) 'San Lucar' / 海滩、科尔多瓦 (NE₃₁) 'Cordoba' / 的 波特罗 (NE₃₂) 'Colt', 托莱多 (NE₃₃) 'Toledo' / 的小客店和其他一些地方都去过, 凭着手脚利索, 勾引过许多寡妇, 糟蹋过几个少女, 还欺骗了几个孤儿, 干了 不少伤天害理的事, 几乎在西班牙所有法院都 挂了号 (MDW₆) 'make himself known'.

CS I, III: 59 y que él ansimesmo, en los años de su mocedad, se había dado a aquel honroso ejercicio, andando por diversas partes del mundo, buscando sus aventuras (NW₂), sin que hubiese dejado los Percheles (NE₂₁) de Málaga (NE₂₀), Islas de Riarán (NE₂₂), Compás de Sevilla (NE₂₃), Azoguevo (NE₂₅) de Segovia (NE₂₄), la Olivera (NE₂₇) de Valencia (NE₂₆), Rondilla (NE₂₉) de Granada (NE₂₈), Playa de Sanlúcar (NE₃₀), Potro (NE₃₂) de Córdoba (NE₃₁) y las Ventillas de Toledo (NE₃₃) y otras diversas partes, donde había ejercitado la ligereza de sus pies, sutileza de sus manos, haciendo muchos tuertos, recuestando muchas viudas, deshaciendo algunas doncellas y engañando a algunos pupilos y, finalmente, dándose a conocer (MDW₆) por cuantas audiencias y

tribunales hay casi en toda España;

EN, I, III: and that he himself in his younger days had followed the same honourable calling, roaming in quest of adventures in various parts of the world, among others the Curing-grounds of Malaga, the Isles of Riara, the Precinct of Seville, the Little Market of Segovia, the Olivera of Valencia, the Rondilla of Granada, the Strand of San Lucar, the Colt of Cordova, the Taverns of Toledo, and divers other quarters, where he had proved the nimbleness of his feet and the lightness of his fingers, doing many wrongs, cheating many widows, ruining maids and swindling minors, and, in short, bringing himself under the notice of almost every tribunal and court of justice in Spain;

10. CH (Liu) I, IV: 32 “骑士大人，”那个商人说，“我以 在场 (DW₇) ‘present’ / 所有 (DW₈) ‘all’ / 王子的名义请求您，让我们承认我们 前所未见 (DW₉) ‘never heard before’ / 前所未闻的事情，实在 于心不安 (DW₁₀) ‘being restless’，而且，这会严重伤害 / 阿尔卡利亚 (NE₃₄) ‘Alcarria’ / 和 / 埃斯特雷马杜拉 (NE₃₅) ‘Extremadura’ 的那些女皇和王后们。

CS I, IV: 74 —Señor caballero —replicó el mercader—, suplico a vuestra merced en nombre de todos (DW₈) estos príncipes que aquí estamos (DW₇) que, porque no encarguemos nuestras conciencias (DW₁₀) confesando una cosa por nosotros jamás vista ni (oída) (DW₉), y más siendo tan en perjuicio de las emperatrices y reinas del Alcarria (NE₃₄) y Extremadura (NE₃₅)

EN, I, IV: "Sir Knight," replied the trader, "I entreat your worship in the name of this present company of princes, that, to save us from charging our consciences with the confession of a thing we have never seen or heard of, and one moreover so much to the prejudice of the Emperresses and Queens of the Alcarria and Extremadura,

As can be seen from the manual re-segmentation of the ten samples selected throughout the Chinese corpus texts, the types of mismatched tokens reveal imbalanced occurrences: while CST failed to recognize correctly as many as thirty-five translated named entities (NEs), only ten instances of dictionary words (DWs) and a single newly created words (NWs) have been mismatched by the application. As for morphologically derived words (MDWs), three out of six mismatched tokens are split words (输了血 ‘blood transfusion finished’; 欠你钱 ‘owing you money’; 挂了号 ‘being registered’), while two are merged words (飞泻 ‘run-discharge’; 繁冗 ‘complicated-elongated’) and one verbal compound formed through affixation (劈下 ‘cut-down’).

I might argue that the reason why words already existing in dictionaries are still being misrecognized is due to the maximal matching principle on which the CST has been programmed to segment. As in the cases of DW3, DW9, DW10, DW11, a legitimate single character word at the head of a textual string is “covered” by another dictionary word collected in CST which contains the former as part of it. Meanwhile, the segmentation problem with dictionary words is supposed to have been caused by the limited compilation of dictionary words of the current version of CST. That is to say, despite the ongoing collection of Chinese words by CST, the number of words already stored is still trivial when compared with the very large number of words needed to carry out a satisfactory Chinese text segmentation.

1.3.3.3 Conclusion

In the course of processing the Chinese translated texts included in the parallel corpus of *Don Quijote* (CCDQ), a higher degree of precision in the segmentation of four-character items can be reported, undoubtedly due to the fact that most of them, including words and phrases, are idiomatic or collocational expressions. Machine-aided segmentation of two-character words, on the other hand, probably due to the technical complexity implied in the automatic discrimination of diverse morphological formations of two-character lexical items, is not so successful. To further explain the nature of the mis-segmentation produced by the automatic segmentation tool, I offer the manual segmentation of ten samples extracted randomly from Liu's translation. As different from machine-based segmentation, the process of manual segmentation is largely based on the researcher's own understanding of the text and her intuition as a native speaker of Chinese. The contrast of the results between the two types of segmentation shows that the two-character segments wrongly identified by the natural language processing tool are relatively large and can therefore hardly be used for the next stage of research. The problems identified in the first stage of Chinese corpus text processing have important implications for the conception of this project as a whole, since only the data generated by the segmentation tool and fully validated should be the basis for the textual analysis.

1.4 Parallel text alignment

Once the evaluation exercise as to the effectiveness and reliability of the segmentation results provided by CST is completed, the construction of CCDQ consequently enters into the next stage: parallel alignment of the selected source and target texts in preparation for the deployment of analyzing tools. In order to facilitate later retrieval of textual information, the alignment of CCDQ becomes especially important given the deficiencies of the Chinese segmentation tool as shown above. That is, the task of aligning CCDQ, however time-consuming and laborious it may be, is crucial if we are to make up for the errors and inadequacies of Chinese text segmentation as shown above.

In fact, these are two mutually complementary procedures. On the one hand, without a successful segmentation of Chinese text strings it would be impossible to process the Chinese corpus with any computational linguistic application. On the other hand, in the absence of an accurate parallel text alignment, the significance of Chinese text segmentation would be reduced to producing simply an extensive wordlist of the Chinese corpus with a number of mismatched tokens. This is obviously not the end result that one would hope to achieve by using corpora to study translations and translating styles.

1.4.1 Initial experiments with automatic aligners

In constructing the sort of linguistic corpora we are concerned with, many researchers would find themselves greatly assisted by various bilingual or multilingual aligners already available.

Unfortunately most of the aligning tools one can think of have been developed for the building of corpora and the study of texts written in similar languages, like, say, alphabetical languages. It is definitely beyond the capacity of such tools to align automatically a typically alphabetical language such as Castilian with a typically ideographic language like Chinese. Unaware of this, I started the alignment of CCDQ straightaway using a popular aligning and concordancing tool available at the time; but I soon realized that things were not going to be that easy. The linguistic tool deployed for the task is known as ParaConc (version 269), the popular parallel concordancer developed by Barlow in 1995.

As can be inferred from its name as a parallel concordancer, the main function of *ParaConc* is to retrieve identical translation equivalents for cross-linguistic study, by furnishing from basic to sophisticated searching functions, such as text search, regular expression search, tag search, as well as other useful textual information, e.g. word and/or character frequency and collocation frequency, etc. An automatic aligning device also constitutes an integral part of the whole program, which has proved to be quite handy with the alignment of a number of language pairs (Barlow, 1995a). Nevertheless, its performance on corpora in non-alphabetical languages like Chinese turns out to be far from satisfactory; and the reason for this lies in the principle on which it has been programmed to work.

The model on which the automatic aligning function has been developed in ParaConc is known as the Gale and Church Algorithm (1993). It is basically a statistical approach to sentence alignment with little linguistic knowledge assistance, e.g. cognate words or bilingual lexicons. It works on the assumption that among a given number of sentence alignment possibilities in the translation, the most likely candidate is the one of similar length to that of the source text sentence. That is, the most salient textual feature that an aligning application prioritizes in identifying two corresponding sentences in a parallel corpus is the similar sentence length shared by the source and target texts (Barlow, 1995b).

The idea of mapping local elements in parallel texts in accordance with the sentence length is derived from the fact that when a language is translated into another that belongs to the same linguistic system, longer sentences tend to be translated into longer sentences and shorter sentences into shorter ones. It is implicit in this presumed aligning model that a given sentence in the source language tends to be translated into an analogous sentence in the target texts by reason of the overall length shared by the two, whilst at the same time given the comparable syntactic structures that the two languages have in common, the syntactic integrity of the original sentence is maintained after being rendered into the translational language.

As a result, this similar sentence-length aligning principle appears to work well with source and target texts written in languages that are morpho-syntactically analogous to each other, like Spanish and French, since generally speaking, when one of them is translated into another, the basic structure of the source text can be reasonably retained, thus maintaining the comparability of the translational language with the original one.

However, it is an idealized aligning environment established on the observation of a rather macro-structural feature of the parallel texts and has therefore very limited potential when applied to two morpho-syntactically different languages, such as early seventeenth-century Castilian and modern Mandarin Chinese. There is a basic lack of comparability between the two languages in question, either in terms of their morphological features or with regards to their syntactic organization, not to mention the length of sentence.

Aligning tools based on the sentence-length principle, such as *ParaConc*, may not require special knowledge of the languages on the part of the corpus developer, but minimal prior alignment of paragraphs is still desirable in that it will help to constrain the automatic search for points of correspondence, including among others, punctuation marks or cognate words. Cognateness is another textual feature that may help an aligner to locate the position of a given translated sentence in the target text. Cognates are words that are related or descended from a common ancestor (have a common origin); in other words, they are etymological variations of the same word, possibly derived from a common predecessor language. For instance, in translating *Don Quijote*, the English translator immediately has at his disposal a whole list of Spanish-English cognates that may well be inserted into his/her work in the first instance (Jarman, 2003), for example, admire/admirar; arrogance/arrogancia; barber/ barbero; battle/batalla; band/banda; capacity/capacidad; capricious/ caprichoso; conflict/conflicto; debate/debate; decadence/decadencia; dialect/dialecto; etc.

All these cognate words, serving as anchor points throughout the target texts, do provide corpus aligning tools with a good start in locating the position of a given translated phrase, as well as in matching up later with the source text sentence. Unfortunately, this kind of first-aid facility afforded by lexical similarities in some language combinations is not remotely applicable in the case of seventeenth-century Castilian and contemporary Mandarin Chinese. To compensate for the lack of such a natural lexical association as exists between cognate languages, increasing efforts have been invested in recent times in compiling electronic bilingual dictionaries of Chinese and English, in order to facilitate the envisaged automatic alignment of parallel corpora written in those two languages. But again not much work has yet been done in the field of Chinese and Spanish: text mining and information retrieval techniques to prepare the ground for the comparative study of this particular language pair remain essentially unexplored.

1.4.2 The need for manual alignment

While *ParaConc* has proved to be reasonably effective in performing a semi-automatic alignment of texts written in most alphabetical languages, it is clearly beyond its capacity to perform such task with the bilingual texts of CCDQ. Consequently, the textual alignment of CCDQ has proved to be a rather arduous task, requiring direct intervention on my part to introduce paragraph and sentence boundary breaks in both the source text and Chinese translations. As a result, a marked feature

of the outcome of the alignment of CCDQ shows that there is no standardized aligning unit, e.g. sentence or phrase, as I had expected at the outset of the construction of CCDQ.

In fact, it has turned out to be impossible even to establish an aligning unit, since there is no such thing as a consistent syntactic or structural correspondence throughout the Chinese corpus and the Spanish corpus of CCDQ. Therefore, the alignment of CCDQ has been done largely at semantic level: a breaking mark is placed where there is a seemingly maximal combination of sentential elements that carry identical semantic units in both the source and target texts. Although great efforts have been made to maintain the syntactic integrity of corpus texts in both languages, retaining a certain flexibility so as to ensure the parallel matching may be of practical use later has been definitely my top priority.

Next, I shall discuss in detail the manual aligning process of CCDQ on the basis of linguistic contrasts between Castilian and Mandarin Chinese, which incidentally led to unexpected technical hitches in the course of textual alignment. Burton Raffel comments in the preface to his translation of *Don Quijote* (Raffel, 1996: xvii): “No one can reproduce Cervantes’s style in English. Not only is his prose uniquely magnificent, but the very music of Spanish, *its syntactical structures*, and the thrust and flavor of its words, are literally untransportable into another language”¹⁶. What can be suggested from Raffel’s remarks is that, if in translating *Don Quijote* the syntactic structures of the language used by Cervantes have already issued notable challenges for its English translators, it would be even more uncertain for anyone attempting the establishment of a definite correspondence between the syntactic structures of Cervantes’s masterpiece and its translation into contemporary Mandarin Chinese. Actually, in aligning the parallel corpus texts of CCDQ, a rather thorny technical issue that has emerged constantly is closely related to locating the sequence of source text segments in their corresponding Chinese translations, since the former, when rendered into modern Chinese, would have undergone a series of considerable structural changes.

Faced with the fact that the two languages in question have very little in common with regards to their syntactic structures, the alignment of the parallel corpus texts of CCDQ inevitably entails a number of specific editing operations, e.g. the breaking-up of Castilian long or multilayered sentences, and the combining of short and less-modified sentences in modern Chinese. Here, it should be noted that such aligning procedures in terms of the rearrangement of textual elements at a level which is above the sentence or clause boundaries as defined by punctuation marks, have been attempted largely drawing on my own understanding of *Don Quijote* and its two Chinese versions, thus it can hardly be standardized, as one would expect from an automatic or machine-aided aligner.

Nonetheless, the manual alignment of CCDQ has not been a deliberate practice of text processing, but it has been carried out as strictly as possible in line with a minimum matching principle in terms of the pairing up of basic semantic units identified in the parallel texts. Here the size of the semantic unit as the basic constituent of the alignment is crucial in that firstly, it should allow a textual

¹⁶ Italics are mine.

sketch to carry a moderately independent meaning in the sentential context, which is to facilitate later comparison of the source and target texts; secondly it had to be of an appropriate length so that it could provide the researcher with sufficient contextual information as shown by concordance programmes in the parallel corpora mining.

1.4.3 Assessment of manual alignment

1.4.3.1 Technical problems caused by different word orders

Notwithstanding the fact that Castilian and Mandarin Chinese share a common syntactical arrangement feature in terms of the order of major sentential constituents, known as subject + verb + object (SVO), there is still a substantial difference between the two languages regarding the flexibility or mobility of grammatical elements in each of the three positions of subject, verb and object (SVO), as well as the placement of other sentence complements within this underlying syntactical structure.

For instance, whereas in a Spanish sentence a subject can be simply omitted in the presence of the inflected predicate verb, this would always be a serious grammatical error in modern Chinese; on other occasions, when a Spanish subject noun can be easily placed past the main verb or to the end of a whole phrase, this linguistic phenomenon would only be acceptable in classical Chinese under specific circumstances, e.g. to emphasize speaker's attitude towards a given proposition; and its function as a rhetorical device in the past is rarely seen in modern Mandarin Chinese. Therefore, when a subject is omitted or altered in the Castilian sentence, deliberately or not, the Chinese translator would have to sacrifice the meaning behind the shift in the position of the Spanish subject, if s/he wishes to maintain the logicity of the target language within the context of a SVO structure.

Compared with a number of Indo-European languages such as Castilian where the long-standing grammaticalization, through verbal inflections, gender and number agreement, etc., has given rise to rigid corresponding relationships among different syntactical categories, a fundamental structuring principle that may account for the proper functioning of contemporary Mandarin Chinese is word order (词序). A prevailing linguistic fact in contemporary Chinese is the crucial role that word order plays in conveying a logical and unambiguous message, which is supposed to be a demonstration of the modestly grammaticalized linkage existing among different sentential elements.

Generally speaking, contemporary Chinese can be categorized safely under the S (subject) V (verb) O (object) type. However, there is a particular syntactic property which sets Chinese SVO apart from that of Indo-European languages: whereas in most SVO-type modern linguistic systems the subordinate clause tends to be placed after the head word, in modern Chinese the subordinate clause is bound to be located in front of the part it refers to.

Accordingly, there is a key issue in translating a complex Castilian sentence into Chinese, i.e.

the repositioning of modifying elements in the source text language when it is rendered into modern Chinese. For instance, in translating a complex sentence found in *Don Quijote*, if we take the main subject as the axis, then most modifying elements tend to follow from left to right, whereas in Mandarin Chinese these referential elements would tend to go the other way round. That is to say, with the main subject as the balancing point, more weight of modification is given to the left of the subject in a complex Chinese sentence.

As a result, the outcome of the Chinese translation of a complex Castilian sentence from time to time shows a reverse order of the original sentence in terms of the position of the subject and its modifiers; or in some cases, in order to adapt the sequence of the original sentence to the linguistic habits of modern Chinese, which avoids the use of extremely long modifying elements prior to the subject, the Chinese translators tend to extract the modifying elements of the original sentence and put them in a single clause immediately after the subject being modified, or by using certain prepositions to express the relations between the subject and its modifier to make the translation more Chinese-sounding. That is, in order to maintain the fluidity of the target language, the Chinese translator feels obliged to make structural alterations, which may well result in considerable changes in word order when translating complex sentences from the source language into the target language.

1.4.3.2 Technical problems caused by different syntaxes

There are various difficulties in translating a lengthy complex Castilian sentence into Mandarin Chinese. One of the main problems arises from the lack of relative pronouns in Mandarin Chinese and its unique habit of using prepositional phrases. While in an extensive Castilian sentence with several parts being modified by subordinate clauses, one has a whole range of relative pronouns to construct a fairly complex sentence: *que, el (s) / la (s) que, cual, el (s) / la (s) cual (es), quien (es), donde, adonde, cuando, como, cuyo*, etc; this kind of complex sentence can never find equivalent structures in Chinese, where even the concept of relative pronoun has been created to help explain linguistic phenomena occurring in foreign languages. It does not constitute an integral part of Chinese syntax.

As regards the placement of prepositional phrases, contemporary Chinese favours the use of prepositional phrase prior to the verb being specified. Another interesting syntactic phenomenon in contemporary Chinese is the coexistence of post-positing and preceding prepositions, with the former of a slightly higher frequency of occurrence. All these linguistic properties of modern Chinese have made the practice of syntactic borrowing quite impossible, a strategy that is slightly more common in the translation between Western European languages.

So, when these multilayered Castilian sentences constructed with a number of relative pronouns and prepositional phrases are translated into Chinese, they tend to get broken down into a number of relatively shorter clauses, and these disaggregated sentential elements are then rearranged or

inserted in specific positions of a sentential sequence that is familiar to the Chinese readership. At times, this process of integration is completed with the assistance of punctuation marks, including a full stop, being placed where appropriate in order to avoid an excessively loose structure of the translation, and thus changing the proportion of source text to translated sentences from 1:1 to 1:2, 1:3 or even higher

All this desegregation and rearrangement of original sentential elements in the Chinese translation adds up to the technical hitches which I have experienced in the course of aligning CCDQ. The following is a good example of the way one single Castilian sentence has been translated into many Chinese clauses. (Spanish) RP_n for any relative pronoun; Sn for any sentence or clause ending with a full stop, semicolon or comma; PRP_n for any prepositional phrase; a single slash for a clausal breaking position; a double slash for a sentential breaking position. Sn and slashes are introduced here to facilitate the comparison between the Spanish and its Chinese translation. (Chinese): ADP_n for any adjective phrase; PP_n for any personal pronoun; Sn for any sentence ending with a full stop; JC_n for any juxtapositional phrase; W_n for a character word or phrase; a double slash for a sentential breaking position.

CS, I, II: 52 Estaban acaso a la puerta dos mujeres mozas, destas que (RP₁) llaman del partido (S₁)//, las cuales (RP₂) iban a Sevilla con unos arrieros / que (RP₃) en la venta aquella noche acertaron a hacer jornada (S₂); //y como a nuestro aventurero todo cuanto pensaba, veía o imaginaba le parecía ser hecho y pasar al modo de lo que había leído (S₃), //luego que vio la venta se le representó que era un castillo / [con sus cuatro torres y chapiteles de luciente plata](PRP₁)/, [sin faltarle su puente levadiza y honda cava] (PRP₂)/, [con todos aquellos adherentes] (PRP₃)/ que (RP₄) semejantes castillos se pintan (S₄)//.

EN, I, II: At the door were standing two young women, girls of the district as they call them, on their way to Seville with some carriers who had chanced to halt that night at the inn; and as, happen what might to our adventurer, everything he saw or imaged seemed to him to be and to happen after the fashion of what he read of, the moment he saw the inn he pictured it to himself as a castle with its four turrets and pinnacles of shining silver, not forgetting the drawbridge and moat and all the belongings usually ascribed to castles of the sort.

CH (Liu) I, II: 19 恰巧门口有两个青年女子,人们称之 (PP₁) ‘call them’ 为风尘女 (S₁) ‘prostitute’//。她们 (PP₂) ‘they’ 随同几个脚夫去塞维利亚, [今晚就投宿在这个客店里] (JC₁) (S₂) ‘tonight they will stay in this inn’//。我们这位冒险家所思所见所想象的, 似乎都变成了现实, 一切都和他在书上看到的一样 (S₃)//。客店在他眼里变成了城堡, 和书上描写的一样 (S₄) ‘just the same way as described in books’, [周围还有四座望楼, 望楼尖顶银光闪闪] (JC₂) ‘there are four turrets around, and their pinnacles of shinning silver’, [吊桥] (W₁) ‘drawbridge’、[壕沟] ‘moat’ (W₂) [一应俱全] (W₃) ‘everything included’//。

CH (Yang) I, II: 18 恰巧客店门口站着两个年轻女人, 所谓 (ADP₁) ‘so-called’ 跑码头的娘们 (S₁) ‘women who frequent the dock’//。她们 (PP₁) ‘they’ 是跟 [当夜在店里投宿的] (ADP₂) ‘staying in the same inn’ 几个骡夫一起到赛维利亚去的 (S₂)//。我们这位冒险家所思所见、所想象的事物, 无一不和他书上读到的一模一样 (S₃)//, 所以这个客店到他眼里马上成为一座堡垒, [周围四座塔, 一个个塔尖都是银光闪闪的] (JC₁) ‘there are turrets around and the pinnacle of each turret is of shinning silver’; 凡是书上写的 [吊桥] (W₁) ‘drawbridge’、[壕沟] (W₂) ‘moat’ 等等, 这里 [应有尽有] (W₃) (S₄) ‘everything included’。

From the sample pair analyzed above, we can see that in dealing with a long Spanish sentence in *Don Quijote*, the Chinese translator adopted a series of strategies to break it down into a number of interrelated clauses or relatively independent sentences of shorter length¹⁷. To be more specific, the kind of sentential manipulation reflected in the target language includes:

- 1) Breaking down the long sentence at the point where a Spanish relative pronoun occurs and starting a new sentence with a Chinese personal pronoun as the subject, e.g. S₁ and S₂;
- 2) Loosening the internal relationship between a main clause and the subordinate phrase which modifies it, by inserting a juxtapositional clause, e.g. JC₁ in S₂;
- 3) Extracting the semantic trunk of a long clause by postponing the modifying parts of prepositional phrases, and converting them later on into juxtapositional elements or single words, e.g. the extraction of S₄ and the alteration of the positions of PP₁, PP₂, PP₃, as well as their subsequent conversions into JC₂, W1, W2 and W₃.

As can be seen, rearranging the Spanish sentential structures in *Don Quijote* is rather a complex process of for the Chinese translator, as well as for anyone who wants to attempt a sentence by sentence alignment between the Spanish texts and their Chinese translations, for many semantic-syntactic parts of the original sentences overlap in the Chinese translations; or when they are rendered into Chinese, their positions are altered to a varying extent under different circumstances to confirm to Chinese grammatical conventions of sentence construction.

Therefore, in an attempt to maintain the integrity of the source language in *Don Quijote*, the only plausible way seems to apply a one-to-multiple model in aligning a Spanish sentence with a cluster of sentences and clauses as it is translated into Chinese. This one-to-multiple model has been developed mainly to cope with intricate lengthy sentences in *Don Quijote*. The commonly used one-to-one sentential alignment is still valid in dealing with simple Spanish sentences and their Chinese translations. The two linguistic points mentioned above only exemplify the considerable difficulties implied in the actual manual alignment of the parallel texts. It has indeed been arduous process which requires a great deal of dedication on my part. The huge efforts made at this early stage, however, have contributed to the construction of an efficient technical platform for the later textual analysis relying heavily upon the use of latest corpus linguistics applications and statistical software.

¹⁷ This does not necessarily imply that Chinese sentences are invariably shorter than Spanish; but in aligning the ST and TT of the CCDQ, the observation that Castilian long sentences get translated into much shorter modern Chinese sentences has turned out to be a common one.

1.5 Conclusion

We have discussed in this chapter the first phase in the construction of a parallel Castilian-Chinese corpus of Cervantes's *Don Quijote* (CCDQ). Due to the extensive manual work implied in the process of the corpus construction, I have just managed to do an in-depth study of the first volume of the novel, which, however, is a complete story in itself, since Cervantes's initial intention was to publish Part I only (Eisenberg, 1983: 3-34). So far as I am aware, this thesis is the first attempt at studying Chinese translations of this Castilian masterpiece within a corpus-based research scheme. This parallel corpus has been designed and constructed with a view to tracking down the stylistic differences in the two Chinese translations of *Don Quijote*. The two Chinese texts incorporated into the parallel corpus have been selected bearing in mind their representativeness as regards the particular period of time in which they were produced and the sort of reception they encountered.

In this sense, CCDQ is first and foremost a specific corpus for the purpose of analyzing and comparing two target texts under consideration. It makes an interesting area of investigation for it tries to elucidate the evolving nature of a translator's style within the context (time and situatedness), i.e. contextually and co-textually as suggested by Malmkjaer (1998: 534-41). Due attention to the source text is also most important in that it provides relevant information which would probably have an impact upon the way the Chinese translators manipulate the language in order to shape an idiosyncratic style of their own.

In the face of very limited technical resources available so far for the construction of parallel corpora in Spanish and Chinese, the conception of CCDQ has been quite a bold attempt on my part. A central theme of the first chapter of the thesis is the introduction and description of the major technical problems that I have encountered in constructing the parallel corpus of *Don Quijote*. I have made efforts in elucidating the developing nature of essential natural language processing tools such as Chinese text segmenters. It is believed that the significance of this corpus project for future studies will be precisely in the way technical and linguistic problems have been identified and pragmatically resolved.

As another one of the most challenging parts of the project, the difficulties of the alignment of CCDQ stem from the striking contrasts between the two languages. In view of their syntactical differences, special efforts have been made to remain flexible in extracting and aligning comparable sentential or clausal units in both the original and its two modern Chinese versions. Meanwhile, the alignment has been carried out on a minimal matching basis, which has been done entirely by myself. All the efforts put into the construction of the parallel corpus have made to facilitate further corpus data retrieval, which will be the topic of the next stage of my research.

Chapter II Corpus Data Retrieval and Classification

... many investigations in language, particularly in areas like stylistics, have remained at a relatively modest level of achievement for a very long time, simply because of the *technical problems involved in validating statements*.

(Sinclair, 2004: 16)¹⁸

2.1 Problem-oriented annotation

As with large-scale comparable corpora, small-scale topic-specific corpora may be fully or partially annotated, depending on the kind of research questions to be addressed. With individually built corpora that require a minimal level of annotation to make the raw corpus texts compliant with analytical tools that have been developed in languages other than the one used in the corpus texts, it is ideal to convert the texts first into Unicode, which may handle different writing or character encoding systems and display them properly in a range of computer programmes.

For some languages, the pre-processing stage of corpus construction may well be more sophisticated than a simple character code conversion, e.g. segmentation for Chinese natural texts (Feng, 2001: 35-42; Zhang & Liu, 2002: 1-7; Ji, 2007a: 93-109). This kind of minimally annotated corpora may be used to conduct simple quantitative textual analysis, such as token/type ratio, keyword indexing, frequency counts of collocations, etc. which are facilitated by various text mining functions provided in well-known standalone corpus software such as WordSmith, ParaConc, or Xaira (SARA)¹⁹.

For further extraction and exploration of linguistic information contained in the corpus texts, a number of marking-up schemes may be applied to raw corpus texts, e.g. part of speech (POS) tagging, syntactic parsing (Goyal and Chateerjee, 2006: 301-8), lemmatization (McEnery, et al. 1997: 219-26), semantic annotation (Thomas and Wilson, 1996: 92-109; Zelle and Ng, 1997: 45-64) and pragmatic or stylistic annotation (Leech and Weisser, 2003: 441-6; McIntyre et al. 2003: 513-22). At the present time, while the first type of linguistic annotation, i.e. automatic POS tagging, has achieved a relatively higher rate of precision with a number of languages (Garside and Smith, 1997: 102-21; Farwell, Helmreich and Casper, 1995: 42-56; Chang and Chen, 1993: 40-7), the remaining types of corpus tagging system listed above still require a considerable amount of human intervention or post-editing to enhance the validity of the output produced mechanically, the annotation sometimes even needing to be done entirely manually (Leech, et al. 1997: 85-101; Semino and Short, 2004).

As Sinclair observes, "... many investigations in language, particularly in areas like stylistics, have remained at a relatively modest level of achievement for a very long time, simply because of the technical problems involved in validating statements" (Sinclair, 2004: 16). As a potential solution to the technical shortcomings of literary studies, I would like to focus on a specific type of corpus

¹⁸ Italics are mine.

¹⁹ Xaira, or its older version SARA, may be downloaded from <http://www.natcorp.ox.ac.uk> or <http://www.xaira.org>.

annotation in the construction of small-scale topic-specific corpora, i.e. problem-oriented annotation (Ji, 2008a). This type of annotation differs from traditional types of corpus tagging systems in that it is highly pragmatic; it is usually developed by individual corpus builders so as to be able to look at the corpus texts from a post-data-extraction point of view. In other words, instead of investing a great amount of energy and time analyzing structures and subsequently machine-aided tagging raw corpus texts, one would first highlight certain linguistic features occurring in the aligned ST/TT texts, or in the target texts in particular, based on the following considerations.

First and foremost, any potential linguistic features selected for corpus-based study should have some structural or formal qualities that would make them easily retrievable from corpus texts and in sufficient quantity to be representative. Sparse data, no matter how interesting they might appear at first sight, are not to be considered in corpus-based textual analysis. Without a sound empirical grounding or sufficient verifiable textual evidence to support one's arguments, the validity of a corpus-based study would be seriously compromised. Furthermore, when the data mechanically extracted from the raw corpus do not amount to a sufficiently large basis for any underlying linguistic patterns to come to light, the methodological significance and usefulness of using corpus techniques and methodologies in textual studies would surely be debatable.

The sort of difficulty that could impede linguistic features from being extracted in sufficient numbers from corpus texts is that (1) the particular linguistic feature selected is itself in short supply in the text; or (2) that it may well occur frequently in the text, but given the current level of technical sophistication of corpus mining software (we shall discuss this in more detail in Chapter III, 3.1), it is not as yet feasible to retrieve it in sufficient numbers. In both cases, one would simply discard the linguistic feature as less useful within a corpus-oriented research scheme, and safely turn one's attention to linguistic features that are automatically retrievable and in sufficient numbers for further quantitative analysis.

Secondly, while being sufficiently abundant is an important prerequisite for problem-oriented annotation, this is not the only condition that a potential linguistic feature has to satisfy. The significance of a corpus-based investigation of parallel corpora largely depends as well on the extent to which the highlighted textual feature may reveal the nature of translated texts, or the peculiar stylistic features of translated texts (Baker, 1993: 233-52; Teubert, 1996: 238-64; Laviosa, 1997a: 289-319; McEnery and Xiao, 2002: 211-29). The term 'translationese', applied to the typical style of translated texts, understood in a narrow sense, may be closely linked to the notion of translation universals (Baker, 1993: 233-52). In this study, however, I shall use the said term 'translationese' in a broad sense, namely any systematic differences uncovered between source/target texts or among target texts within a certain comparative framework, e.g. a parallel corpus which contains the translations of a single source text at different points in time (Malmkjær 1998: 5434-41).

Such differences as may be elicited from translated texts may be reflected at various linguistic

levels, e.g. phonetic, topological (Saldanha, 2005), morpho-syntactic, lexico-grammatical (Kenny, 2000: 93-104), phraseological (Ji, 2007b), etc. They may not necessarily require or even be able to fit in the theoretical framework on a scale as large as that of translation universals (always bearing in mind that the existence of such universals is still very much in the balance), but they may be approached with good results from the perspective of related branches of applied linguistics, such as stylistics (Wynne, 2005: 59-64; Saldanha, 2007), quantitative sociolinguistics (Rickford, J. and McNair-Knox, F. 1994) or cognitive linguistics (Semino and Short, 2004).

The conceptual extension of translationese will bring substantial benefits to the enrichment of the new paradigm of translations studies as a whole, as rightly predicted by Laviosa a decade ago: "... [t]he corpus-based approach is evolving, through theoretical elaboration and empirical realization, into a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description and the practice of translation" (Laviosa, 1998: 474-9). A broader view on the nature and behaviour of translated texts will not only expand our theoretical horizon, but also encourage the study of a much wider variety of linguistic phenomena identified in translations by following the corpus-based research scheme.

This in turn will have a direct impact on what may be seen potentially as a linguistic feature for problem-oriented annotation. As can be expected, within the scope of the enriched concept of translationese, any linguistic issue may be highlighted for problem-oriented annotation in individual corpus projects, so long as it may help reveal systematic differences between the ST/TT or TT/TT pairs, as a result of their linguistic contrastiveness or time factors that would surely have a role to play when comparing two target translations of the same source at different points in time.

Once the automatic retrieval of textual data is completed, there comes the classification of data which would have to be done prior to the data extraction in most corpus annotation schemes. The rearrangement of procedures in parallel corpus construction has important implications for small-scale individual corpus projects in terms of cost-effectiveness. It may also help broaden the range and scope of questions that corpus-based Translation Studies may be able to address in the current state of language engineering.

In this respect, I believe this study offers a fairly novel way of investigating the usefulness and productivity of problem-oriented annotation in corpus-based translation studies. It has not only greatly facilitated the generation and categorization of textual data in the exploration of the parallel corpus of *Don Quijote*, which would otherwise have been quite impracticable under the current situation of automatic corpus annotation, but it has also helped expand existing areas of corpus-based translation research from being largely lexico-grammatical to a new and more challenging lexical domain, which has come to be known as multiword, idiomatic expressions (Storrer and Schwall, 1995: 35-50; Bannard, 2007: 1-8) or phraseology (Dirk, 2001, 97-107; Nuccorini, 2002; Stubbs, 1995: 23-55; 2001 and 2002:

215-44; Sinclair, 2002 and 2004²⁰; Rosamund, 2005 and 2007) in corpus linguistics or corpus-based language studies.

2.2 Automatic extraction of four-character expressions

For many corpus-based translation studies, including those in search of universals or genre- or translator-specific styles, a key factor to the success of corpus-based textual analysis is to what extent the data gathered by this methodology may be seen as large enough and sufficiently regular in their distribution so as to allow a proper quantitative analysis of the texts under discussion. Towards the end of her corpus-based study of English translations of hapax legomena found in the German source texts, Kenny does not hesitate to point out the limitations of her own conclusions on the question of normalization in English translations, which has been mainly due to the insufficiency of the corpus data she has been able to gather:

I have, however, been wary of generalizations in this and the last chapter, and have tendered explanations for observed phenomena only very gingerly. There are a number of reasons for such hesitancy. The first relates to the amount of evidence on which judgments are made: for many of the tendencies tentatively observed in GEPCOLT, the available evidence was sparse. (Kenny, 2001: 210)

To overcome the methodological drawbacks that early corpus-oriented translation scholars have experienced, one has to set in motion a full-scale investigation of the corpus texts looking for relevant, measurable linguistic information. My own research concentrates on four-character expressions (FCEXs). They have been highlighted in the current study for a number of reasons, in particular, technical convenience offered by its unique structural properties in the automatic extraction of corpus data, as well as the important pragmatic and aesthetic functions it assumes in modern Chinese (Chen, 2003: 5-9). This four-character morphological pattern is closely related to Chinese idioms, though as will be shown later in this thesis, the compositional nature of four-character phraseology is far more complex than including idioms only.

Generally speaking, given the phonetic peculiarity of Mandarin Chinese (e.g. each character has four tones) this four-character pattern lends itself well to the creation of the melodious rhythm of the language, thus adding an important poetic dimension to the language embodied in this particular syntactical structure. Secondly, the number four is a special symbol in traditional Chinese culture which tends to link phrases made up of four characters with stylistic elegance and formality. As a result, this four-character pattern is widely and purposely exploited in Chinese formal and literary writings, in which the appearance of four-character words or phrases plays a unique role in enhancing the idiomaticity and aesthetic value of the language.

²⁰ The term coined by Sinclair referring to the phraseology of phraseology is phraseogonomy (Sinclair, 2004: 177), which shows Sinclair's theoretical orientation to this topic; while other scholars like Stubbs or Rosamund have approached the issue from a more applied perspective such as quantitative studies of phraseology (Stubbs, 2002) or corpus-based descriptions of phraseology (Rosamund, 2005 & 2007)

In retrieving all the four-character segments from the two Chinese subcorpora of CCDQ, the parallel concordancer known as *ParaConc* has proved to be quite useful. Among the four searching categories provided by *ParaConc*, we have text search, regular expression search, tag search, and parallel search. The regular expression search differs from the rest in that it requires the user to encode his or her searching instructions by using a set of meta-characters. For example, the meta-string `\bsp[eo]a?k[sei]?n?g?\b` represents the search for all forms of the lemma SPEAK, where `\b` `\b` stands for word boundary; `[a-z]` for any lower case letter, `[ab]` for possible alternatives and `?` for zero or one instance of the symbol specified. In this meta-expression all the conjugation forms of SPEAK, including spoke, speaks, spoken and speaking, are encoded.

The regular expression search function may seem to be overwhelming in the first instance, since it is more abstract and synthetic than the normally used word search. This is especially the case with the exploration of texts in non-alphabetic languages, i.e., Mandarin Chinese, by making use of this particular function of *ParaConc*, since very few instructions or reference materials are available at this moment in time regarding the way to write meta-language for searching regular expressions in non-alphabetical languages.

Moreover, the establishment of a well-defined set of string operators²¹ usually requires a number of trials with different searching symbols selected in the light of their respective meta-functions. This is because a regular-expression search is basically a string search, which in looking for a certain text string throughout the corpus disregards other extra-linguistic factors of the textual context, e.g., word boundary or white space between any two alphabetical words or Chinese orthographic characters. Whilst this particular aspect of the regular-expression search function has proved to be essential in defining a set of string operators, it has also greater implication regarding the mining of segmented Chinese texts.

After the performance of segmenters of Chinese natural texts, which are word boundary unspecified, a given text string is marked by a white space left between every two words identified as a legitimate token by the Chinese Segmentation Tool. Another extra-linguistic feature of segmented Chinese texts that has proved to be important in establishing the best possible formula for the extraction of four-character items by using *ParaConc* is the white space left by CST within a segment. As a matter of fact, this kind of internal space within a segment is so easy to miss that I did not become aware of its existence until I happened to identify a working set of string operators for the search of Chinese four-character tokens quite by accident.

The searching order that I have been able to identify after a number of tests may be very simple, but it has proved to be quite efficient in separating four-character tokens from other types of segments in the Chinese corpus texts, and has therefore saved me a great deal of manual work: `\S{7}\S`

²¹ For a full list of string operators, please refer to *ParaConc Manual* at <http://www.athel.com/paraconc.pdf>, p. 59

(\S\ stands for word boundary and the digit seven refers to the length of the phrase or word). The reason behind the figure seven standing for the overall string length of a four-character token is closely linked with the somehow inconspicuous presence of white space within a segment as mentioned above. For example, in the four-character idiom 津津有味 (doing something in earnest), while each orthographic character occupies a space, there is also an additional whitespace in between each two characters; as a result, the total span of this four-character word is therefore seven digits long.

By using the regular expression search function furnished by *ParaConc*, more than two thousand four-character segments have been retrieved automatically from Liu's translation of *Don Quijote* and the reported data for Yang's translation is around one thousand and seven hundred. Scrolling quickly down the concordance lines, some mis-segmentations and four-character-long transliterations have been screened out manually to enhance the validity of the global statistics as provided in the following table. I should point out at this stage that no corpus software is perfect especially when it comes to such a difficult technical issue as Chinese text segmentation. The quantitative linguistic data that I have retrieved do tell us something interesting about the differences between the two translations; one cannot be expected to be exhaustive, otherwise she would have to fall back to the traditional manual text analysis, which in turn would be difficult to make quantitative. In the table below, raw frequency is used to distinguish between observed frequency and normalized frequency, where the latter is an important indicator for the comparison of corpora of different sizes.

Table II Global statistics of four-character segments in Yang's and Liu's translations

CHINESE SUBCORPORA	RAW FREQUENCY	NUMBER OF TOKENS	FREQUENCY PER 10K WORDS
Yang (1978)	1, 798	154,082	116.691
Liu (1995)	2, 247	149,650	150.150

A practical technique that might be used here to assess the significance and frequency of occurrence of FCEs in Yang's and Liu's translation is to compare these normalized frequency of FCEs with the relevant statistics extracted from a larger comparable corpus, ideally composed of Chinese translation texts produced during the same period of time. Such a comparison, however, is not really possible given the different phraseological categories highlighted in my parallel corpus CCDQ as against those encoded in Chinese translational corpora already online.

Most of existing Chinese monolingual or parallel corpora have been constructed and annotated by means of the standard character marking-up scheme, i.e. Peking University Tagset (PUT)²², which covers fifty word classes, ranging from purely grammatical items, e.g. parts of speech (POS), to distinctive lexical or semantic categories, such as idioms (i) and fixed expressions (I). Despite the wide coverage of the PUT system, it does not single out four-character expressions as an individual item for tagging purpose. As will be shown in the next section, the term four-character

²² For Peking University Tagset, see http://bowland-files.lancs.ac.uk/corplang/babel/PKU_tagset.htm

expressions refers to a whole range of Chinese phraseological types, including among others, Chinese idioms, fixed expressions, morpho-syntactically patterned short phrases or words, foreign transliterations, and others.

That is, although there is certain overlapping between Idioms (i) and Fixed Expressions (which also tend to be in four-character patterns), as two constituent elements of the PUT system, and four-character expressions as a whole, the latter embodies a much larger and complex phraseological domain in Chinese lexis. Therefore, the statistics provided by a comparable corpus, i.e. the Babel English-Chinese Parallel Corpus²³, which has also been built under the PUT scheme, can hardly offer any significant information regarding the distribution characteristics or frequency of occurrence of FCEs at an average level.

Table III displays the normalized frequencies of two closely relevant phrasal categories, idioms and fixed expressions in the Chinese comparable translation corpus. They make up two important sub-categories within the phraseological domain of four-character expressions. For reasons explained above, a comparison of the occurrence of FCEs in the Chinese target texts of CCDQ and the occurrence of idioms and fixed expressions in the Babel translational corpus cannot possibly yield any direct information regarding the quantitative significance of FCEs in the two Chinese translations of *Don Quijote*.

Nevertheless, if we produce the normalized proportions of Chinese idioms and fixed expressions in the comparable corpus, the result may give us an approximate idea of the percentage represented by these highly conventionalized phrasemes within the confines of the broader phraseological category of four-character expressions. That is, instead of using the comparable corpus to estimate the relative significance of the linguistic feature in the two Chinese target texts, which is not technically plausible in my study, the comparable corpus has been exploited as an indicator of the internal complexity of the composition of four-character expressions.

Table III Idioms and fixed expressions in the Babel English-Chinese Parallel Corpus

PHRASAL CATEGORY	RAW FREQUENCY (TOKENS)	FREQUENCY PER 10K WORDS
Idioms	936	32.6
Fixed Expressions	785	27.3
Total	1, 721	59.9

Table III shows the normalized frequency of idioms and fixed expressions in the comparable corpus draws closely to sixty. This roughly represents half of the frequency of four-character segments that have been automatically retrieved from Yang's and Liu's work, respectively (Table II). This in turn seems to suggest that idioms and fixed expressions, which in their majority are four-character long in

²³ The Babel Parallel Corpus, which contains a total of 287, 412 Chinese words, is freely accessible at <http://bowland-files.lancs.ac.uk/corplang/babel/babel.htm>.

Chinese, only make up around half of the phraseological category of FCEs. To have a closer view of the inherent complexity of FCEs, we shall proceed to conduct a pilot typological study of Chinese four-character expressions, which will draw upon both previous studies of Chinese idiomatic expressions and a close inspection of quantitative linguistic data extracted from the parallel corpus CCDQ.

2.3 A pilot classification of Chinese FCEs

As Baroni & Bernardini point out, in the current state of affairs in the field of corpus linguistics, there seem to be two major paths to textual feature categorization with the assistance of computational linguistic tools. First, there are “linguistically sophisticated methods”, such as the one developed by Biber (1995) for register analysis, which entails “a detailed tagging/parsing system with human post-editing and disambiguation; ... the selection of meaningful linguistic features and ... the functional interpretation of the dimensions underlying the various factors, in order to develop a detailed multidimensional analysis based on a large set of co-occurring lexico-syntactic features” (Baroni and Bernardini, 2006: 262).

Secondly, there is the possibility of machine-learning text categorization, which is characterized by the study of a few popular choices of textual features, including function words, textual statistics (e.g. average sentence length, lexical richness measures), and knowledge-poor surrogates of a full syntactic parse, such as n-gram and part-of-speech (POS) information (Ibid). The model for textual feature retrieval and categorization developed in this study shows some similarity with the ones described by Baroni as “linguistically sophisticated methods”, in that the encoding and interpretation of highlighted textual features, i.e. Chinese four-character expressions (FCEs), are placed within a linguistically rich analytical context as furnished by the typological classification of Chinese FCEs to be introduced shortly.

This unique four-character pattern in Chinese phraseology has always been connected with Chinese idioms or *Chengyu* as we say in Mandarin Chinese. It is estimated that in modern Mandarin Chinese, the proportion of four-character items in idiomatic expressions may be as high as ninety percent or even more (Xu, 2006:108). Consequently, in many cases people tend to identify (quite frequently, unconsciously) the concept of four-character expressions with *Chengyu*, or may even treat the two terms as freely exchangeable options.

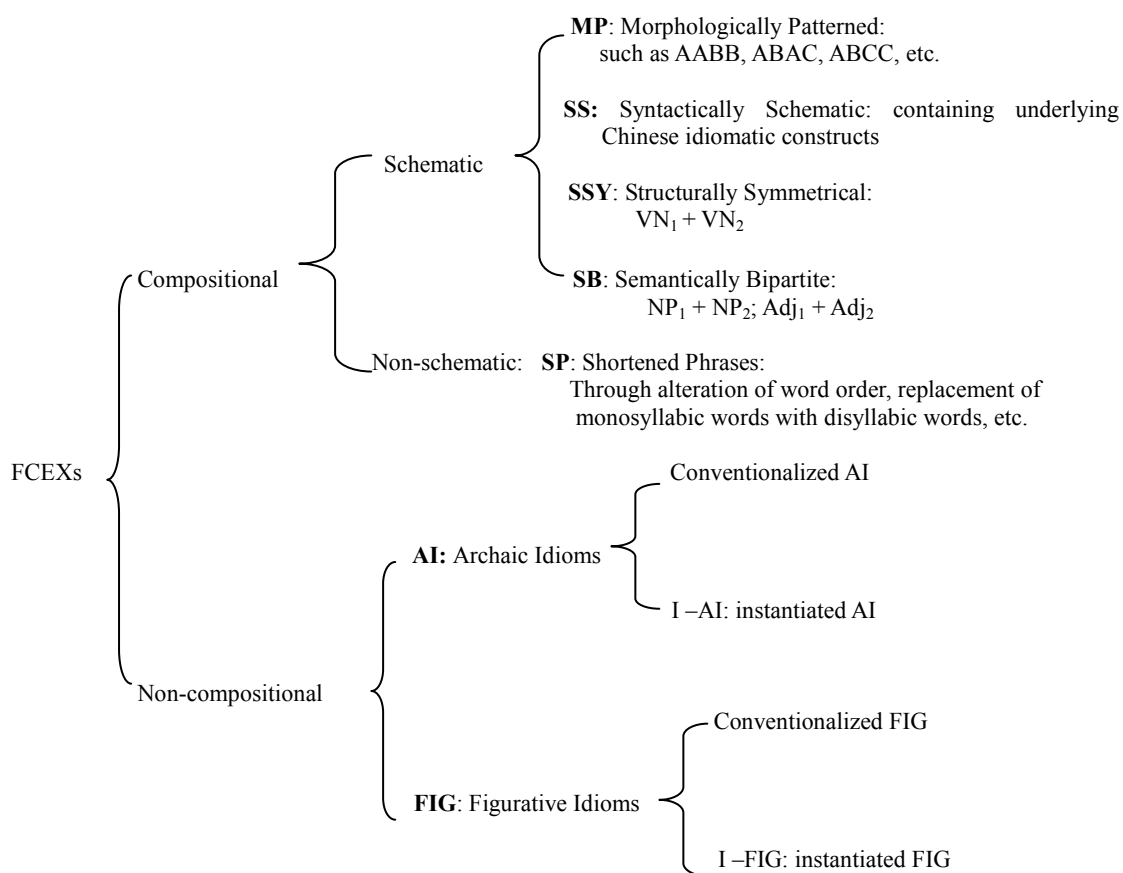
However, such conventionalized notion can hardly reflect the compositional nature of four-character expressions, a very complex phraseological category in Chinese representing a heterogeneous linguistic system of phrasal units centred upon this underlying morpho-syntactic structural pattern of Chinese. Despite the fact that not much research has been done on the etymological origin of this generalized four-character pattern, it is widely taken for granted that the said four-character pattern is a perfect manifestation of the prosodic musicality, structural equilibrium or symmetry, and semantic

succinctness of Chinese (Xu, 2006: 107). In this sense, we may regard this four-character pattern of writing as a widely favoured and highly flexible rhetorical device very typical of Chinese phraseology.

For the purpose of my study, whose main objective is to detect and analyze the stylistic use of four-character expressions in two recent Chinese versions of *Don Quijote*, I propose the following classification of four-character expressions based upon the lexico-syntactical configuration of some four thousand four-character items extracted from the two Chinese translations under consideration (see Diagram II).

Due to the definitional complexity and possible confusion implicit in any classification, I have decided to use simple, descriptive terms. Popular categories of phraseology such as idiomatic and formulaic expressions, collocations, fixed or semi-fixed expressions, tend to be rather elusive in the actual practice of manual classification and may not necessarily cover the same ground as do four-character expressions in Chinese. By using descriptive terms to differentiate between various types of Chinese four-character expressions, one would expect to minimize the attendant dangers of subjectivity when processing data.

Diagram II A classification of Chinese four-character expressions retrieved from CCDQ



According to their structural configuration, four-character expressions are classified into two major categories: compositional and non-compositional FCEs. The dividing line between

compositional and non-compositional phrases is drawn in terms of the semantic compositionality of an FCEX. An FCEX is seen as compositional where, according to the intuition of the researcher or a native speaker of Mandarin Chinese, the breaking down of an FCEX into a sequence of individual characters does not affect the meaning of the FCEX as a whole. On the other hand, if the splitting of an FCEX may lead to a substantial change in the meaning it has, it is classified as a non-compositional FCEX. Compositional FCEXs can then be further divided into schematic and non-schematic subclasses depending on whether there is an underlying structural or syntactical pattern within a compositional FCEX. Chinese schematic FCEXs are highly flexible in that they may have several formal representations. In my examination of the four-character segments extracted from the two Chinese translations, I have been able to identify four recurrent types of schematic FCEXs: morphologically patterned words (MPs), syntactically schematic phrases (SSs); structurally symmetrical phrases (SSYs) and semantically bipartite phrases (SBs).

Non-schematic FCEXs mainly refer to shortened phrases which may be obtained by means of word order alteration, and/or by substituting monosyllabic words in ancient Chinese for disyllabic words in modern Chinese, etc. Non-compositional FCEXs include Chinese archaic idioms and figurative idioms where each subcategory contains both their conventionalized representations and instantiated variants. I should point out at this stage that this taxonomy is not intended to cover the compositional range of four-character expressions in Mandarin Chinese as a whole, but rather to test its validity within the confines of four-character items frequently used in the two modern translations under investigation.

2.4 Redefining Chinese Idioms

A very important compositional part of four-character expressions is Chinese idioms. As a quick comparison of the figures provided in Table II and III shows, idioms make up between one quarter and one fifth of the entirety of Chinese four-character expressions. On the other hand, more than ninety per cent of Chinese idioms are composed of four characters. The structural resemblance has caused substantial confusion between the two phraseological concepts, as they have been taken as two largely exchangeable notions by many Chinese people. To further my investigation on the phraseological profiles of the two Chinese translations, I should first clarify such misconception, which in turn prepares the ground for the further classification of the linguistic data retrieved from the parallel corpus.

Idioms, as different from other phraseological categories in my study, are those Chinese four-character expressions whose morpho-syntactic features show some aspect of idiomaticity (Moon, 1998: 6-9). The concept of idiomaticity is no less complex and elusive than that of conventionality, and yet we require some defining properties which will allow us to draw some sort of provisional defining line between idiomatic phrases and loosely combined lexical units. However, it is quite impossible to come to a definition of what an idiom is or to what extent a particular expression may be said to be

‘idiomatic’.

Most attempts at defining idiomatic expressions end up being circular with the defining properties that make an idiom tending finally to hark back to the question of idiomaticity. Nunberg et al. (1994: 491-538) highlight six defining properties of idioms ranging from central to more peripheral, i.e. conventionality; inflexibility; figuration; proverbiality; informality and affect. In my study, such definition of idiomaticity is re-examined and further modified in light of the behaviour of idiomatic expressions found in Chinese corpora.

A recent and popular definition of Chinese idioms is the one given by the Chinese linguist Shi Shi. Shi defines as idiomatic, “those expressions that have been long conventionalized in Chinese; and normally with fixed structural patterns and compositional constituents; they usually have specific meanings which cannot be inferred from their literal meanings” (Shi, 1979: 35). Shi’s definition of Chinese idioms highlights the first three dimensions of idiomaticity in the study of English idioms by Nunberg et al. Taking this working definition as my point of departure, I shall proceed to show that the notion of idiomaticity is a very complex issue that requires therefore a very discrete and dynamic description of idiomatic expressions in real life language which has not as yet being provided, at least in Mandarin Chinese.

2.4.1 Idiomatic variants

Shi’s definition of Chinese idioms also brings into focus the entrenched morpho-syntactical structures, as well as the fixed semantic components that seem to characterize many Chinese idioms. Like the notion of conventionality, the second point made in Shi’s definition of Chinese idioms is largely prescriptive and barely allows any definitional flexibility. For instance, it does not account for important linguistic facts with respect to the dynamic behaviour of Chinese idioms in real communicative contexts, e.g. the phenomena of Chinese idiomatic variants (Xu, 2006).

As a matter of fact, the topic of idiomatic variants has tended to be deliberately avoided, not least on account of the remarkable compositional or structural adaptability of many idiomatic expressions in real contextual situations which would turn out to be too complex to be explained within the theoretical framework of most traditional grammars. The highly flexible use of Chinese idioms is mainly reflected in the number of syntactical forms they can take on, which may be classified according to the way these variants occur, i.e. by abbreviation, by extension and by changing the order of words as follows:

- (1) Abbreviation refers to the way in which a Chinese idiom, usually structured in a four-character pattern, is condensed by the extraction of certain monosyllabic morphemes, which sometimes are adjacent and sometimes separated by other morphemes located in between.

For example, the use of 蛇-足 ‘serpent-feet’ as a condensed expression of the idiom 画蛇添足 ‘draw-serpent-add-feet’ which is a metaphorical expression meaning doing things twice or repeating things, is realized through the extraction of the second and fourth noun morpheme, 蛇 ‘serpent’ and 足 ‘feet’, from the other two symmetrical verbal morphemes, 画 ‘draw’ and 添 ‘add’. The reason for the retention of the noun morphemes, instead of the verbal morphemes of the idiom, can be tentatively deduced from the fact that the figurative meaning of the whole idiom, which is an ironic reference to redundant actions, is already implicit in the contradictory noun morpheme pair of 蛇-足 ‘serpent-feet’.

As for the extraction of adjacent morphemes in an idiomatic expression, a good example is the use of 敌忾 ‘enemy-resist’ as the abbreviated expression for the idiom 同仇敌忾 ‘together-hostile-enemy-resist’. The structure of this idiom can be divided into two semantic parts. As suggested in the English translation of each morpheme only when people share the same hostile attitude towards a certain target, which is the literal meaning of the first two morphemes, can they join together to resist it as a common enemy, which is the literal meaning of the last two morphemes. As a result, the use of the last two morphemes 敌忾 ‘enemy-resist’ as the elliptical expression of the whole idiom is probably based on the notion that the meaning of the first two morphemes may be readily inferred from the last two. Similar idiomatic abbreviations are 劳燕 ‘lao (a kind of bird) and swallow’ in 劳燕分飞 ‘lao-swallow-separate-fly away’; 黔驴 ‘Qian (a place) -donkey’ in 黔驴技穷 ‘Qian-donkey-tricks-exhausted’; 鳌头 ‘turtle-head’; in 独占鳌头 ‘only-occupy-turtle-head’; 效颦 ‘mimic-bend one’s brows’ in 东施效颦 ‘Tongshi-mimic-bending one’s brows’; 告老 ‘retire’ in 告老还乡 ‘retire-return to one’s hometown’; 家长 ‘family-stories’ in 家长里短 ‘family-stories-community-news’; 蓬荜 ‘shabby house’ in 蓬荜生辉 ‘shabby-house-shine’.

- (2) Extension refers to the process of morpho-syntactical variation by which the underlying structure of a Chinese idiom is maintained while the addition of modifying or restrictive elements turns the normally four-character long idiom into an extended idiomatic expression.

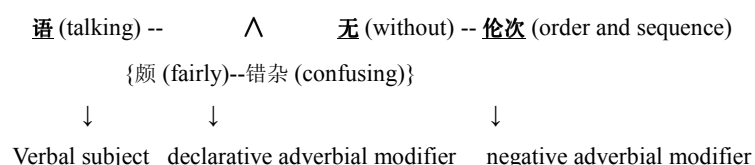
This particular method by extension may be compared to a similar derivational process in English whereby parts of the meaning of an idiom may be adjusted by means of adjectives or relative clauses, e.g. *leave no (legal) stone unturned* or *John got the job by pulling strings that were not available to anyone else* (Nunberg, et al, 1994: 500). To illustrate this specific type of Chinese idiomatic-variant formation, three representative samples are given below where flexible use in a real (authentic) communicative context brings to light a subtle gradation in structural versatility.

In the first example of idiomatic extension 语-颇错杂-无伦次, the Chinese four-character idiom 语-无伦次²⁴ (talking-without-order-sequence), which is normally used to describe someone in a disorderedly state of mind, is modified by the insertion of a short phrase 颇错杂 between the first and second morphemes. Assuming that the syntactic modifications of idioms have been made possible

²⁴ A hyphen is placed among parts of idioms or short phrases to facilitate the understanding of their syntactic structures or grammatical relations: as in 语-无伦次, a hyphen is introduced between subject and modifier of the idiom; and in 颇-错杂, a hyphen is used to separate adverb from adjective.

largely due to the semantic instead of the syntactic relations held by their constituent parts (Nunberg, 1994: 505), we could argue for the rationality of the insertion of the short phrase 颇-错杂, which means ‘fairly confusing’, within the underlying structure of the idiom as {subject + modifier}, since such a syntactic extension constitutes no more than a qualitative extension of adjectival morphemes without in any way affecting the figurative connotation of the idiom.

Diagram III Idiomatic variation – extension (1)

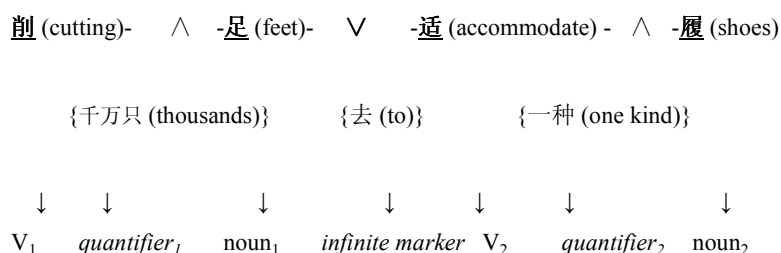


As shown in the Diagram III, since the declarative adverbial modifier, which is to be inserted immediately after the subject as a verbal noun meaning ‘talking’, carries the same semantic weight as the negative adverbial modifier which forms part of the idiom, it follows that such a syntactic modification will not affect the meaning of the whole idiom (talking without logic or order) but it helps to draw attention to the specific quality of the subject.

The second example of Chinese idiomatic extension may be considered quite analogous to some English idiomatic changes characterized by the quantification of certain parts of an idiom which hardly affect the figurative reference of the idiom as a whole, like in the expression *to touch a couple of nerve*, where the addition of *a couple of* to the idiomatic part *nerves* may be restricted to modify the degree of the subject noun of the VP idiom *touch a nerve* (Nunberg, 1994: 501).

The Chinese four-character idiom selected in this case is 削-足-试-履 (cutting - feet - trying on - shoes), which means either cutting one’s feet to fit the size of shoes, literally, or being inflexible with the reality of the situation, figuratively. The extended phrase which contains the particular idiom as its underlying structure is 削 - 千万只 - 足 - // 去 - 适 - 一种 - 履²⁵, and the semantic relations between parts of the idiom are analysed in the following diagram:

Diagram IV Idiomatic variation – extension (2)



²⁵ In this case, a double slash is used to separate VP clauses within a single short phrase.

From the illustration above we can see that the syntactical structure of the idiom 削足-试履 as $VN_1 + VN_2$ is kept intact while the process of idiomatic extension unfolds at two semantic levels. Firstly, it is the quantification of two noun morphemes of the idiom: 足 ‘feet’ is modified by 千万只 ‘thousands’ and 履 ‘shoe’ is restricted by 一种 to a certain kind of mode. Obviously, in this case, the use of somehow contrastive quantifiers to modify the noun idiomatic morphemes serves to accentuate the ironic tone of the figurative meaning of the idiom.

Secondly, the seemingly trivial addition of the infinitive marker between VN_1 and VN_2 represents a significant move towards loosening the entrenched syntactic structure of the idiom to a freely composed short phrase. It must be noted here that the stretching of the syntactic structure of the idiom is not pursued at the expense of the figurative meaning of the idiom; on the contrary, the introduction of the word 去 ‘to’ highlights the implicit purpose-action relation between the two VN parts and thus turns the highly condensed four-character idiom into a relatively freely composed phrase with overt semantic relations among the two actions involved, namely 削足 ‘cutting one’s feet’ and 试履 ‘to fit the size of the shoes’.

In fact, it would seem that, among the main instances of the flexible use of Chinese idioms, the only option available to the syntactic variants of the idiom 削足-试履 is the extension process. We could try to reproduce the elliptical expressions of this idiom by imitating the process of idiomatic abbreviation as happens with the flexible use of the idiom 画蛇-添足 and 同仇-敌忾 explained above. Then, the three likely combinations of idiomatic morphemes would turn out to be 削足 ‘cutting—feet’, 试履 ‘try on- shoes’, and 足履 ‘feet—shoes’: while the first two belong to the category of the extraction of adjacent morphemes as 敌忾 in 同仇-敌忾, the last morpheme pair recalls the abbreviated expression of 蛇足 for the idiom 画蛇-添足.

A closer look at all three of the suggested modes of abbreviating expressions of the idiom 削足-试履 brings me to the conclusion that none of them can be used as an adequate way of reducing the idiomatic chunk, since the semantic content of each morphemic pair only represents part of the literal meaning of the idiom, and without the presence of other idiomatic parts, there is no foundation on which to build up the figurative meaning of the idiomatic expression as a whole.

The third example of idiomatic extension exhibits the maximum level of flexibility that may be achieved in semantic relations among parts of a Chinese idiom by bringing in additional elements to form part of the extended syntactic structure of the idiomatic expression. The idiom selected here is 鱼目-混珠 ‘fish eyes - mixed with-pearls’, which means passing off fish eyes for pearls or passing off a fake for the real thing, in its figurative sense.

Quite evidently, understanding any idiomatic expression in any language would be dependent to a great extent upon the level of familiarity with the appropriate context. That is, if the constituent

parts of an idiom can be subsumed within the textual context, it would signal correspondences between the components of the structured denotation of the context and the various parts of the idiom; and this in turn would become an essential prerequisite for the realization of the kind of idiomatic extension possible as exemplified in 鱼目-混珠²⁶:

- (i) 也许有人担心, 允许人人可以参加考核, 鱼目 ‘fish eyes’ 如果 混 ‘mixed with’ 珠 ‘pearls’ 怎么办?
- (ii) 其实不怕, 只要严格掌握考核这一关, 鱼目 ‘fish eyes’ 是 混不进 ‘will not be mixed with’ 珠群 ‘groups of pearls’ 的.
- (iii) 现在怕的是, 正当国家急需 珠宝 ‘treasures’ 的时候, 珠宝 ‘treasures’ 却被埋在 鱼目 堆 ‘pile of fish eyes’ 里.

- (i) Someone may worry about the fact of allowing everyone to attend the exam -- what if “fish eyes” are passed off for “pearls”? (Passivization of the idiom)
- (ii) Actually it does not matter; provided that the selection process is strictly controlled, “fish eyes” would not be passed off for “pearls in groups”
(Passivization of the idiom and quantification of the idiomatic noun morpheme).
- (iii) What is afraid of is that when the State needs “treasures” (disyllabicalization), “treasures” are buried in piles of fish eyes.
(Free use of parts of idioms; quantification and disyllabicalization of idiomatic noun morphemes)

The three sentences listed above appear in a single paragraph and are selected from a news report published in the official organ of China *People's Daily* in 1999. As we can see, the idiomatic extension of the idiom 鱼目-混-珠 is reflected in its disintegration into a larger textual context, which definitely goes beyond the boundary of sentence or clause by including passivization of the idiom and displaced use of the idiomatic parts, namely 鱼目 ‘fish eyes’, 混 ‘mixed with’, and 珠 ‘pearls’, as well as the quantification and disyllabicalization of the idiomatic noun morphemes, i.e. 鱼目 ‘fish eyes’, and 珠 ‘pearls’. This has been made possible by virtue of the favourable context, which helps the readers to establish correspondences between two sets of entities, namely the qualified candidates and pearls; and the non-qualified candidates and fish eyes.

- (3) Another interesting feature in relation to syntactic or structural variation in Chinese idioms is the change in word order.

Possible reasons that may account for the sequential alteration of idiomatic morphemes are, firstly, specific phonetic requirements in Chinese poem writing, like in [取义成仁] 今日事, 人间遍种自由花 ‘Today I will give my life to our enterprise of saving our country from foreign invasion, which will give millions freedom’, the shift of the first and second half of the idiom 成仁-取义 ‘save enterprise – give life’ has been simply due to the need of maintaining the phonetic prosody of the verses. Secondly, changes in word order in the flexible use of Chinese idioms can also be motivated by pragmatic needs. Like the alternation of 狗 (dog) and 人 (man) in the idiom 狗-仗-人-势 (dog- rely on- man- power), which

²⁶ Source text is selected from *People's Daily*, March 1, 1999

means relying on one's power to bully others, can be used to express the speaker's sarcastic attitude towards the man in a privileged position.

2.4.2 Idiomaticity and figurativeness

In defining Chinese idioms, Shi made a third observation: "... [they] usually have specific meanings which cannot be inferred from their literal meanings" (Shi, 1979: 35). It sounds common sense; furthermore, it is widely accepted among linguists of different persuasions as one of the main attributes of idiomatic expressions. But it is, nevertheless, arguable, especially when referring to Mandarin Chinese, since a large number of Chinese idioms simply do not call for a figurative interpretation. In other words, the semantic composition of such idioms may be literally mapped onto their constituent parts.

The Chinese linguist Tao was the first to bring to the fore idiomatic expressions such as 空前 - 绝后 'rare before – extinct in the future' or 移风 - 易俗 'change habits – alter customs', which do not hold any figurative meaning, but are still treated as legitimate items in most Chinese dictionaries. Tao starts his typological study of Chinese idioms by closely observing these non-figurative Chinese four-character phrases which he generally refers to as Chinese 'archaic idioms'. He points out that Chinese idiomatic expressions encompass a whole range of long established phrases: some of them have a figurative dimension of meaning, while others do not, and their diachronic status in the Chinese lexicon is precisely reflected in the fact that they contain morphemes or words from ancient Chinese. Consequently, Tao proposes a taxonomic approach to idiomatic expressions which brings together these non-figurative idioms as parallel idioms alongside generally recognized figurative idioms.

Chinese archaic idioms are defined by Tao as long-established idiomatic expressions that do not involve usually an element of figurative meaning and thus are semantically more transparent when compared with Chinese figurative idioms. Nevertheless, as Tao himself suggests, being non-figurative is not the same as being semantically transparent. That is because the syntactical composition of archaic idioms normally implies by definition the use of archaic functional or content words, as well as grammatical structures very typical of ancient Chinese. As a result, without a good knowledge of the lexicon and syntax of ancient Chinese, it would be quite impossible to understand these archaic idioms, even for a native speaker, for ancient Chinese and modern Chinese are two very different linguistic systems.

2.5 A classification of figurative idioms (FIGs) and archaic idioms (AIs)

2.5.1 Figurative idioms

To facilitate the manual retrieval of idioms from automatically extracted FCEs, I propose a separate typological framework for figurative idioms and for archaic idioms, both largely drawn on

previous studies in the field (Tao, 2002: 130-37), and on the evidence collected from large-scale Chinese corpora (Ji, 2008b: 67-82). Important structural adjustments, however, have been made in order to minimize any inconsistency that may inevitably arise from human intervention in the actual practice of categorization. In fact, given the current state of affairs in the art of natural language processing, a hybrid model which combines both computer-assisted corpus data extraction and human-judgement-based lexical knowledge representation seems to be a most sensible strategy in linguistic classification and annotation (Ooi, 1998: 80-2).

One could say first that a figurative idiomatic expression in Mandarin Chinese, or FIG for short, is a type of conventionalized phrasal unit that may include, among other figures of speech, metaphor, metonymy, simile, synecdoche and hyperbole in their semantic configuration. Based precisely on the semantic representation of the figurative meaning of FIGs, Tao (2002: 134-5) has identified six major categories of Chinese figurative idioms.

First, a figurative Chinese idiom may represent the target B in a typical metaphorical relation: (A) → B, where A is the source and B is the target. Here the whole figurative item may well be a concrete object, person, state, process like 晴天霹雳 ‘thunder on a clear day referring to unexpected accidents’; 冰清玉洁 ‘the purity of ice and jade which refers metaphorically to one’s innocence, especially young females’. Secondly, a figurative idiom could appear as a complete simile in itself marked by the linking word 如 or 似 and meaning *like* or *as*; for example, 福如东海 ‘fortune as massive as the east sea’, 似水年华 ‘past time as running water’.

Thirdly, there are metaphorical idioms in Mandarin Chinese, where the realization of the figurative meaning of the item is attained via a number of functional connections between the source A and the target B, such as the structure of {<m.> B (adj.) + <s.> A (n.)} in 风烛残年 ‘remaining time of one’s life as the burning candle in the wind’. The symbol of <m.> refers to the semantic function of the target B as metaphorical; <s.> refers to the source A. In this idiom, the metaphorical part B 风烛 ‘burning candle’ is an adjective; it is used to modify the subject or the source A 残年 ‘the remaining part of one’s life’. Similarly, in 笔走龙蛇 ‘calligraphy as the move of serpent or dragon’, the grammatical relationship between the metaphorical part and the source part may be described as {<s.> A (n.) + <m.> B (adj.)}, where 笔走 ‘the moving of the brushes’ is a noun and it is modified by the metaphorical expression 龙蛇 which is a noun used as adjective referring to the creeping movements of dragon and serpent.

According to Tao, the fourth category of figurative idioms refers to some relatively more complicated FIGs, whose semantic structure may also be approached as bipartite, analogous to that of the metaphorical FIGs described above. The configuration of their figurative meaning, however, is less conspicuous in that the source A is apparently absent from the underlying semantic structure of the idiom, which comprises the target B and the adjective or adverbial element modifying the metaphorical part B, like 饱经 ‘have undergone ... sufficiently’ in 饱经风霜 or 烂熟 ‘very familiar with’ in 滚瓜烂熟. It is

worth mentioning that the interaction between the two parts that constitute the underlying structure of a figurative idiom, i.e. the metaphorical part and the modifying part, is largely grammatical and subject therefore to a series of syntactic rules; and that the establishment of the semantic linkage between the target and the implicit source is to a great extent obtainable only pragmatically. Such pragmatic association is sometimes remarkably context-dependent, like the conceptual pairing of 宝刀 ‘treasured sword’ with ‘very experienced person’ in 宝刀不老, and sometimes, context-free or: in other words, institutionalized like the conventionalized use of 风霜 ‘wind and frost’ to refer to accumulated personal experience.

Next, hyperbole is also a well-exploited rhetorical device in the semantic composition of Chinese idioms, where the figurative meaning of the various items derives from exaggerating the actual quality of the source entity, e.g. 火冒三丈 ‘rage of three zhang²⁷ high’; 花天酒地 ‘sky covered by flowers and ground covered by wine; excessive indulgence’. Last, we have Chinese figurative idioms whose figurative meaning is activated by reference to the origin of the allusion, which may be historical events, personal names, connotations, literary quotations, to name but a few. Examples of figurative idioms under this subclass are 筚路蓝缕 ‘exploring unknown territories with crudely made cart and very poor cloth; extreme hardship’; 弓杯蛇影 ‘the reflection of the arch in the glass resembles the shape of a snake, which thus may unnerve the people drinking the wine held in the glass’, etc.

As shown above, with the exception of the last type, whose figurativeness refers back to the origins of the alluded event or name, all the other five categories are defined strictly in line with the use of different rhetorical tropes in their semantic representations. Figurative idioms grouped under the fourth subclass show some irregularities as regards their marked semantic-syntactic features; however, they can still be treated safely as a kind of metaphor, where the absence of the source can be reasonably predicted in a fully informative context or in some cases, provided that they are well and long established in Chinese, a proper understanding of these source-unspecified idioms does not even require any special efforts from the native speaker to contextualize them, since the reading of these idioms at both literal and figurative levels has long been an established practice.

2.5.2 Archaic idioms²⁸

Chinese archaic idioms are idiomatic expressions that usually do not involve a figurative element, and so they show a different type of semantic compositionality when compared with most Chinese figurative idioms. Nevertheless, for Chinese archaic idioms, being non-figurative is not the same thing as being semantically transparent. That is because the morpho-syntactic configuration of these archaic idioms entails by definition the use of archaic functional or content words, as well as syntactical structures typical of ancient Chinese. As a result, without a good knowledge of ancient

²⁷ Zhang is a traditional measure of length; a Zhang is equal to 2.35 meters.

²⁸ Archaic idioms in the Chinese translations of *Don Quijote* will be studied in detail in the following chapters, which compare the use of archaisms in the source and two target texts. Further explanations of archaic idioms are provided in Chapter IV and V.

Chinese lexis and syntax, it would be very difficult to understand these archaic idioms, even for a native speaker, since ancient Chinese and modern Mandarin Chinese are two very different linguistic systems. According to the Chinese linguist Tao, there are at least seven subtypes of Chinese archaic idioms (Tao, 2002: 135-137):

- (1) Archaic idioms that contain content words and their structures in ancient Chinese, e.g. [匪夷] 所思 ‘perplexing’, [安]步当车 ‘walking steadily as sitting inside of a car’, 不[经]之谈 ‘not serious matters’, 博闻强[识] ‘extremely good at memorizing things’, 长[歌]当哭 ‘reading poems with emotion as crying’.

Content words refer to a broad range of lexemes which include nouns, adjectives, verbs, etc. These content words are in opposition to other kinds of functional items like prepositions, conjunctions, which constitute another major word category, i.e. functional words. Orthographically speaking, content words in ancient Chinese (written in simplified Chinese) show a striking affinity with content words in modern Chinese, and it is hence quite difficult to distinguish between the two solely based on their formal appearance in writing. There is however an efficient way in which we may reasonably make out the difference between the two without having to resort to referential material all the time.

The way to ascertain the diachronic status of a content word is to examine how it functions syntactically in a given textual context. That is because in modern Chinese the functional role of a content word is relatively fixed when compared to its counterpart in ancient Chinese. For example, if a content word works as an adjective, then it should serve as an adjective in a consistent manner, no matter the textual context in which it may appear. This, however, is not the case with ancient Chinese, where the grammatical roles of content and functional words remain still largely unmarked.

For example, among the archaic idioms listed above, the morphemes shown in squared brackets all exhibit certain syntactic underspecifity when understood within the context of ancient Chinese. In [安]步当车 ‘walking steadily as sitting inside of a car’, the first morpheme 安 which is normally used as an adjective is now used as an adverb to modify the second morpheme serving as the main verb. Similarly, in 长[歌]当哭 ‘reading poems slowly and with emotions as crying’, the second morpheme 歌 may be roughly equalled with the lemma *sing* and is usually used in its sense as a noun *song*; however, it is exploited in this case as a verb. Such shifting in syntactic functionality as generally seen in ancient Chinese is probably due to the fact that Chinese characters are not inflected, as a result of which the grammatical function of a character-word is hard to tell from the morphological changes at the end of a word.

- (2) The functional under-specificity of content words in ancient Chinese gives rise to another important linguistic feature known as the functional flexibility of content words in ancient Chinese.

Words with a variable syntactical function were typical in ancient Chinese by virtue of the

functional flexibility provided by the morphological structure of monosyllabic words; the interchange of different functions embodied in a single word was fairly normal. The fact that the syntactical functions of content words in ancient Chinese may vary according to the context is supposed to be conditioned by the scarcity of content words with a discrete semantic reference. Such peculiar linguistic behaviour usually only takes place in the early stages of language development. The most frequent types of content words with varying syntactical functions are nouns or noun phrases used as verbs and situational adverbs; also, adjectives may be used as verbs, and verbs and adjectives used as nouns.

Like in the following idioms, which share the same structure as [朝 X 暮 X]: 朝-思-暮-想 ‘think about something day and night’; 朝-欢-暮-乐 ‘enjoy something day and night’; 朝-生-暮-死 ‘born in the morning and die in the evening’, the fixed first and second noun morphemes which mean morning and afternoon respectively, are used as adverbs to describe the time in which the two actions as represented by the second and fourth morphemes take place; or in 厚 ‘thick’ -此 ‘this’ - 薄 ‘thin’ - 彼 ‘that’; 安 ‘settle’ - 邦 ‘country’ - 定 ‘stabilize’ - 国 ‘state’, where the highlighted adjective morphemes in each case are used as verbs in their respective material sense. Consequently, the underlying structure of the two idioms may be illustrated as VN₁ + VN₂. Based on observations like this, a second type of Chinese archaic idioms has been identified by Tao which contain words characterized by their shifting syntactic functions in ancient Chinese, e.g. [祸]国[殃]民 ‘harm – country – imply – people’ (n. → v.); 不[期]而遇 ‘without – date – meet’ (n. → v.); 骇人[听][闻] ‘astonishing – hear – notice’ (v. → n.); 敬而[远]之 ‘respect – and – far – personal pronoun’ (a. → v.); [厚]此[薄]彼 ‘thick – this – thin – that’ (a. → v.)²⁹.

It is interesting to note in this regard that in identifying the first two types of Chinese archaic idioms, Tao’s definition has basically drawn upon the unusual syntactic behaviour of ancient Chinese, i.e. the instability in the grammatical functions assumed by certain content words/ morphemes in ancient Chinese. In the actual categorization of four-character items retrieved from the parallel corpus of *Don Quijote*, these two types of archaic idioms may therefore be conflated and classified under a single category as archaic idioms with content morphemes showing changing syntactical functions, which are a consequence of the morphological singularities of ancient Chinese.

- (3) There are archaic idioms which contain functional words or pronouns and their structures in ancient Chinese, such as 反求[诸] ‘like’ 己, [诸] ‘like’ 如此类, 得其所[哉] ‘final exclamatory particle’, 长[此]以往 ‘this’, 多[此]一举 ‘this’.

Highlighted characters are equal to 这 in modern Chinese, which means “this” or “these”. Things turn out to be easier with the detection of archaic idioms as defined by the use of purely functional words in ancient Chinese. This is because when compared with content words in ancient Chinese, the number of functional words found in ancient Chinese is rather limited. In differentiating between content words in modern Chinese and their counterparts in ancient Chinese, one would still

²⁹ a. for adjective; n. for noun; adv. for adverb; v. for a verb.

need to have a closer look at their syntactical behaviour in the actual context.

However, when it comes to the identification of archaic idioms constructed with purely functional words in ancient Chinese, this can be easily done by consulting a dictionary of ancient Chinese grammar. The most frequently used functional words in ancient Chinese amount to no more than a handful of monosyllabic morphemes: 之、其、者、所、或、莫、焉、诸、曾、且、乃、非、勿、毋、弗、于、以、为、与、因、而、则、虽、然、唯（惟）、盖、夫；也、矣、乎、焉、哉；欤（与）邪（耶）、耳.

The functional roles that these words play in ancient Chinese are not difficult to discern in texts, since most of them are no longer in use in modern Chinese; and their syntactic roles are substituted by another set of functional words in modern Chinese, which are mainly disyllabic words such as 的, 得, 地, 等, 所以, 虽然, 但是, 而且, 即使, 并且, 然而, 同时, 就是, 并非, etc.

- (4) There are archaic idioms that contain the negative structure of ancient Chinese, e.g. 概[莫] ‘cannot’ 能外, 人[莫] ‘cannot’ 予毒, 居心[叵] ‘difficult’ 测, 大而[无] ‘not’ 当, 悔[不] ‘not’ 当初. Highlighted morphemes may be translated as 不 in modern Chinese, which means “no”, “not” or “without”, etc.

The spotting of archaic idioms defined by the use of negative structures in ancient Chinese can be done by highlighting the negative functional words in ancient Chinese, which represent an even smaller amount of lexical items in their totality: 莫, 非, 叵, 无, 不, 勿、毋.

- (5) There are also archaic idioms that contain morphemes existing only in formal ancient Chinese, like 毁家[纾] ‘alleviate’ 难, 聊以[卒] ‘end’ 岁, 暴[殄] ‘reduce’ 天物;; 面面[俱] ‘all’ 到.

Archaic idioms containing morphemes in formal ancient Chinese are the ones that can be most easily recognized, because morphemes that are used solely in formal ancient Chinese tend to be much more orthographically complex, as compared with words in modern Chinese.

- (6) Another special type of archaic idioms contains figurative words or morphemes in classical Chinese, such as 改换 [门庭] (courtyard → family status); 别具 [匠心] (the mind of a gardener → design). Noting here that the presence of figurative words in this kind of archaic idioms make them look like figurative idioms.

- (7) There are, finally, archaic idioms with a strong reference to traditional Chinese notions or ideological matters, e.g. 披麻戴孝 (traditional Chinese funeral ceremony); 安贫乐道 (Taoist proposals for spiritual life).

Such a detailed typology of Chinese idioms as proposed by Tao seems to be informative enough to provide us with an integrated framework for the classification of four-character items

extracted from the parallel corpus. However, it should be noted that the last two types of archaic idioms proposed by Tao, namely archaic idioms that contain figurative morphemes (AI-VI), and archaic idioms that include words with a strong reference to traditional Chinese ideological or socio-cultural issues (AI-VII), may well give rise to certain definitional confusion in the process of manual classification of Chinese idioms.

The reason for this is that the sixth type of archaic idioms has similar qualities to those of most figurative idioms by virtue of the fact that they share with them the figurative dimension of idiomaticity; and thus, can hardly be separated from the latter. As for the last type (7) of archaic idioms in Tao's typology, it could be said that they are actually quite analogous to the last type of figurative idioms, where the etymological origin or socio-cultural connotation of the idiom serves as the referent source of the figurative meaning of the item. Therefore, out of practical consideration for further corpus data processing, I propose to incorporate the last two types of archaic idioms into the family of Chinese figurative idioms so to avoid potential inconsistency of classification which may creep in as a result of the definitional ambiguity in Tao's original taxonomy of Chinese idioms.

2.6 Conclusion

The aim of Chapter II is to provide a detailed discussion of two major procedures in the exploration of the parallel corpus of *Don Quijote*. The two important steps carried out at this stage are the automatic extraction of four-character expressions from the parallel corpus and the establishment of an annotation scheme for the manual classification of the automatically retrieved textual data. These two practices play a major role in the full use of the parallel corpus thus built; and a detailed explanation of the way they have been pursued to achieve their goal has important implications for future research in a field where there has not been as yet sufficient debate on methodologies.

In the first half of Chapter II, I made some remarks regarding the identification of potential linguistic features in corpus-based translation research. It has been noted that the real purpose in adopting a corpus-based approach to translation is to uncover underlying linguistic patterns in the target texts compiled in the form of electronic files. Two important attributes of any potential textual features to be studied within the corpus-oriented scheme are that (1) it should be sufficiently large in numbers and widely distributed across the corpus texts in question; that (2) it should have some formal or structural characteristics to make it easy to extract automatically, with the current technical sophistication of corpus mining software. Compliance with these two basic requirements makes the data amenable to further exploration of translational corpora, i.e. the classification or annotation of the linguistic data thus extracted.

In my study of the stylistic profiles of the two Chinese translations, a very special kind of phrasal unit very typical of Chinese, has been highlighted, namely four-character expressions. These four-character expressions represent the most conventionalized part of Mandarin Chinese and are

usually linked with Chinese idioms, given that more than ninety per cent of items collected in Chinese idiom dictionaries are four-character long. It is generally believed that this highly condensed four-character pattern, which most probably has its earliest origin in formal literary or historical writings in ancient Chinese, reflects a unique phrasal predilection of the Chinese people. The underlying four-character pattern of phrasing, therefore, has important aesthetic values for the appreciation of literary writing style in modern Chinese.

Furthermore, this largely conventionalized length-specific form of writing has made a large-scale automatic extraction of four-character expressions from corpora an easily executable exercise with a variety of existing text mining tools. The two distinctive features of four-character expressions comply with the two technical requirements set out at the beginning of Chapter II: they should be relevant in terms of the quality of the language and there should be plenty of them. Such properties of four-character expressions make them an ideal candidate for the study of the two Chinese literary translations, which have proven to differ from each other significantly at the phraseological level.

Another major issue discussed in Chapter II with regards to further processing of corpus data is known as linguistic-knowledge annotation. Despite their numeric importance, the massively generated linguistic data are raw data which can hardly show any interesting information about the target texts under investigation. That is to say, machine-produced data need to be further classified to mark out their significance in the textual organization of translation. This actually leads to the second major issue discussed in Chapter II, which is the development of a topic-specific annotation scheme for the purpose of individual corpus projects. Instead of following a more routine approach to the corpus annotation, such as part-of-speech (POS) tagging, syntactic parsing, I have attempted to explore a highly pragmatic annotation type known as problem-oriented annotation.

The problem-oriented annotation has several advantages over traditional annotation types, which are among others, its cost-effectiveness in quickly building up small-scale translational corpora, and its versatility in dealing with a whole variety of research issues in translation studies, ranging from lexico-syntactical, semantic, and phraseological to typological, etc. This kind of problem-oriented annotation sits well with the current situation of corpus-based translation research, where the fast expansion of the discipline tends to cause an increasing gap between the various linguistic aspects of translational texts that it attempts to address and explain, and the highest level of automatic corpus annotation that current corpus engineering has achieved so far, which unfortunately lags behind largely at the POS level.

In exploring a system of annotation more responsive to solving problems, I have come to the idea of a special typological framework for Chinese four-character expressions. Based on the data collected from the parallel corpus, as well as previous studies on Chinese idioms, an open-ended descriptive taxonomy of Chinese four-character expressions has been formulated. Such a tentative categorizing scheme includes seven major types of conventionalized phrasal units composed by four-

characters, which are morphologically patterned phrases, syntactically schematic phrases, structurally symmetrical phrases, semantically bipartite phrases, shortened phrases within the four-character pattern, and lastly, four-character idioms. The first six types of four-character expressions may be relatively easily recognized in processing the corpus data manually, since the discriminating criteria suggested here are quite straightforward and less-prone to definitional ambiguity.

The identification of the last type of four-character phrases, namely Chinese idioms, has turned out to be more complicated. This has been largely due to the two features quite unique in Chinese idioms, i.e. their structural versatility and double dimension of idiomaticity including figurativeness and archaism. Such semantic-syntactic particularities of Chinese idioms have been rarely discussed in the literature, a clear description of which however holds the key to a successful manual annotation of the Chinese texts compiled in the parallel corpus.

For the purpose of my research, I have revisited the notion of Chinese idioms and made important observations with respect to the behaviour of Chinese idiomatic expressions in naturally-occurring contexts. This in turn has provided general guidance to the classification of four-character expressions retrieved from the parallel corpus of *Don Quijote*. The linguistic data thus annotated will be then used in the next stage of the project, which deals with the quantitative exploration of the textual patterns hidden behind the sorted data.

Chapter III General Phraseological Patterns in Yang's Translation

3.1 Quantifying style in corpus-based Translation Studies

Quantifying style, or stylometry, is one of the best established methodologies in western literary studies. But it has rarely been practised on translation texts in exploring literary stylistics or authorship attribution. As a result, many scholarly works on individual translators' style seem to have based their judgements on limited excerpts randomly and irregularly selected from parallel source/target texts. The introduction of corpus methodologies to descriptive translation studies has brought with it renewed attention to issues of generation and codification of quantitative data from electronically-stored translation material. It is now generally held that a corpus-driven or corpus-based approach to translation studies could yield data by following a more empirically-based line of enquiry into the stylistic nature of literary translation.

From initial discussions on the ontological status of a translator's style (Mikhailov & Villikka, 2001: 378-85), through to the outline of a potentially prolific methodological framework for the study of stylistic patterns in translated texts (Baker, 2000: 241-66), and finally to the actual practice of quantifying the use of highlighted linguistic features in search of possible stylistic traits in translations (Saldanha, 2005), it would seem that such a research agenda seeking to elicit the visibility of translators in their work is steadily gathering momentum, reshaping the way we see translation as an act of re-creation in its own right, and the new status of translators, as legitimate creative writers assimilating foreign elements into their own cultures, is becoming more recognised as a result.

In the course of developing a solid line of empirical research in corpus-based translation studies outlined above, I have argued for an interdisciplinary approach to style in translation studies, which, as will be shown in this thesis, points to a potentially very productive direction for future research. It is believed that an interdisciplinary study of translation style will benefit greatly from the cross-fertilization of research methodologies and insights obtained from different although related disciplines, such as computational stylometry (Hoover, 2003), corpus phraseology (Moon, 1998), textual statistics (Stubbs, 2001), quantitative sociolinguistics (Biber & Finegan, 1994; Eckert & Rickford, 2001) and cognitive stylistics (Semino & Culpeper, 2002; McIntyre, 2006).

In fact, the growing interaction and complementation among these subject domains has already prompted the development of an emerging field known as corpus stylistics (Wynne, 2005; Hoover, Louw and Culpeper, 2008). This thesis hopes to prove that corpus-based translation studies in general, and translational stylistics in particular, should and will be able to benefit from such a general trend in current literary studies, which is characterized by its interdisciplinarity and growing reliance on computational techniques in pursuing textual studies.

Despite the huge research potential offered by empirical stylistics, as well as the promising

results obtained so far in this regard, the type of quantitative analysis pursued in early works exhibits limited versatility and a certain lack of sophistication in the argumentation. For example, preferred linguistic choices by translation stylisticians are, among others, function words, punctuation marks, syntactic or part-of-speech (POS) information, etc. On the other hand, semantically-rich textual elements, such as lexis or phraseology, are hardly given much consideration. This aspect of language is therefore in need of much greater attention if we are to advance beyond the rather impressionistic and over cautious approach of traditional stylistics.

Formidable technical problems obviously stand in the way of a corpus-based quantitative analysis of relevant lexical or phrasal categories which would help with the processing of translational corpora and the attribution to a translator of a particular style. Nevertheless, as will be shown in the present study, despite the improvements made to many current corpus linguistics tools, this technical hurdle in the quantification of translators' style is not far from being resolved. In fact, it is suggested that the infrequent use of corpora in the study of translators' style could actually be sorted out by adapting a more flexible and pragmatic attitude bearing in mind that the mechanical inspection and human observation surely complement one another.

Secondly, in order to advance further in the field of corpus-based translation stylistics, one needs to be very clear about the difference between descriptive data and inferential statistics (Oakes, 1998: 1). The generation of descriptive data involves the extraction and summation of countable linguistic events within corpus texts, from which one might infer the characteristic use of language by a particular author. In fact, many corpus-based stylistic studies on translation have remained at this level. The conclusions thus obtained are based on *observed* patterns in a limited amount of textual information, rather than on a wide base of general distribution of quantitative linguistic features throughout the corpus texts especially selected for investigation.

The methodological limitations of a traditional approach soon become obvious when translation researchers are faced with the processing of linguistic data in hundreds or in thousands, instead of tens. In this situation, it is necessary to resort to certain statistical techniques to uncover such underlying linguistic patterns in translations as may be hidden behind an agglomeration of corpus data. Such corpus processing procedures actually entail what Oakes refers to as inferential statistics, which may among other functions, answer questions, formulated as hypotheses, regarding whether and to what extent one author's style is different from another's. Given its wide application in corpus linguistics and in other related fields in the social sciences, it seems that such methodology may well be worth testing in Translation Studies.

In this chapter and the next one, which focus on the use of four-character expressions (FCEs) in the two contemporary Chinese versions of Cervantes's *Don Quijote*, I should endeavour to address the two main methodological issues in corpus-based translation stylistics mentioned above, namely, the relative lack of debate on the importance of semantically-rich linguistic units in quantifying translating

style, and the need for testing methods and techniques adapted from corpus statistics to help with the uncovering of stylistic traits in translations. It is hoped that this study, which aims to expand the current methodological framework for translation stylistics, will help in the development of this fast growing area of research.

3.2 Sampling

Faced with a very large number of four-character segments detected in Yang's and Liu's translations, accounting for all the instances extracted from CCDQ would appear to be quite impractical. The manual assignment of each four-character item to the relevant four-character category would also turn out to be extremely time-consuming and error-prone, especially considering the internal complexity of the typological framework established for Chinese four-character expressions in the previous chapter.

The quantitative nature of the corpus data thus gathered requires a probabilistic approach to the identification of linguistic patterns which may lie behind four-character expressions in the two Chinese translations under consideration. In other words, since it is not practical to study all the examples of four-character expressions retrieved from CCDQ, we would need to select a fraction of the database which will then be used to represent the entire corpus as a whole. The significance of sampling for corpus-based textual analysis lies in the fact that it may provide important information regarding the specific characteristics of a large population, the internal complexity of which may not be at all perceptible to human inspection and manual analysis.

To ensure that samples are chosen at random, while still being representative in describing the distribution of FCEX across the two target texts, an online random number generator has been employed³⁰, which will help extract samples automatically from the two translations (see Appendix I & II). A basic characteristic of this kind of computer-assisted technique is that the numbers generated by automatic randomizers, once computed, will be soon discarded by the programme and the same data would never be used again. This is to secure the randomness of the sequences of digits thus produced.

Due to the large size of FCEXs retrieved from each translation, some three hundred four-character expressions have been selected randomly from each translator's work, which roughly represent some fifteen per cent of the entire database of FCEXs. All the FCEX instances included in the test sample sets are then classified according to their structural or semantic properties and incorporated into a cross-tabulation, which exhibits the distribution of FCEXs across the three sample sets in Yang's and Liu's works, respectively.

³⁰ see <http://www.randomizer.org/>

Table IV Distribution of FCEX types in Yang's and Liu's work

No	Code	Yang		Liu		Row Total
		Observed	Expected	Observed	Expected	
1	SP	18	18.5	19	18.5	37
2	I-FIG	6	10.5	15	10.5	21
3	SS	15	15.5	16	15.5	31
4	SSY	20	16.5	13	16.5	33
5	MP	62	42.5	23	42.5	85
6	SB	63	45	27	45	90
7	I-AI	7	18.5	30	18.5	37
8	FIG	57	66	75	66	132
9	AI	52	67	82	67	134
Column Total	FCEX	300	300	300	300	600

N.B. SP: Shortened phrases; I-FIG: instantiated figurative idioms; SS: Syntactically schematic; SSY: Structurally symmetrical; MP: Morphologically patterned; SB: Semantically bipartite; I-AI: Instantiated archaic idioms; FIG: Conventionalized figurative idioms; AI: Conventionalized archaic Idioms; FCEX: four-character expressions

To see whether there are any significant differences between the two translations in terms of the distribution of main four-character expression types, we shall conduct the statistical procedure, χ^2 (chi-square) test, which is widely used in corpus linguistics and social sciences to evaluate statistically significant differences between proportions for two groups in a dataset. It is a non-parametric test and has the great advantage of not depending on the population being normally distributed. However, it should be noted that the use of χ^2 test in quantitative analysis will be inappropriate, if any expected frequency is below one or if the expected frequency is less than five in more than twenty per cent of the contingency table (Dawson & Trapp, 2004: 153-4). The expected frequency can be found using the following formula:

Formula I Expected frequency calculation in χ^2 test

Expected value = Row total x Column total / Grand total of items

The expected values for the observed frequencies have also been incorporated into Table IV. As we can see, in none of the cells has the expected value turned out to be less than five, which suggests the χ^2 test may be safely deployed with the data presented in Table IV. Now, we may proceed to compute the chi-square value for the contingency table, by using the formula shown below:

Formula II Chi-square test

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

In this formula, O stands for the observed frequency and E, for the expected value. The interpretation of the calculated χ^2 value needs the contrast of the actual result with a set of predefined

critical values at different levels (Oakes, 1998: 27). This is because the chi-square test, just as many other statistical tests, is performed under the null hypothesis that there is no significant difference between the proportions for the two groups. To prove that the initial assumption does not apply to the current study, the computed χ^2 value has to be larger than the threshold value, normally set at five per cent. The computation of the data shown in Table IV gives a χ^2 value of 61.164 (Preacher, 2001)³¹, which is much larger than the critical value 26.12 at the 0.001 level³², suggesting that the differences between the two translations are indeed significant.

There are a number of factors that may explain such significant differences between Yang's and Liu's work. Firstly, it is worth noting that there is a fifteen-year gap between the two Chinese versions: while Yang's work was officially launched to the public in 1978, Liu's version did not come out until 1995. These fifteen years have witnessed important changes in the evolving of Chinese: it has been a critical moment in modern China, in which the country has entered a new age of vigorous cultural and economic contacts with the external world never seen before.

To appreciate the changes the Chinese language may have experienced within this particular chronological frame, it would be ideal to compare two similarly constructed monolingual corpora of Chinese, with particular attention being paid to the differences in the distribution of four-character expressions, or mainly idioms, across a range of text genres. The corroboration of any working hypothesis or theory with regard to these changes largely depends on the availability of relevant corpus material. For the time being, however, it seems that we are still not in a position to describe quantitatively any linguistic changes that Chinese may have experienced during these fifteen years, for there is no comparable monolingual corpus available for the time span under investigation.

Nevertheless, we may yet get a flavour of the rapid changes taking place in China between the 1990s and early twenty-first century, especially with regards to the noticeable shifts in the distribution of idioms across different text genres, as formal versus fictional. Also, the finding uncovered in such a comparative study may reduce the possibility that the differences simply reflect personal preferences. This is now feasible thanks to the availability of two monolingual Chinese corpora, i.e. LCMC (Lancaster Corpora of Mandarin Chinese, 1990s) and UCLA Chinese Corpus (early 21st century), which have been compiled under the same sampling scheme and are therefore largely comparable for the purpose of studying linguistic change.

3.3 The time factor in the use of idioms in contemporary Mandarin Chinese

Languages are always on the move, rhythm and tempo may vary considerably as time goes by, but change goes on across time and space in a vast continuum. This section sets out to throw light on possible changes in contemporary Mandarin Chinese seen through a comparative study of idiomatic

³¹ Preacher, K. J. (2001) Available from <http://www.quantpsy.org>

³² The appendix of chi-square critical values used may be found in Oakes (1998), p.266

expressions in two large-scale modern Chinese monolingual corpora, i.e. Lancaster Corpus of Mandarin Chinese, also known as LCMC (1990s), and the UCLA Chinese Corpus (early 2000s). The two corpora have been constructed following the same sampling framework as that of the Brown or LOB corpus, and are thus essentially comparable³³. As two widely distributed monolingual corpora of modern Chinese, both LCMC and UCLA Corpus have been built to address the increasing need for large-scale comparable corpora to do contrastive language studies, usually in combination with purposely-built specific corpora of much smaller size.

The very interesting results presented here suggest that idioms have experienced a significant change in the last decade of the twentieth century. This is supported by corpus evidence showing that when compared to the LCMC, the later UCLA corpus shows a noticeable reduction in the use of idioms across many Chinese text types or genres. The diminishing presence of idioms in a variety of established Chinese text genres is not at all a trivial linguistic discovery, when taking into account the vital position held by idioms in Chinese lexicographical or phraseological studies. Here, I hope to show that a quantitative examination of the two consecutive Chinese monolingual corpora will bring us closer to an understanding of the changing nature of Chinese idioms viewed diachronically.

Table V Distribution of idioms in LCMC versus UCLA Chinese corpus³⁴

Code	Text Type	Raw Frequency (LCMC)	Raw Frequency (UCLA)
AD	Adventure/Martial Arts Fiction	338	300
ES	Essays and Biographies	931	363
GF	General Fiction	290	223
HU	Humour	108	76
MY	Mystery/ Detective Fiction	266	493
NED	News Editorials	369	111
NREP	News Reportage	484	236
NREV	News Reviews	249	117
PL	Popular Lore	501	171
RE	Religion	112	7
REP	Reports/Official Documents	108	36
RO	Romantic Fiction	378	263
SC	Science (Academic Prose)	344	51
SF	Science Fiction	45	255
SK	Skills/Trades/Hobbies	244	9
Total	Total	4767	2711

Table V exhibits the raw frequency of idioms in different text genres, which is an initial comparison of the two monolingual corpora. However, it should be noted that the first impression that we may gather from such a comparison may turn out to be misleading, given the different sizes of the two corpora. While the LCMC contains one million tokens³⁵, the current version of the UCLA corpus

³³ See http://www.lancs.ac.uk/postgrad/xiaoz/papers/corpus%20survey.htm#_Toc92298888

³⁴ Last access to LCMC and UCLA corpus was on February 8, 2008

³⁵ See <http://bowland-files.lancs.ac.uk/corplang/lcmc/>

holds 687, 634 running words in its collection³⁶. As a result, it would be rather difficult to tell from the outset whether the two corpora genuinely differ from each other in terms of the distribution of idioms across the fifteen text types.

To bring to the surface implicit linguistic patterns lying behind the raw frequencies, the method I am using here is the statistical procedure known as Pearson's moment-product correlation test, which has produced the important statistical results shown in Table VI. Pearson's correlation test is widely used in corpus linguistics to test the strength of association between different corpus texts. It does not assume any causal relationship between the variables under test and may only deal with continuous data. It expresses the strength of correlation numerically through the correlation coefficient, *R*, which varies from minus one to one as the maximum values at two extremes.

Table VI shows in the first place that the mean frequency of idioms in the LCMC is as high as 317.8, which is much higher than that of the UCLA corpus. The computed coefficient of the correlation model is approximately 0.435, whose further interpretation requires the consultation of the index of Pearson's coefficient critical values set at different significant levels³⁷. As a normal practice in corpus linguistics, we opt for the five per cent as the threshold level to measure the strength of correlation between the two Chinese corpora. Given that we do not see an obvious reason to assume or hypothesize the existence of a strong relationship between the two corpora in advance, we shall check the computed coefficient value with the critical value at the two-tailed non-directional category, which is always more prudent than using the one-tailed directional value³⁸.

Table VI Summary of Pearson's correlation test

STATISTIC	VARIABLE X (LCMC)	VARIABLE Y (UCLA)
Mean	317.8	180.73
Biased Variance	4, 4198.69	18634.86
Biased Standard Deviation	210.23	136.51
Covariance	13364.37	
Correlation	0.435	
Determination	0.19	
Degrees of Freedom	13	
Number of Observations	15	
Critical value for Pearson's test (two tailed at 5% level)	0.514	
Significance (Y/N) (two tailed at 5% level)	No (no significant correlation between two corpora)	

The mechanics of the Pearson's correlation test is that we start the statistical procedure by assuming a null hypothesis which treats the two corpora as having no relationship at all. In order to subvert the default hypothesis, the computed coefficient must be equal or greater than the critical value. However, as Table VI shows, the correlation coefficient obtained from the two Chinese corpora, which is as low as 0.435, is definitely below the threshold value at the critical five per cent, which is 0.514.

³⁶ See <http://bowland-files.lancs.ac.uk/corplang/ucla/>

³⁷ See Appendix 8 The Pearson's product-moment correlation coefficient, Oakes, M. (1998), p. 267

³⁸ Personal discussion with Dr. Stefan Evert, computational linguist at Osnabrück University, Germany

This suggests that despite the many similarities shared by the two corpora, such as the same sampling framework and the same language type, standard Mandarin Chinese, they indeed differ from each other in terms of the frequency of occurrence and the relative distribution of idioms.

Given the general assumption of a high degree of conventionality in Mandarin Chinese, the fact that the rapid changes happened within a short period of time would surprise many people (Xiang, 1979: 112-21). But it does happen: these changes are taking place as shown by the variations in the use of Chinese idioms across different text genres. To have a clear view of the shifts in the use of idioms in the UCLA corpus, as well as the general patterns of distribution of idioms in each corpus, the raw frequencies summarized in Table V have been normalized into frequencies per 10k words.

To facilitate access to the numerical information provided in Table VII, these results have been used to draw a histogram in which the two coloured sets of bars represent the distribution of idioms across different text types in the LCMC (grey bar) and the UCLA corpus (black bar), respectively. As may be seen from the graph, several important patterns regarding the evolving nature of Chinese idioms in written texts seem to emerge³⁹.

Table VII Comparison of normalized frequencies between LCMC and UCLA corpus

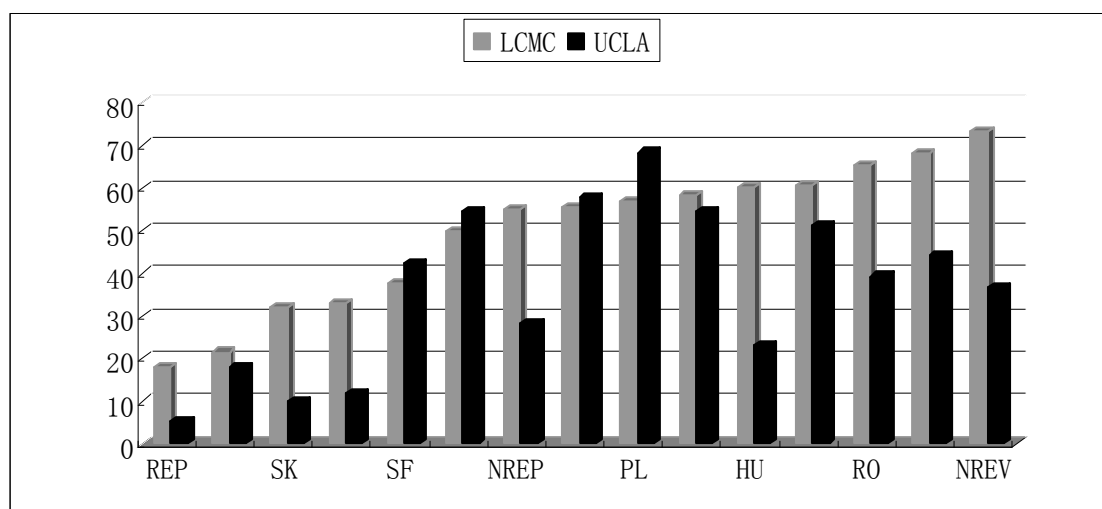
CODE	FREQUENCY PER 10K WORDS (LCMC)	FREQUENCY PER 10K WORDS (UCLA)
REP	18	5.45
SC	21.5	18
SK	32	10
RE	32.9	11.7
SF	37.5	42.5
GF	50	54.4
NREP	55	28.10
MY	55.4	58
PL	56.9	68.4
AD	58.3	54.5
HU	60	23
ES	60.5	51.1
RO	65.2	39.3
NED	68.3	44.4
NREV	73.2	36.6
Total	47.7	39.5

The grey bar which embodies the LCMC shows a general trend to run above the black bar representing the UCLA corpus. This fits in well with the descriptive statistics shown in Table VII, where the mean frequency of the LCMC is twice that of the UCLA corpus. It seems to suggest that at the general level, the language recorded in the LCMC is richer in idioms than the material compiled in the UCLA, constructed some ten years later. However, the term “idiomaticity” is a very complex concept, which may well have different connotations in different text types or genres. In a previous study based on the evidence collected from large-scale modern Chinese corpora, i.e. the Modern

³⁹ Both the LCMC (Lancaster Corpus of Mandarin Chinese) and the UCLA corpus have been constructed with material collected from sources of written texts, e.g. online electronic libraries, or electronic texts on the web.

Chinese Corpus⁴⁰, it has been noticed that there are three main features which characterize Chinese idioms: conventionality, the presence of an element of figuration or archaism, and potential structural flexibility (Ji, 2007c).

Diagram V Comparison of normalized frequencies between LCMC and UCLA corpus



There are five perceivable low points in the set of black bars, which take place (1) in the first genre, Reports or official documents (REP); (2) between the Skills/trades or hobbies (SK) and Religion (RE) categories; (3) in one of the middle text type, News Reportage (NREP); (4) Humour (HU) and (5) in the last category, News Review (NREV). Among these six divisions, which have been highlighted because of the observed sharp decrease in the use of idioms in the relevant text genres, we can see there are invariably non-fictional text types. As mentioned above, the practical functionalities of Chinese idioms may be approached either by its figurative or archaic attributes. Then, in the light of the distinctive discursive features or writing conventions of these highlighted Chinese text genres, we could assume that with regards to Chinese text types which use idioms as an important rhetorical device to enhance the formality of the language style, such as News Reportage (NREP), News Reviews (NREV), Religious (RE), Reports and official documents (REP), the language used in these text genres seems to have been evolving very noticeably towards a more informal style.

On the other hand, in text genres where idioms may be explored as figurative tropes, such as in the case of Humour (HU), the significant drop in the use of idioms in a time span of approximately ten years seems to suggest that the semantic transparency of such texts has been enhanced considerably. At the same time, the two peaks featured in the red bars seem to suggest that the rhetoric or aesthetic value of idioms in Chinese fictional or popular writings has been steadily enhanced, which is well represented by the two small peaks in the red bars as above its blue counterpart Science Fiction (SF), General Fiction (GF) and Popular Lore (PL).

⁴⁰ The Modern Chinese Corpus may be accessed at http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp?dir=xiandai

These findings uncovered through a comparative study of two Chinese corpora separated by approximately ten years would require further explanation within a broader socio-linguistic framework to allow us a fuller understanding of the evolving nature of Chinese language in the last quarter of the twentieth century. However, on this basis, one might suspect that similar trends regarding the changing nature of Chinese idioms (as described above) may have also been active in the seventeen years' period between 1978 to 1995, the publication dates of the two Chinese versions under consideration. If such hypothesis works, it would appear that Liu's increased use of idioms in his version of *Don Quijote* Part I, which will be explained in some details further on, has basically followed the general trend in the rise of the use of idioms in Chinese fictional genres. This in turn would seem to suggest that Liu's increased use of idioms implies an intention on his part to assimilate the source text language into the target linguistic habits, which presumably has been achieved by injecting more idiomatic elements into his way of writing.

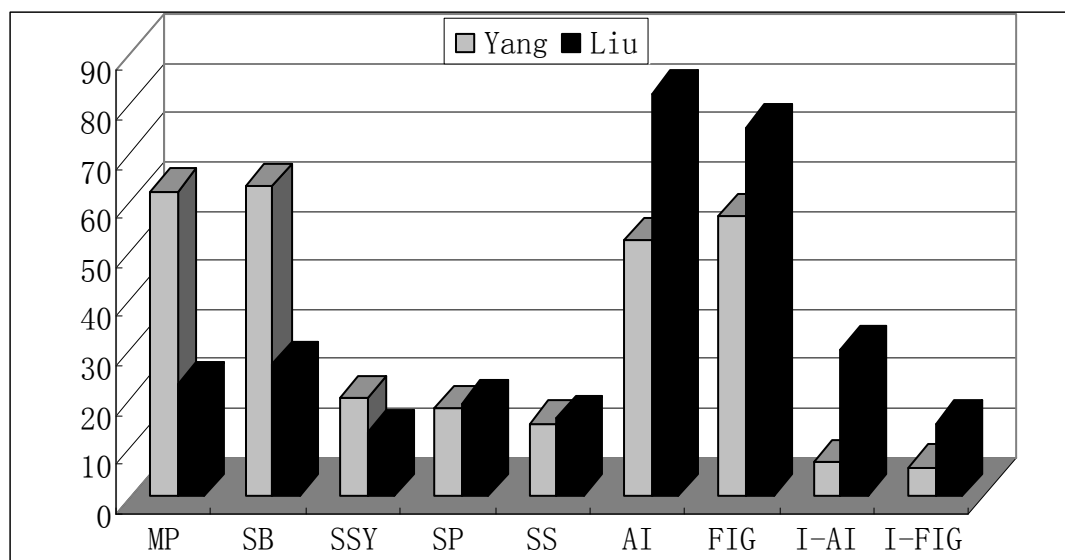
In the following sections of this chapter and in the next chapter, I shall examine more closely the different use of Chinese four-character idiomatic expressions made by Yang (1978) and Liu (1995), respectively, by focusing on three general patterns underlying the two Chinese versions of *Don Quijote*: Yang's preferred use of Chinese morpho-syntactically patterned four-character expressions, such as semantically bipartite phrases, morphologically patterned phrases, and structurally symmetrical phrases; (2) Liu's higher use of Chinese figurative idioms; and finally, (3) Liu's idiosyncratic use of Chinese archaic idioms (these last two points will be dealt with in Chapter VI).

3.4 Yang's Phraseological Preference for SBs, MPs and SSYs

Going back to Table IV, we can observe the general pattern of distribution of the main types of FCEs in the testing sample sets. It gives us a first impression of numerical contrasts between Yang's and Liu's work regarding their different phraseological profiles. I should point out that the reason why we have used raw frequency, instead of normalized proportions to plot the histogram is due to the fact that the same size of the sample set has been selected from each translation. The figures shown in the cross-tabulation are then used to generate a bar chart, which helps visualize the trends hidden behind the abstract data.

As can be seen from the graph displayed above, a significant number of consistent patterns underlying Yang's and Liu's use of FCEs in their different versions begin to emerge. The major difference between the paired bars occurs when the grey bars (Yang) run higher than their black counterparts (Liu) across the initial three categories, i.e. morphologically patterned FCEs (MPs), semantically bipartite FCEs (SBs) and structurally symmetrical FCEs (SSYs). Such findings surely reveal an important aspect of Yang's phraseological profile.

Diagram VI Identification of general contrastive patterns between Yang's and Liu's work



The relatively higher use of these three FCEX categories by Yang would seem to exhibit some similarity in the way the three categories operated by modifying phraseological patterns at a structural rather than semantic level. This in turn may be said to respond to an initial and original intention to assimilate the source text to the linguistic characteristics of the target language. In this section, we shall first concentrate on this particular profile of Yang's work, by studying in details Yang's use of MPs, SBs and SSYs in relation to the original, and also in comparison with Liu's choice of other Chinese phraseological categories.

3.4.1 Semantically Bipartite FCEXs (SBs)

(1) ST containing repetitions or bi-synonymous structures⁴¹

ES1 ... sin poder hallarla, vinieron a parar a un prado lleno de fresca yerba, junto del cual corría un arroyo apacible y fresco: (Chapter XV, I)

EN: they came to a halt in a glade covered with tender grass, beside which ran a pleasant cool stream that invited and compelled them to pass there the hours of the noontide heat

Y: 后来他们走到一片碧油油的草地上，旁边有一条平静清澈的溪水。

L: 最后来到一片绿草如茵的平地上，旁边有一条清澈的小溪缓缓流淌。

ES2... pero don Quijote, coligiendo por su huida su miedo, alzándose la visera de papelón y descubriendo su seco y polvoroso rostro, con gentil talante y voz reposada les dijo: (Chapter II, I)

EN: but Don Quixote, inferring their fear from their flight, raised the pasteboard visor, revealing his dry, dusty face, and in a gallant manner and reassuring voice, he said to them

Y: 露出一张又干又瘦、沾满尘土的脸，斯文和悦地说：

⁴¹ In this part, I use a box to highlight Yang's use of morphologically patterned FCEX and an underline to mark Liu's literal translation or plain explanation.

L: 便掀起纸壳做的护眼罩，态度优雅、声音平缓地对她们说：

The use of SBs by Yang in translating ES1 and ES2 resemble each other in that the perceived phraseological patterns of the source texts, i.e. the juxtaposition of synonymous adjectives or nouns, serve as a linguistic model upon which the Chinese translator may build similar constructs given the availability of particular linguistic properties of the target language, i.e. the semantically bipartite four-character expressions. The use of a SB structure as [Adj₁ + Adj₂] by Yang in translating ES2 seems to have been motivated by the source linguistic feature as containing two juxtaposed parts [gentil talante + voz reposada] within a prepositional phrase initiated by *con*.... However, due to the very limited space provided by the four-character pattern, Yang has had to drop off the Chinese translation of the two nouns in the source texts, i.e. *talante* (manners, mode) and *voz* (voice), a phrasing strategy which thus requires her to be highly selective in bringing in adjectival elements to fill in the four-character pattern at a later stage.

To make up for the loss of information, inevitable in this case, one would need to choose Chinese adjectives which not only match the meaning of *gentil* (gentle) and *reposada* (peaceful), but also tend to be used as collocations of the nouns *manners* (talante) and *voice* (voz), in order to maintain the semantic integrity of the source text and not to cause any substantial loss of the original information. In fact, this is exactly what Yang does as a skilful translator and creative writer -- the two adjectives 斯文 (well-mannered and educated) and 和悦 (soft and pleasant) tend to be used to describe one's behaviour without referring to the subjects that they are modifying. In contrast, Liu's translation has broken up the prepositional structure of the source text into common phrases.

(2) Others

ES1 “La razón de la sinrazón que a mi razón se hace, de tal manera mi razón enflaquece, que con razón me quejo de la vuestra fermosura”. (Chapter I, I)

EN: The reason for the unreason to which my reason turns so weakens my reason that with reason I complain of thy beauty.

Y: “你以无理对待我的有理，这个所以然之理，使我有理也理亏气短；因此我埋怨你美，确是有理。”

L: 许多地方这样写道：“以你无理对我有理之道理，使我自觉理亏，因此我埋怨你漂亮也有道理。”

In the case of ES1, the SB structure used in Yang's translation is [理 (reason) 亏 (insufficient)] + [气 (courage) 短 (short)]. It has been modelled on the four-character pattern to enforce the feeling of frustration as conveyed by the source verb *enflaquecer*. Such a rhetorical use of the four-character morpho-syntactic pattern greatly enhances the formality of the protagonist's utterance. However, the parodic connotation of such phraseological pattern is triggered off when embedded in a wider sentential context that largely resembles similar comical and verbose writings in the target language, as hinted at by the use of pseudo-archaic expressions like 所以然 (therefore + archaic suffix) and 确是 (actually is).

The phraseological pattern of Liu's translation, on the other hand, seems to be less

pragmatically charged than Yang's approach, though its deliberate use of monosyllabic words as typical of ancient Chinese to substitute disyllabic words in modern Chinese somehow helps to build up a formal style of language in his work. This actually signals Liu's implicit intention to modify the register of the target language by injecting more archaic elements. I shall talk in more detail about this particular linguistic feature of Liu's version in the following chapter.

3.4.2 Morphologically Patterned FCEs (MPs)

The preferred use of morphologically patterned FCEs is another major feature of Yang's translation. As shown in Diagram IV (Identification of general contrastive patterns between Yang's and Liu's work), it is in fact the phraseological feature that most differentiates the two translators. As shown in the example given below, morphologically patterned FCEs provide the Chinese translator with a versatile phraseological device, where certain source grammatical structures or linguistic expressions in the source text are not readily available or common in the target language. In assimilating the source language into the target language, in line with the different functions, Yang's use of FCEs may be divided up into the following four modes of operation: i.e. repetition, plurality, modification of degree of emphasis and bi-synonymous structures.

(1) Repetition

ES1.... porque querer vuestra merced que vaya con él por estas soledades, de día y de noche, y que no le hable cuando me diere gusto es enterrarme en vida. (Chapter XXV, I)

EN: Your grace wanting me to go with you through these deserted places by day and by night without talking whenever I feel like it is like burying me alive.

Y: 您要我跟着您 日日夜夜 在这种荒僻的地方奔走，想跟您说话又不能够，这简直是活埋了我。

L: 您让我跟您 日夜兼程，在荒郊野岭奔走，想跟您说话的时候还不能说，这简直是活埋我

The use of the MP [日 (day) 日 (day) + 夜 (night) 夜 (night)] by Yang in translating ES1 is in accordance with the source text feature as it contains a repeated prepositional structure [*de ...de...*]. In contrast with Liu's use of archaic idiom [日夜 (day and night) 兼 (both, archaism) 程 (travelling)], Yang's translation has retained the morphological construct of the source text expression, while converting it into the typical four-character pattern of Chinese writing. The phraseological repetition used in the target text assumes an important communicative function in conveying Sancho's complaints after long and useless adventures with don Quijote. Such an attitude is somehow belittled in Liu's use of Chinese archaic idiom, which is supposed to reinforce the formality of the squire's speech rather than to express his dissatisfaction.

(2) Plurality

ES1 Déjeme, iré a ensillar a Rocinante, y aparéjese vuestra merced a echarme su bendición, que luego pienso partirme, sin ver las sandeces que vuestra merced ha de hacer, que yo diré que le vi hacer tantas que no quiera más. (Chapter XXV, I)

EN: Let me go and saddle Rocinante, and let your grace get ready to give me your blessing, for I plan to leave right away without seeing the crazy things your grace is going to do, though I'll say I saw you do more than anyone could wish.

Y: 因为我打算马上动身，您还得干些什么 疯疯癫癫 的事，我都不瞧了。我会对她说，我看见您干了多少多少疯傻的事，叫她听不下去。

L: 不打算再看您要做的 那些蠢事 了。我会把我所看到的一切都告诉她，一点儿都不会漏下。

ES2 Otras algunas menudencias había que advertir, pero todas son de poca importancia y que no hacen al caso a la verdadera relación de la historia, que ninguna es mala como sea verdadera. (Chapter IX, I)

EN: A few other details were worthy of notice, but they are of little importance and relevance to the true account of this history, for no history is bad if it is true.

Y: 此外还看到些 枝枝节节，不过都无关紧要；故事只要真实就好，那些末节是无足轻重的。

L: 还有一些 琐闻，不过都无关紧要，并不影响故事的真实性。所有琐闻都是真实的。

The use of morphologically patterned FCEs has been shown to perform one other important grammatical function in translating *Don Quijote* into Chinese, which probably derives from the contrasting nature of the source and target languages. As typically inflected language, the plurality of nouns in Spanish may be easily implemented – with the exception of irregular nouns –, by simply adding an *s* to the end letter of a word. However, this is obviously not the case with the pluralization of Chinese nouns, whose grammatical representation is often realized through the use of definite/indefinite quantifiers or relevant determiners prior to the nouns being modified.

In fact, this is exactly what Liu has opted for in translating the plural form of nouns in ES1 and ES2: 那些蠢事 (those stupidities) for *las sandeces*; 一些琐闻 (some trivialities) for *algunas menudencias*. As opposed to the conventional way of translation adopted by Liu in these two cases, the use of MPs by Yang should be seen as a novel and original attempt at restructuring the target text. In order to express the plural form of nouns in the target text, Yang has taken recourse to the morphological repetition of adjectives, which is a typical phraseological device of Chinese, to achieve similar grammatical changes in the target language: [疯疯 (mad-mad) + 癫癫 (insane-insane)] for *las sandeces*; [枝枝 (branch-branch) + 节节 (knot-knot)] for *algunas menudencias*.

(3) Modification of degree of emphasis

ES1.... si las señoras del coche, que hasta entonces con gran desmayo habían mirado la pendencia, no fueran a donde estaba y le pidieran con mucho encarecimiento les hiciese tan gran merced y favor de perdonar la vida a aquel su escudero (Chapter IX, I)

EN: ...given Don Quixote's blind rage, if the ladies in the carriage, who until that moment had watched the battle

with great dismay, had not approached him and implored him most earnestly that he do them the favour and grant them the boon of sparing the life of their squire.

Y: 车上几个女眷一直在哆哆嗦嗦看打架，这会儿亏得她们赶来，恳求堂吉诃德宽宏大量，手下留情

L: 幸亏车上那几位一直在惊恐地观战的夫人来到唐吉诃德面前，哀求他大发慈悲，饶恕她们的侍从。

ES2 convidaron a los dos, con muestras de muy buena voluntad, con lo que tenían. (Chapter XI, I)

EN: and with displays of goodwill invited them both to share what they had.

Y: 他们把几张羊皮铺在地下，一转眼就摆上了朴素的便饭，诚诚恳恳邀请两位客人同吃。

L: 这时，牧羊人把锅从火上端了下来，在地上铺了几张羊皮，迅速摆上一张旧桌子，非常客气地请两人共同进餐。

The exploration of the morphological pattern of MPs in Yang's translation of *Don Quijote* has shown a third dimension which relates to the modification or intensification of emphasis. In translating both ES1 and ES2, the use of MPs, i.e. morphological repetition of adjectives by Yang, is motivated by the appearance of *gran* (short form of *grande* meaning great) and *muy* (very) in the source texts. In the case of ES1, the formation of the morphologically patterned four-character expression 哆哆嗦嗦 is attained in two steps: firstly, the deconstruction of the disyllabic word 哆嗦 (tremble or shiver) into two monosyllabic morphemes 哆 and 嗦; then, through the duplication of each morpheme, the original disyllabic word is extended into a typical four-character expression which is much more descriptive than its root word. By the same token, the configuration of the MP 诚诚恳恳 (sincere and earnest) has followed the same procedure of deconstruction and reconstruction. On the other hand, Liu's translation is basically literal, using similar Chinese adverbs to translate the Spanish modifiers. As we can see, here the exploration by Yang of the functional versatility of MPs is again to increase the expressiveness or vividness of the target text, which represents a highly skilful manipulation of the target language.

(4) Bi-synonymous structures

ES1 -Mire vuestra merced bien lo que dice y mejor lo que hace -dijo Sancho-, que no quería que fuesen otros batanes que nos acabasen de abatanar y aporrear el sentido. (Chapter XXI, I)

EN: "Your grace, be careful what you say, and more careful what you do," said Sancho, "for you wouldn't want this to be more fulling hammers that end up hammering and battering our senses".

Y: 桑丘说：“您说话得仔细，干事更得仔细啊。我但愿别又是捶打得咱们昏头昏脑的研布机之类。”

L: "那个东西您可得看清楚," 桑乔说, "但愿别又是一些刺激咱们感官的研布机。"

ES2.... Veamos si le iguala o hace ventajas el de embestirse dos galeras por las proas en mitad del mar espacioso, las cuales enclavijadas y trabadas, no le queda al soldado más espacio del que concede dos pies de tabla del espolón (Chapter XXXVIII, I)

EN: And if this seems an insignificant danger, let us see if it is equalled or surpassed when the prows of two galleys collide in the middle of the wide sea, for when they lock and grapple, the soldier is left within no more than two feet of plank on the ram of the ship

Y: 当时两只船难分难解，战士只能用船头二尺宽的撞角作立脚之地。

L: 两只船碰撞在一起，战士们只能在船头的冲角上有两尺宽的立足之地。

The use of morphologically patterned four-character expressions by Yang in ES1 and ES2 is derived from the source text feature as containing bi-synonymous structures of verbs (*abatanar y aporrear*) or past participles (*enclavijadas y trabadas*). Nevertheless, despite the formal resemblance between the bi-synonymous structures of the source text and the morphologically bipartite pattern of Yang's translation, a significant substitution of semantic content has taken place when the source text is translated into Chinese as shown in Yang's work: [昏 (confused) 头 (head) + 昏 (confused) 脑 (mind)] for *abatanar y aporrear el sentido* (beat and pound the sense); [难 (difficult) 分 (separate) + 难 (difficult) 解 (split)] for *enclavijadas y trabadas* (interlocked and joined).

In the case of 昏头昏脑, the MP is used to describe the confused state of mind after the beating and pounding of senses, rather than a direct translation of the source text expression. Therefore, it should be seen as a cause/effect extension of the source text content, which is very different from the phrasing strategies analyzed above. In the case of 难分难解, the semantic structure of the MP is a double negation of the act of separating or splitting. From a different point of view, this actually illustrates the closeness between two ships at sea as described in (las proas) *enclavijadas y trabadas*.

The reason for the kind of structural modifications which occurs in Yang's translation of ES1 and ES2 may be found in the very limited space provided by the four-character patterns, which thus requires the translator to be highly economical with morphemes or words to achieve the maximum communicative effect. When following a target-language-orientated approach to translation, in order to optimize the assimilation of source linguistic elements into the target language, it would seem that the target linguistic structures may well gain priority over a faithful transmission of the source text reference.

3.4.3 Structurally Symmetrical FCEXs (SSYs)

Yang's use of structurally symmetrical four-character expressions may seem to be quite arbitrary in many cases. Such use does not even seem to have been triggered off by certain source text features. Whereas Liu, by following a more literal approach to the source texts has been able to give a rendering which sound quite natural in the target language.

ES1 Pero lo que a mí más me fatigaba era el ver ir a pie a Zoraida por aquellas asperezas (Chapter XLI, I)

EN: What most troubled me was seeing Zoraida walking on that harsh terrain...

Y: 我最难受的是瞧索赖达一脚高一脚低的登山越岭。

L: 不过，最让我难受的就是看着索赖达在这崎岖的路上行走。

In translating ES1 *ir a pie por aquellas asperas* (walking along rugged path), while Liu's translation has been largely literal, Yang has used a typical SSY [登 (climb) 山 (mountains) + 越 (cross

over) 岭 (mountains)] to describe the hardships and difficulties undergone by Zoraida during her escape from her homeland to Spain together with the *cautivo* (the captured one).

ES2 ¿cómo es posible que pone vuestra merced en duda el casarse con tan alta princesa como aquésta? (Chapter XXX, I)

EN: How can your grace have any doubts about marrying a princess as noble as this one?

Y: 跟这样高贵的公主结婚还有什么 推三阻四 的?

L: 同这样一位高贵的公主结婚，您还有什么可 犹豫 的?

Similarly, in translating ES2 *poner en duda* (put ... in doubt), whereas Liu has translated the source text simply as 犹豫 (doubt), Yang has created a SSY [推 (push) 三 (three) + 阻 (impede) 四 (four)] to describe the hesitation in which don Quijote finds himself over marrying Dulcinea or Dorotea. It should be pointed out that the use of the second and fourth morphemes, 三 (three) and 四 (four), within the structure of the SSY follows from the idiomatic construct [...三 ...四], whereby *three* and *four* are used as adverbs rather than nouns and have an implicit derogative connotation when modifying verbs, such as doing things carelessly, behaving dishonestly, remaining indecisive in making a decision, etc.

ES3 Dijo que, así como Luscinda se vio en su poder, perdió todos los sentidos; y que, después de vuelta en sí, no había hecho otra cosa sino llorar y suspirar, sin hablar palabra alguna. (Chapter XXXVI, I)

EN: He said that as soon as Luscinda found herself in his power, she had fallen into a deep swoon, and when she regained consciousness she had done nothing but weep and sigh and had not spoken a single word

Y: 据他说，陆莘达瞧自己落在他手里，就晕死过去；清醒之后，只是 淌眼抹泪，唉声叹气，没说过一句话。

L: 卢辛达发现自己到了费尔南多手里，顿时晕死过去，醒来后，也只是 边哭边哀叹，一句话也不说。

The SSY [淌 (drip) 眼 (eyes) + 抹 (wipe off) 泪 (tears)] used by Yang in translating the single verb *llorar* (cry) is a good example of her preferred modification of the source text at a morpho-syntactic level. The formation of the SSY, however, has been more flexible this time in that it is achieved through splitting the disyllabic word 眼泪 (tears) into 眼 (eyes) and 泪 (tears) and then filling in the underlying four-character pattern with two monosyllabic verbal morphemes, 淌 (drip) and 抹 (wipe off), which tend to be used as collocations with 眼泪 (tears).

The construction of the SSY as such makes it clear that the intention of the translator is to put more emphasis on the source text *llorar*, by transforming a single verb into a sequence of actions and thus making the target text more vivid and descriptive. Such a phrasing strategy may fit in well with the scenario in which Cardenio recalls with great pains the sorrow and sadness endured by his beloved Lucinda when she found herself in the arms of Fernando, the man from a powerful aristocratic family who enforced her to marry him regardless of the long commitment between the two lovers.

ES4 Decíanme mis padres que en sola mi virtud y bondad dejaban y depositaban su honra y fama, y que considerase la desigualdad que había entre mí y don Fernando, y que por aquí echaría de ver que sus

pensamientos, aunque él dijese otra cosa, mas se encaminaban a su gusto que a mi provecho (Chapter XXVIII, I)

EN: My parents would tell me that their honor and reputation had been placed for safekeeping in my virtue and chastity, and I should consider the difference in the rank between me and Don Fernando, which would allow me to see that his thoughts, although he said otherwise, were directed more toward his pleasures than my benefit ...

Y: 我爹妈对我说，他们全靠我的贞洁来保全他们的声名体面。他们叫我别忘记自己和堂费南铎的门第太不相称；只要明白这一点，就能看出他尽管嘴里说得天花乱坠，心上是只图寻欢取乐，并没有顾到我的幸福。

L: “父母常常对我说，我的品行牵涉到他们的声誉，他们要我注意到我同费尔南多之间的差距。从这儿可以看出他们考虑的是他们的好恶，而不是我的利益。当然，这是另外一回事了。

In the example quoted above, there is a great deal of difference between Yang's and Liu's translation of *se caminaban a su gusto* (his intentions had for their aim his own pleasure). On the one hand, Liu's translation is 他们考虑的是他们的好恶 (what they (Dorotea's parents) were worrying about was their self-gratification); on the other hand, Yang has translated it as 寻欢取乐 (seek pleasure and make merry), referring to Fernando's indulgence in sensual pleasures. The divergence of understanding between the two translators may have been caused by the complex sentence structure of the original text and the sophisticated commentary made by Dorotea, which involves both her attitude towards her parents' discretion at maintaining the family's honour at the expense of her interest, and her disappointment at the dissolute nature of Fernando's love.

ES5.... que ella sin ser forzada ofrecía, por todas las partes de su fértil y espacioso seno, lo que pudiese hartar, sustentar y deleitar a los hijos que entonces la poseían (Chapter XI, I)

EN: ...for she, without being forced, offered up, everywhere across her broad and fertile bosom, whatever would satisfy, sustain, and delight the children who then possessed her.

Y: 她不用强迫，她那丰厚宽阔的胸膛，处处贡献出东西来，使她的儿女能吃饱喝足，生存享乐。现在这群儿女做了妈妈的主人了。

L: 而她却心甘情愿地用富庶辽阔的胸膛所拥有的一切来喂养和愉悦那些拥有她的儿女们。

The use of SSY by Yang in translating ES5 is slightly different from the cases analyzed above in that for the first time, there seems to be certain structural analogy between the source and target text. The juxtaposition of the three source text verbs *hartar, sustentar y deleitar* (satisfiate, sustain and please), when translated into Chinese by Yang, is embedded in two SSYs [吃 (eat) 饱 (full) + 喝 (drink) 足 (sufficient)] and [生存 (survive) 享乐 (enjoy)]. Meanwhile, Liu's translation has been on the main word by word.

3.5 Conclusion

In Chapter III, my main objective has been to process the linguistic database mined from the parallel corpus with a view to bringing to the surface the underlying phraseological patterns in the two Chinese translations of *Don Quijote* (I). Due to the very large number of four character segments extracted automatically from the two Chinese target texts, only three hundred samples have been randomly selected from each Chinese translation for comparison purposes. The statistical procedure

Pearson's Product-Movement test shows that despite the acknowledged relationship between the two Chinese versions of *Don Quijote*, they differ significantly from each other in terms of the distribution of the four-character phraseological categories highlighted in the current study.

Two possible sociolinguistic factors have been put forward to explain such divergence between the two target texts. First, it is assumed that given the fast evolving nature of the Chinese language, the gap of nearly twenty years may well have given rise to different linguistic habits as identified in the two Chinese translations. Secondly, it is suggested that the contrasting use of typical Chinese phraseological devices between the two translators may also be a reflection of their individual language style.

To test the first hypothesis, I compared the distribution of four-character idioms in two diachronically successive Chinese monolingual corpora, i.e. the LCMC and the UCLA corpus. The comparison of relevant corpus material revealed that there is a general trend in the later UCLA corpus in which while the presence of idioms in non-fictional text genres, e.g. reports, official documents or news reportage is sharply reduced, the use of idioms in fictional genres such as science fiction, general fiction and popular lore is greatly enhanced. It is an important finding regarding the evolving nature of the Chinese language. However, since the time span that the two monolingual corpora cover, i.e. from early 1990s to early twenty-first century, does not quite fall within the time span covered by the two Chinese translations, i.e. from late seventies to nineties, the insights gained from a comparable corpus study may still be used as reference material, rather than providing a direct answer to our initial research inquiry – a situation which is commonly seen in corpus-based language research as it is constrained to a great extent by the availability (or lack) of suitable corpus material.

Another compelling explanation for the observed differences between Yang's and Liu's use of four-character expressions must surely be their personal stylistic preferences. A distinctive phraseological feature that separates the two Chinese translations is Yang's higher use of morpho-syntactically patterned four-character expressions, which include semantically bipartite phrases, morphologically patterned phrases and structurally symmetrically phrases. A close look at Liu's use of these three typical Chinese phrases seems to point in the direction of a deliberate liberal and substantial reduction in the deployment of these phraseological devices.

First, the use of morphologically patterned phrases (MPs) in Yang's work exhibits a maximum degree of consideration towards the source text features. It shows that the insertion of MPs in the Chinese target texts has played a major role in adjusting the Castilian grammatical or functional particles to the target linguistic systems by such means as repetition, plurality, different degree of emphasis, use of bi-synonymous structures, etc. Given the huge linguistic contrasts between the two languages, the use of MPs reflects a highly skilful manipulation of the target language on the part of the Chinese translator, Yang.

ES camina, no se esfogue con la tardanza (prepositional phrase) el fuego de la cólera que tengo, y se pase en amenazas y maldiciones la justa venganza que espero (Chapter XXXIV, I)

EN: Go, run, hurry, make haste; do not allow delay to cool the flames of rage that I feel or the righteous vengeance I hope for to dwindle into mere threats and curses

Y: 我期待着一场理直气壮的报复呢，你别拖拖拉拉 (verb) 泄了我的火气，弄得一场报复化作几句恫吓和咒骂

L: 你快去叫他，免得耽误了时间 (present perfect, verbal phrase)，让我的怒火熄灭，把我这种正义的报仇仅仅转为几句吓唬和诅咒的话。

The use of verbal MP [拖拖 (pull-pull) + 拉拉 (drag-drag)] in Yang's translation of the original text is probably due to a lack of similar prepositional expression for *con la tardanza* (with the delay) in Chinese. As a result, in translating the source text expression Yang has transformed radically the typical Spanish reflexive verbal phrase *esfogarse el fuego* (the fire goes out) into an infinitive structure in Chinese [you (Leonela, the maid) do not 拖拖拉拉 (delay) to put off the fire of my anger], shifting the subject from the inanimate entity *fire* to the servant. Liu's translation has basically followed the same grammatical change and the only difference is that the central part of Liu's translation is a conjugated verb in the present perfect tense, instead of a morphologically patterned FCEX in indefinite form as found in Yang's translation.

Secondly, Yang's use of semantically bipartite four-character expressions (SBs) displays certain flexibility or originality in transferring the source text information into the target linguistic system. In some cases, Yang's choice of SBs seems to have been triggered off by source text expressions containing repetitive or bi-synonymous structures; whereas at other times, there is hardly any connection between the source and target texts at the linguistic level.

ES Viendo esto, hablé a doce españoles, todos valientes hombres del remo, y de aquellos que más libremente podían salir de la ciudad... (Chapter XLI, I)

EN: Consequently I spoke to twelve Spaniards, all of them valiant oarsmen who could leave the city without difficulty

Y: 我听了就去找到十二个西班牙人，都是身强力壮的划手，可以自由出城的

L: 我就找了十二个西班牙人，他们都是划船能手，人也勇敢，而且都能自由出城。

The use of semantically bipartite phrases by Yang in these two cases resemble each other in that the semantically symmetrical structures used in the target texts do not appear to have a formal equivalent in the source language as the parallel texts show. In translating the source text, the use of the SB structure [身 (body) 强 (strong) + 力 (strength) 壮 (grant)] by Yang turns out to be more descriptive than the literal translation by Liu, thus enhancing the tone of confidence of the speaker – the *cautivo* (the captive), recalling his adventurous escaping from the territories of the moros.

Furthermore, a point worth making in this regard is Yang's apparently liberal and rather arbitrary use of structurally symmetrical four character expressions (SSYs). In all of the five samples

selected randomly from the parallel corpus, Yang has used this highly elaborate Chinese rhetorical device when translating original excerpts which do not show any consistent phraseological pattern. One may conclude, therefore, that Prof. Yang is deliberately intent on modifying and polishing her version of *Don Quijote* to make it sound in tune with the unique linguistic habits of contemporary Mandarin Chinese.

Chapter IV General Phraseological Patterns in Liu's Translation

4.1 Liu's phraseological preference for figurative idioms (FIGs)

It is clearly apparent from Diagram VI (see Section 3.4) that in Categories FIG (figurative idioms), I-FIG (instantiated figurative idioms), AI (archaic idioms) and I-AI (instantiated archaic idioms), the red bar (Liu) shows a general tendency to run above its blue counterpart (Yang). The more frequent use by Liu of Chinese idioms, e.g. archaic expressions and figurative idioms (both including their variants), seems to suggest that in translating *Don Quijote*, Liu has delivered a kind of writing that carries more idiomatic weight than Yang's. The types of FCEXs preferred by Liu, i.e. archaic expressions and figurative idioms, represent together a much more conventionalized version in the target language. Such finding seems to follow the general pattern that has been discovered through the comparison of the two chronologically-separated Chinese corpora, i.e. the enhanced use of idioms in modern Chinese fiction.

This may well be a rather significant finding as regards the special nature of the language style as moulded in Liu's work. If we regard the cross-linguistic phenomena reflected in Yang's higher use of the first three FCEX categories as a structural modification operating at the surface level – given that the changes made are largely morpho-syntactic –, the kind of cross-linguistic transformation we observe in Liu's writing seems to favour a higher level of language manipulation, as indicated by his increased use of idiomatic expressions including figurative and archaic idioms.

Don Quijote, as an early seventeenth-century Castilian masterpiece, is well known for its rich vein of figurative expressions, idioms, proverbs, intertextual references to Greek and Latin classic, as well as Biblical events and so on. Unfortunately, more often than not, the figurative element in Cervantes's language is well hidden to the gaze of a contemporary target audience. This is, obviously so, when taking into account the huge cultural, geographical and diachronic gap between Cervantes's reading public and Liu's contemporary Chinese readers. It is precisely this huge gap that Liu is attempting to bridge by exploring Chinese figurative idioms to create a reading experience which his readers might be more familiar with.

I shall try now to analyze in some detail the different use of figurative idioms by two translators in the face of a large variety of idiomatic expressions very present in the source text, such as cognate and non-cognate metaphors, metonymy, personification, Latinisms, intertextual references to classical authors and events, proverbs, etc. A rather interesting impression that comes to mind when first contrasting idiomatic expressions in the source text and the use of figurative idioms in the two modern Chinese translations is that while the figurative nature of the source text has generally proved to be a direct factor in triggering off the use of figurative idioms in the target texts, it cannot explain fully all the instances of figurative idioms identified in the two Chinese translations. For example, in a number of cases where the source text involves no more than a single grammatical item like an

adjective or verb or noun or prepositional phrase, its corresponding translation may well be a typical Chinese figurative idiom.

In the next chapter (5) in order to gain a better understanding of the distribution pattern of figurative idioms in Liu's work, in relation to the source text as well as to Yang's first Chinese version of the novel, I will attempt to analyse these distribution patterns by statistically modelling the corpus data. But before that, we shall first have a closer look at the way Liu handles Chinese figurative idioms to deal with a large variety of source text figurative expressions, such as Chinese/Spanish (non-) cognate metaphors, metonymy, classical references, historical events, etc.

4.2 Cognate metaphors

Most observers will agree today that all human languages are essentially metaphorical in that the use of this rhetorical device (figure of speech) forms an integral part of the cognitive mechanism by which knowledge acquired in one domain may normally be used to explain the unknown in another (Lakoff, 1993: 205-51; 1995: 177-213). As we have discussed in Chapter II when trying to establish a typological framework for Chinese four-character expressions, idioms is a very complex phraseological concept which can be hardly broken down into a definite set of dimensions as languages may vary one from another. In my view, to put it simply, idioms may be seen as a general term to refer to figures of speech which transfer meaning from one domain to another, from the physical to the abstract. An idiomatic expression may convey the journey from word to concept, a part to the total, e.g. metonymy and synecdoche.

Inherent to the cognitive complexity, a cognate metaphor is a special phenomenon within the general field of figurative language. A cognate metaphor may be defined as such wherever the selection of conveyor or signifier referring to the object being metaphorized is almost identical in two languages. Cognate metaphors are much more prevalent among languages in close proximity, etymological, geographical or timewise, than among languages with little or no visible connection such as Spanish and Chinese. It may therefore be said that the awareness of the availability of cognate figurative idioms in the target text may well differ considerably from one translator to another. Constant resorting to this linguistic device surely indicates a conscious effort made on the part of the Chinese translator to incorporate the source text elements into the semantic/cognitive network of the target language.

1) Value

ES1 ¿Qué me ha de suceder -respondió Sancho-, sino el haber perdido de una mano a otra, en un estante, tres pollinos, que cada uno era como un castillo? (Chapter XXVI, I)

EN: "What else could have happened," responded Sancho, "except that from one moment to the next, in an instant, I've lost three donkeys, each one has sturdy as a castle?"

L: "怎么回事?" 桑乔说, "转眼之间我就丢了三头驴。每头驴都价值连城。"

Y: 桑丘说: "出了什么事吗? 我一换手、一眨眼的功夫, 丢失了三匹驴驹子, 每一匹都抵得一座大房子呢。"

First of all, it appears that metaphors relating to value of concrete objects by means of plots of land or units of money are quite similar between the two languages. In translating relevant source text metaphors, there are certain figurative idioms that are readily accessible to the Chinese translator. For example, in translating ES1, *cada uno era como un castillo* (each one was worth a castle), Liu has used a typical Chinese figurative idiom 价值连城 (worth several cities), while Yang has translated it plainly as 抵得一座大房子 (worth a big house).

2) Obedience or show of respect

ES2 Anselmo le replicó que aquél era su gusto, y que no tenía más que hacer que bajar la cabeza y obedecelle (Chapter XXXIII, I)

EN: The innkeeper wanted to bar the gate as soon as he saw them outside, but the blanket tossers did not agree, for they were people who would not have cared an ardite even if Don Quixote really had been one of the knights errant of the Round Table.

L: 安塞尔莫说他愿意这样，她应该做的就是俯首听命。

Y: 安塞尔莫说，他爱这么安排，她只消依顺就行

In describing one's submission or obedience to others, both Spanish and Chinese may draw for the "carried over" meaning on the bodily gesture *bow one's head*. Yang has translated the source text *bajar la cabeza y obedecelle* (lower her head and obey him) through a plain explanation 她只消依顺就行 (it will be all right so long as she obeys). By contrast, Liu has introduced the Chinese figurative idiom 俯首听命 (bow one's head and obey submissively) into his version, which has tactically enhanced the originality of the target language without sacrificing a great deal of the source text content. Also, the figurative idiom 俯首听命 has an implicit derogatory connotation in a Chinese textual context in the way it is often used to describe the negative and unconditional subordination that a woman owes to the man in a traditional husband and wife relationship.

The perceived pragmatic function in the target text echoes the socio-cultural event described in the source text. The source excerpt appears towards the opening part of Chapter XXXIII *Donde se cuenta la novela del curioso impertinente*, where Anselmo, obsessed by the idea of putting to the test the moral integrity and fidelity of his wife Camilla, after persuading his best friend Lotario to try to seduce Camilla, he subsequently instructs Camilla to entertain Lotario while he is away, deliberately. Camilla, "como una mujer discreta y honrada", despite her opposition to such arrangement, feels obliged to follow her husband's instructions. Here, the use of the Chinese figurative idiom 俯首听命 transcribes the situation described in the source text into a traditional Chinese hierarchical and possessive relationship between husband and wife.

4.3 Non-cognate metaphor

When translating source texts which do not appear to evoke cognate metaphors in Chinese, as

in the cases shown below, Liu's first choice is to substitute the source text expressions with typical Chinese figurative idioms, notwithstanding the different conceptual schema operating in establishing some kind of correlation between the concrete event or object and the abstract concept arising from it. Such conceptual transformation has been possible largely due to the fact that although localized figurative expressions of abstract concepts may be influenced by socio-cultural factors of a particular language community, and the view of the physical world outside may vary in substance or degree both geographically and diachronically, the mental or conceptual experience of all human beings is basically similar (Mac Cormac, 1998: 67).

In this section, we may observe that (1) Liu seems to be more concerned with an immediate assimilation of the idiomatic expressions from the source text, as well as the socio-cultural context which positions them in the real world, through a spontaneous use of typical Chinese figurative idioms; and (2) Yang seems to be more in favour of a literal approach to the source text, but usually provides a translator's note explaining the actual meaning of the original metaphor. As a result, when compared with Liu, Yang's work sounds more like a literal translation, as evidenced by her minimal interference with the semantic configuration of the original in the target context and the addition of very detailed annotations to make her version more accessible to prospective readers.

(1) Moral integrity

ES3 Porque mi Dulcinea del Toboso osaré yo jurar que no ha visto en todos los días de su vida moro alguno, así como él es, en su mismo traje, y que se está hoy como la madre que la parió; (Chapter XXVI, I)

EN: Because I shall go so far as to swear that my Dulcinea of Toboso has not in all her days seen a single Moor just as he is, in his own clothing, and that she is today as she was on the day she was born.

L: 我敢发誓，托博索的杜尔西内亚这辈子从未见过一个穿着摩尔人衣服的摩尔人。她至今仍 守身如玉。

Y: 我可以打赌，我的杜尔西内娅台尔托波索一辈子也没看见过一个穿摩尔服装的地道摩尔人，她现在就象生她的妈妈一样。(作者不止一次用这话来形容童贞女子(参看本书第一部第九章)，有的本子把这句话改为"就象她刚从妈妈肚里生出来的时候一样"，但这样是改掉了原文。作者好象是故意这样说的。)

Moral integrity is a recurrent motif in *Don Quijote*, as the idealization of Dulcinea constitutes the central strategy in the process of don Quijote's self-realization as an archetypal medieval knight-errant. To stress Dulcinea's maidenhood, don Quijote returns to the expression *como la madre que la parió* (as the day she was born). Rico notes that such an expression was a typical joke current at the time to refer to young ladies in books of chivalry (Rico, 1998: Chapter XXVI, note 8). As a conventional allusion in the literary genre, the expression *como la madre que la parió* is often used to emphasize an assertion, or in extreme cases, deprived of its semantic meaning, is used purely as an expression of surprise or exclamation (Ibid: Chapter IX, note 18).

Despite its modern currency as an ethically-loaded expression in Spanish, it does not have a valid counterpart in Chinese, where the figurative representation of a woman's spiritual purity may be

verbalized by comparing it to structurally stable and valuable objects such as precious stones or jades. Well aware of the cognitive-linguistic difference existing between the two languages, Liu has translated the source text into a typical Chinese figurative idiom 守身如玉 (keep one's reputation as spotless as jade or cherish one's integrity like a precious jewel). The pragmatic function performed by the Chinese figurative idiom in the target context not only has removed the socio-cultural barriers to understanding implicit in the original expression but most importantly, it may well evoke in readers' mind similar images of chaste women in traditional Chinese literary works.

Yang, on the other hand, has rendered it as 她现在就象生她的妈妈一样 (she is now like her mother who gave birth to her), which makes it inaccurate as a translation. The reason for Yang's mistranslation may be found in the subsequent translator's note, where she goes on to explain: "the author uses this expression more than once to refer to maidens, see Part I, Chapter IX, in some versions it is translated as *she is just like when she was first born from her mother's womb*. However, this is an alteration of the original text. It seems that the author has written as such deliberately, if it is not due to a misunderstanding of the source text expression.

ES4 Soy la que, encerrada en los límites de la honestidad, vivió vida contenta hasta que, a las voces de tus importunidades, y, al parecer, justos y amorosos sentimientos, abrió las puertas de su recato y te entregó las llaves de su libertad (Chapter XXXVI, I)

EN: I am the one who, secluded and surrounded by virtue, lived a happy life until, heeding your urgent words and what seemed to be fitting and loving sentiments, opened the doors of her modesty and handed you the keys to her freedom...

L: 我过去深居闺阁，无忧无虑，直到后来，在你似乎正当的纠缠骚扰下，向你敞开了我贞洁的大门，把我的自由的钥匙交给你，以身相许，结果得到的却是忘恩负义。

Y: 我向来贞静，日子过得快活，直到我听了你的央求，看了你表面上正当热烈的情感，才败坏了操守，把身心交付给你。我落到目前的境地，又看到你这会儿的情况，知道你全没有把我放在心上。

The Dorotea-Fernando episode plays a key role in the development of another main motif of the novel, i.e. the triumph of love over social constraints. Dorotea, is from a well-off peasant family and her parents are "labradores, gente llana, sin mezcla de alguna raza mal sonante, y, como suele decirse, cristianos viejos rancios; pero tan ricos que su riqueza y magnífico trato les va poco a poco adquiriendo nombre de hidalgos, y aun de caballeros". On the other hand, Fernando, while living in the same area as Dorotea, is the younger son of a local notable, "un duque, que le hace uno de los que llaman grandes en España".

The source text excerpt under consideration belongs to the opening part of Chapter XXXVIII, where Dorotea laments Fernando's betrayal so soon after the first night the two lovers spent together. In her autobiographical style of narration, Dorotea uses a euphemistic expression to refer to the most significant event in her life, her surrender to Fernando: *la que te entregó las llaves de su libertad* (she is the one.... who gave you the keys to her freedom). Although it may not be impossible to infer the literal meaning of this metaphorical expression from the sentence context, to refer to maidenhood

with keys to one's freedom is quite unusual in the context of the target cultural background. In facilitating the understanding of the source text, Yang has provided a simple explanation 把身心交付给你 (give my body and soul to you), implicitly referring to the pre-nuptial relationship between the two lovers.

To solve this problem, Liu has compromised in his approach, combining a literal translation with a Chinese figurative idiom 把我的自由的钥匙交给了你, 以身相许 (FIG) (I have given you the key to my freedom + FIG: pledge to marry someone). The use of the Chinese figurative idiom 以身相许 in the target text has a strong pragmatic connotation. It is often used to convey a maiden's pledge of love and loyalty to her lover. But the result of their love affair may well turn out to be tragic given the young lady's lack of experience or Fernando's betrayal. In other words, the pragmatic function of the Chinese figurative idiom is twofold: in the first place, it depicts the innocent and accommodating image of Dorotea for love's sake, who seems to be overwhelmed by the fervent love of an upper-class man; and secondly, its use in the target text is foreboding of the later abandonment of the young girl by Fernando once his desire is gratified.

(2) Destiny & philosophy of life

ES5 Mas, como pocas veces, o nunca, viene el bien puro y sencillo, sin ser acompañado o seguido de algún mal que le turbe o sobresalte (Chapter XLI, I)

EN: But good seldom or never comes pure and simple, unaccompanied by some evil to throw it into disarray, and as our fortune would have it.

L: 可是好事很少或从来没有一帆风顺的, 总要伴随一些节外生枝的事情。

Y: 可是 一竿子到底的好运是绝无仅有的, 好运总穿插着坏运, 吉凶总相伴相随。

ES5 comes from Chapter XLI *Donde todavía prosigue el cautivo su suceso*. It is the second half of the interpolated autobiography of the famous Leonese captive, which largely draws on Cervantes's own life, his participation in the Battle of Lepanto (1571) and his subsequent capture and arrest by Turkish soldiers and long captivity in Algeria. In this chapter, the captive recalls his adventurous escape from Algeria with his beloved Zoraida, an attractive and devout Moorish woman. The return journey by sea was full of thrills and unexpected dangers, among others, sea storms, the chasing of Turkish soldiers, the encounter with French pirates, the confrontation with Castilian patrolling horseback riders, etc.

The captive in Chapter XLI, after a short period of sailing in a tail wind, foretells their encounter with French pirates, who were notorious for their greediness: "...éstos, sin duda, son cosarios franceses, que hacen a toda ropa". The expression used in this excerpt (ES5) is made of two subordinate parts: *el bien puro y sencillo* (pure and unadulterated good) and *el bien acompañado o seguido de algún mal que le turbe o sobresalte* (good attended or followed by some disturbing or socking evil). The personification of abstract concepts such as good or evil exemplifies the captive's mentality as a Christian and accentuates the great uncertainty which accompanies him throughout his

flight from captivity.

Despite the wide currency of exemplary themes against the cultural-religious backdrop of the source text, the meta-linguistic shapes that the source language takes to represent the notions of good and evil are not common in the target language. As a result, in translating ES5, Liu has employed two Chinese figurative idioms that tend to be used under similar circumstances. They have been constructed within very different schemes: 一帆风顺 (a full sail of tail wind), which means everything proceeds successfully; and 节外生枝 (the growth of branches over knots), which means present troubles giving rise to new ones. Unlike Liu, Yang has compromised in his approach to the source text figurative expression, mixing literal translation, regional dialect and plain explicitation, with a view to conciliate her desire to facilitate understanding of the target text with the maximum retention of the original expression in the source text.

ES6 Duró este engaño algunos días, hasta que, al cabo de pocos meses, volvió Fortuna su rueda y salió a plaza la maldad con tanto artificio hasta allí cubierta, y a Anselmo le costó la vida su impertinente curiosidad (Chapter XXXIV, I)

EN: The deception lasted some months until Fortune spun her wheel, the wickedness they had concealed with so much skill was made public, and Anselmo's reckless curiosity cost him his life.

L: 这个骗局持续了一段时间。几个月后，命运女神扭转乾坤，他们精心设置的骗局昭然若揭，

Y: 事情一时上没闹破，直到几个月后，命运的轮子转了过来，掩盖得非常巧妙的丑事就传扬开了。

The various figurative shapes that different languages adopt when trying to explain the functioning of the universe and life circles lies at the heart of individual cultural entities. According to Cárcer and De Sobies (1916: 233), the *rueda de la fortuna* refers here to the “inconstancia y poca estabilidad de las cosas humanas en lo próspero y en lo adverso”. The comparison of fortune to a rolling wheel to describe the uncertainty of life characterizes the language Cervantes used in *Don Quijote*: sometimes in the words of the author as a comment on a discourse event, sometimes spoken by one of the characters in the novel.

The source text expression *volvió Fortuna su rueda* (Fortune gave a turn to the wheel) appears towards the end of Chapter XXXIV *Donde se prosigue la novela del Curioso impertinente*. In this passage, it is a third-person narrator who rather gleefully comments upon the imminent disclosure of the well-concealed love affair between Camilla, Anselmo's wife, and Lotario, Anselmo's best friend. In Chapter XLVII, Sancho uses a similar expression in a sequence of proverbs addressed to the priest: “...pero ya veo que es verdad lo que se dice por ahí: que la rueda de la Fortuna anda más lista que una rueda de molino, y que los que ayer estaban en pinganitos hoy están por el suelo”. Here, Sancho is warning the priest that his mistreatment of don Quijote by confining him in a cage fastened to an ox-cart was not a good idea; no good could come out of it.

As far as I know, there is no similar expression in Chinese that could bring together the capricious nature of Fortune and the circular motion of the wheel. Hence, a literal translation of the

source text, just as Yang did, would appear to be at odds with the linguistic habits of the target language, though it is not impossible to grasp its literal meaning within the context. Liu, on the other hand, has created a joint figurative phrase to make up for the lack of equivalent expression in the target language: 命运女神 扭转乾坤 (FIG) (The Goddess of Destiny + FIG: move heaven and earth; bring about a radical turnabout in the course of events).

It should be pointed out that Liu's idiosyncratic use of 扭转乾坤 when translating ES2 does not seem to be stylistically appropriate on this occasion, for the figurative idiom 扭转乾坤 is often associated with decisions or actions taken to change the course of history while rarely applies to daily-life settings. Moreover, the figurative idiom 扭转乾坤 has a positive semantic connotation referring to a critical or momentous turn in history or a turn for the better in one's life. Such a use of the figurative idiom in the target textual context somehow manages, though not ideally, to depict the well-kept love affair between Anselmo's wife and the fact that his best friend is about to be thrown into the open, which eventually leads to Anselmo's personal tragedy.

ES7 ... si ya vuestra merced no gusta que su padre y mi señor la dé al otro mundo, porque no se puede esperar otra cosa de la pena con que queda por vuestra ausencia. (Chapter XLIV, I)

EN: ... unless you wish to see your father and my master in the next world, which is all that can be expected, considering the grief your absence has caused him.

L: 假如你不愿意让你的父亲即我的主人到极乐世界去的话。你的出走给你父亲带来的痛苦已经让他悲痛欲绝了。

Y: 除非你愿意把你的爸爸、我们的主人赶出人世；他为你出走，伤心得只有死路一条了。

The above excerpt comes from Chapter XLIV entitled *Donde se prosiguen los inauditos sucesos de la venta*. It tells the story of a young lord don Luis falling in love with doña Clara, who used to live with her father, Judge Juan Perez de Viedma, next door to him. He had followed doña Clara in disguise to beg for her hand in marriage and her father said he would consider the proposal, though their early age and different social status presented obvious difficulties for their love. Meanwhile, four horsemen had arrived at the inn looking for don Luis and ordered him to return home with them, for he had left home without his parents' permission and his father was therefore very seriously concerned about his whereabouts and well-being.

To give greater emphasis to his entreaties, the servant of don Luis' father has used the phrase *la dé al otro mundo* (to leave this world) as a euphemism for *death*. In translating the source text expression, Liu has imaginatively brought in a Chinese figurative idiom 极乐世界 (Sukhavati; Pure Land; Western Paradise), which dates its etymological roots back to Sanskrit or to Buddhist texts and has been assimilated into Chinese as a euphemism. Yang, on the other hand, has translated it as 把你的爸爸、我们的主人赶出人世 (expel your father and our master from this world), where the tone of the servant's words seems to be less considerate than the way Liu renders it.

(3) Feelings and emotions

The very different way in which the two languages refer to mental activities and emotional feelings, e.g. hesitation, love, fear, anger, etc., represents also a very challenging area for Chinese translators, while at the same time furnishing me with a wealth of first-hand material which should help greatly in the process of describing and analyzing the special characteristics that separate a translator's phraseological style from another's.

a. Fear

ES8 Digo que oyeron que daban unos golpes a compás, con un cierto crujir de hierros y cadenas, que, acompañados del furioso estruendo del agua, que pusieran pavor a cualquier otro corazón que no fuera el de don Quijote. (Chapter XX, I)

EN: I say that they heard the sound of rhythmic poundig, along with a certain clanking of irons and chains that, accompanied by the clamorous fury of the water, would have put terror in any heart other than Don Quixote's.

L: 他们听到的是一种铁锁链有节奏的撞击声，还伴随着水的咆哮声，除了唐吉诃德，任何人听到这种声音都会 毛骨悚然。

Y: 那是有节奏的敲打声，夹杂着铁片和铁链的碰擦声，再加上汹涌的水声。除了堂吉诃德，谁听了都会害怕的。

Early seventeenth-century Castilian and contemporary Mandarin Chinese use different means of representation to refer to feelings like fear or anger. ES8 above comes from Chapter XX which goes under the title *De la jamás vista ni oída aventura que con más poco peligro fue acabada del famoso caballero en el mundo como la que acabó el valeroso don Quijote de la Mancha*. It tells the story of don Quijote's disappointment over his imagined encounter with six fuelling-mill hammers. Continuing their thirsty travelling through the night, don Quijote and Sancho Panza arrived at a waterfall where they heard regular and heavy blows accompanied by the sounds of the waterfall. The third-person narrator observes that the terror caused by the strange thuds, has put the fear of God into their hearts: *pusieron pavor a corazón que no fuera el de don Quijote*.

To express abstract feelings such as terror into something really tangible (physical), the verb *poner* (put) comes in useful in the source text. However, in contemporary Mandarin Chinese such a literal translation of the source text would be most unlikely owing to the fact that the process of conceptualization of figurative speech in Chinese is rather different. Therefore to render the sentence ES8 into Chinese, Liu has brought in a typical figurative idiom which physically describes the effect of extreme horror upon the human body, i.e. 毛骨悚然 (fine hair standing on end and cold to marrow, the inside of the spine, the backbone feels cold; exceedingly frightened). Meanwhile, Yang has provided a plain explanation of the original metaphor: 谁听了都会害怕的 (anyone who hears these sounds will be scared to death).

b. Anger

ES9.... el cual, viéndose tratar mal de aquel villano malandrín, puesta la cólera en su punto y crujiéndole los huesos de su cuerpo, como mejor pudo él, asió al cuadrillero con entrambas manos de la garganta (Chapter XLV, I)

EN: “Nothing flows from her, vile rabble,” Don Quixote replied, burning with rage, “Nothing flows from her, I say, but amber and delicate musk; and she is not blind or humpbacked but as upright as a peak of the Guadarramas.”

L: 唐吉珂德 见这个坏蛋 竟敢 如此 对待 自己，立刻 气得 七窍生烟！他用 双手 紧紧 掐住了 团丁的 脖子。

Y: 堂吉珂德 瞧这 混蛋 对自己 撒野，火气冲天，浑身的骨头都要爆裂了。他 拚命 用两手 卡住 那 巡逻队员 的脖子

ES9 comes from Chapter XLV *Donde se acaba de averiguar la duda del yelmo de Mambrino y de la albarda, y otras aventuras sucedidas, con toda verdad*. In this chapter we witness how a member of the Santa Hermandad⁴² becomes aware that don Quijote was actually the person who set free the galley slaves in a previous chapter and therefore has to be brought to justice. On the contrary, don Quijote felt totally vindicated in setting them free from the clutches of the Santa Hermandad for the sake of social justice. The confrontation between don Quijote and La Santa Hermandad symbolizes once again the constant of conflict between the protagonist's world of chivalry, and the real world, in this case the forces of law and order in the service of the Crown (Fajardo, 1994: 46).

When don Quijote is accused of breaking the law he is enraged, utterly bursting with indignation *puesta la cólera en su punto y crujiéndole los huesos de su cuerpo* (worked up to the highest pitch of wrath, and all his joints cracking up with rage). For lack of a similar expression in Chinese, Liu uses a Chinese figurative idiom 七窍生烟, which means that one seethes with rage or that smoke bursts out of man's seven apertures, i.e. eyes, ears, nostrils and mouth. Etymologically the notion of 七窍 (seven apertures in the human head) dates back to *Zhuang Zi*, named after its author (BC 369- 268), who was the founder of Daoism, the most influential philosophical school in Chinese history.

The importance of the term 七窍 at the very centre of Chinese philosophy and culture lies in the fact that Daoism is the native philosophical system in China, unlike Buddhism which was imported from India. Furthermore, this notion illustrates most effectively the unique mode in which the Chinese people perceive and comprehend the complex mechanism at work in human emotions. Not surprisingly, the term 七窍 is commonly found in classical writings or textbooks of traditional Chinese medicine explaining a whole variety of clinical syndromes, and the idiom 七窍生烟 is therefore one of the most expressive ways of describing the feeling of rage and anger taking hold of don Quijote.

By contrast in Yang's translation of the passage we are looking at is a mix of a contextually instantiated variant of four-character expression 怒气冲天 and a literal translation of the second half of the original metaphor: 火气冲天 (I-FIG), 浑身的骨头都要爆裂了 (his anger flushes the sky [I-FIG], the

⁴² La Santa Hermandad is a rural police force created in the late fifteenth century by the Catholic Monarchs (Ferdinand and Isabella) to maintain law and order in the countryside.

bones of his whole body are about to rupture). Such a combined translation strategy is not a plain or dull rendering of the source text expression in the least. On the contrary, Liu's direct use of an expressive and very typical Chinese figurative idiom brings about a flavour to the target text that is definitely more idiomatic and vivacious.

4.4 Other figurative speeches in the ST

(1) Metonymy

Metonymy is an important figurative trope which consists in taking a characteristic feature of an object for its whole. For example, the substitution of tongue as oral muscle for the more abstract concepts of language or dialect; or the replacement of Westminster as a borough of the city of London for the Parliament of the United Kingdom, which is located in that area. The sample cited above illustrates the contrastive phrasing strategies taken by Yang and Liu in translating the source text metonymy which does not happen to have a formal equivalence in the target language.

ES10 que con toda solicitud y advertimiento miraba por la honra porque no pareciese mal al vulgo ocioso y a los ojos vagabundos y maliciosos la entrada de un mozo rico, gentilhomme y bien nacido, y de las buenas partes que él pensaba que tenía, en la casa de una mujer tan hermosa como Camila; (Chapter XXXIII, I)

EN: ... with the greatest care and vigilance, and contrived to pare and whittle down the number of visiting-days agreed upon, to prevent the idle vulgar herd with their wondering, malicious eyes from gossiping about the visits of a well-born young gentleman with the qualities he believed he possessed to the house of a woman as beautiful as Camilla.

L: 他知道自己有些优越条件，因而在一些游手好闲、别有用心的小人看来，一位如此富有、英俊而又出身高贵的小伙子出入一位像卡米拉这样漂亮女人的家，一定是件很有意思的事。

Y: 因为象他这样一个富贵公子，自己知道颇有几分人材，如果经常到卡蜜拉那样漂亮夫人的家里去，那些吃了闲饭没事干的人就不免歪言散语，恶意中伤。

The source text metonymy quoted above (ES10) is selected from Chapter XXXIII *Donde se cuenta la novela del Curioso impertinente*, which tells a moral tragedy derived from the complicated triangle relationship between Camila, Anselmo and Lotario. ES10 is located in the opening part of Chapter XXXIII, where Lotario finds himself in the impossible situation of wanting to honour his long-established friendship with Anselmo, while keeping a discrete distance with the latter to safeguard the reputation of his friend as a married man. That is because, as the third-person narrator comments, the frequent visits of Lotario, as a single, rich and highborn gentleman, to Anselmo's home would provide material of gossips to those *idle eyes* (ojos vagabundos).

Here, the reference to parts of the human body, i.e. eyes, to point to a certain group of people is a typical metonymic use of language, which does not really bear any validity in the target language. To resolve this problem at a meta-linguistic level, Liu has brought in a Chinese figurative idiom which is also metonymic in nature and functions as an appropriate substitute to the original metonymy in the target language: 游手-好闲 (free hands idle about); meanwhile, Yang has provided a plain expression: 吃

了 闲饭 没事干的人 (people without much to do after free meals).

(2) References to classical events, Latin expressions, proverbs, etc.

ES11 el cual alcanzado, a muchos hemos visto que, habiendo pasado por estas Sirtes y por estas Scilas y Caribdis, como llevados en vuelo de la favorable fortuna, digo que los hemos visto mandar y gobernar el mundo desde una silla (Chapter XXXVII, I)

EN: ... and once they obtain this, many are those we see who, having sailed over those Scyllas and Charybdises as if borne on the wings of favourable fortune, govern and rule the world from an armchair.

L: 他们 含辛茹苦，一旦达到了目标，就好像插上了时来运转的翅膀，开始坐在椅子上统治世界，

Y: 咱们看见许多人 历尽艰难困苦，到了这一步，就飞黄腾达了；咱们看见他们坐在安乐椅里统治世界

ES11 is selected from Chapter XXXVII *Que prosigue la historia de la famosa infanta Micomicona, con otras graciosas aventuras*. In this chapter, don Quijote embarks upon the comparison between a career as a writer and serving in the army. In the source text context of ES12, don Quijote portrays the life circle underscoring men with aspirations for personal wellbeing and success in their chosen official careers: they normally start off in a dire state of impoverishment, short of decent cloth or proper food, following a rough and hard course, they stumble here, fall there, and get up once more to fall again: “*Por este camino que he pintado, áspero y dificultoso, tropezando aquí, cayendo allí, levantándose acullá, tornando a caer acá*”. However, once all those difficulties have been overcome, their life sees a sudden qualitative change and their socio-economic status improves substantially, they rule and govern the world from a chair, their hunger turned into satiety, their cold into comfort, their nakedness into fine clothing, their sleep on a mat into idle repose in holland and damask: “*mandar y gobernar el mundo desde una silla, trocada su hambre en hartura, su frio en refrigerio, su desnudez en galas, y su dormir en una estera en reposar en holandas y damascos*”.

To describe the hardship that men of letters normally have to undergo to attain their professional aspirations, don Quijote uses the expression *por estas Sirtes y por estas Scilas y Caribdis*, which are geographical names in Greek mythology. In *Las Frases del Quijote*, there is a very detailed explanation of the origin of this expression: “... [e] ntre Scila y Caribdis. exp. fig. con que se explica la situación del que no puede evitar un peligro sin caer en otro⁴³” (Cárcer and De Sobies, 1916: 243). Three important aspects may be read into the meaning of this expression: (1) the obstacles underlying an enterprise; (2) the close proximity between the two obstacles; and (3) the inevitability of being dangerously faced with one or the other. As can be expected, such culturally loaded figurative expression in the source text is not readily available in the target language. Chinese translators would have to devise or identify some equivalent phrase on the basis of their personal understanding of the original and their particular style in the target language.

In translating ES11, Liu has exploited the Chinese figurative idiom 含辛茹苦 (containing

43 (D. O En.CO H. NO AM. NO) = Bancos y escollos de las costas de Africa é Italia que los poetas pintaron como muy temibles á los navegantes: significan aquí generalmente cualquier peligro. (CLEM.)

pungent stuff in mouth and chewing bitterness), which means enduring great hardships. If we were to map the figurative meaning of the Chinese idiom on to the scheme of the meta-function of the source text figurative expression as outlined above, we would find the Chinese figurative idiom has just covered the first ‘meaning’ of the expression, namely the obstacles facing an enterprise, without any semantic reference to the second and third aspects of the whole meaning implied in the source expression.

Similarly, Yang has translated ES11 literally as 历尽 艰难困苦 (experience all kinds of hardship), which is a verbal phrase containing a bipartite structure composed by two synonymous nouns 艰难困苦 (difficulty + poverty). As we can see, due to an obvious lack of comparable figurative expressions in the target language, both Yang and Liu have opted for a partially similar paraphrase of the source text figurative expression and the only difference between them in terms of the phraseological profile of their translation consists in that while Liu has proved to be consistent with his customary preference for typical Chinese idiomatic expressions, whereas Yang has again largely followed her usual literal approach to the source text.

ES12.... pero pensar que hemos de volver ahora a las ollas de Egipto, digo, a tomar nuestra cadena y a ponernos en camino del Toboso, digo, a tomar nuestra cadena y a ponernos en camino del Toboso.... (Chapter XXII, I)

EN:... but to think that we’re going back to the flesh-pots of Egypt, in other words picking up our chain again and setting for El Toboso, is like thinking it’s night-time already when it isn’t yet ten in the morning – it’s like trying to get figs from thistles.

L: 但是, 如果以为我们现已回到了 太平盛世, 可以拿着锁链去托博索了

Y: 您如要我们这会子 回到埃及的肉锅旁边去 ①, 换句话说, 要我们扛着这副链子到托波索的大道上去 (①《旧约全书》《出埃及记》第十六章 第三节, 记以色列人在旷野里挨饿, 埋怨说, 宁愿在埃及肉锅边吃得饱足时死去。“回到埃及的肉锅边去”, 一般指恋念过去丰足的生活, 这里指办不到的事)

ES12 is selected from Chapter XXII *De la libertad que dio don Quijote a muchos desdichados que mal de su grado los llevaban donde no quisieran ir*. In this chapter, don Quijote sets free a group of galley slaves from what he sees as an unjustified punishment for their alleged crimes and the mistreatment by their guards. Emulating chivalric traditions, don Quijote demands that the criminals that he has rescued from their captors present themselves to his beloved Dulcinea and extol his noble action to her – a request which is seen by those ungrateful slaves as an impossible mission. They repudiate don Quijote’s request for paying respect to Dulcinea as a form of repayment to his kindness. Ginés de Pasamonte, who is the leader of this motley crowd, and also the most outspoken, utterly refuses to comply with don Quijote and dismisses the whole thing as totally ludicrous, referring in the process to a famous biblical event in the *Old Testament*, i.e. the Israelites’ Exodus from Egypt. Cárcer and De Sobies (*Las Frases del Quijote*, 1916: 175) comments on this occasion:

Las ollas de Egipto... [a]lude a ... cuando murmurando los hijos de Israel contra Moisés porque les había traído al desierto, se acordaban de cuando estaban en Egipto sentados junto á las calderas ú ollas llenas de carne ...

To make up for the lack of a similar biblical reference in the target language, Liu has introduced the Chinese figurative idiom 太平盛世 (peaceful and prosperous age), which alludes to the Tang Dynasty (AD 618-907), known as the most prosperous dynasty in Chinese history. Yang, on the other hand, gives a literal translation 回到埃及的肉锅旁边去 (going back to the flesh-pots of Egypt), which is then followed by a very detailed footnote explaining the etymological origin of the source text expression in the *Old Testament* and an additional paraphrase to make sure the Chinese readers understand the expression correctly: “回到埃及的肉锅边去”，一般指恋念过去丰足的生活，这里指办不到的事 (“going back to the flesh-pots of Egypt”, normally refers to indulge oneself in the past life of abundance and prosperity; here it refers to do impossible things).

ES13... porque los buenos amigos han de probar a sus amigos y valerse dellos, como dijo un poeta, usque ad aras; que quiso decir que no se habían de valer de su amistad en cosas que fuesen contra Dios. (Chapter XXXIII, I)

EN: Good friends should verify their friendship and help each other usque ad aras, as the poet put it; which means that they mustn't use their friendship for purposes that are offensive to God [usque ad aras: Plutarch ascribes the phrase (“as far as the altar”) not to a poet but to the orator Pericles]

L: 好朋友之间应该彼此信任，就像一位诗人说的，光明磊落，不应该利用友谊做违反上帝意志的事情。

Y: 良朋好友之间的依赖和利用，应该象诗人所说的：‘能供在祭坛上’^①。

就是说，不该利用友谊干违反上帝的事。（“能供在祭坛上” (usque ad aras) 这是纪元前五世纪古希腊政治家贝利克雷 (Pericles) 的话，见普鲁 (Plutarch) 《杂文集》 (Moralia) 《论羞愧》 (De la malavergenza) 一文中塞万提斯记错了，以为是一个诗人的话。

ES13 is from Chapter XXXIII *Donde se cuenta la novela del Curioso impertinente*, which tells a moral story of how the long established male friendship between Anselmo and Lotario is eventually destroyed by Anselmo's jealousy and possessiveness over Camila, his attractive and newly-married wife. In an attempt to dissuade Anselmo from his perverse fantasy (tempting fate), Lotario, who repeatedly warns against the dangers of such a deceitful plan, quotes an aphorism from one of Plutarch's moral tracts (fifth century B.C. In his edition of *Don Quijote* Rodríguez Marín (p. 350) observes that, “*hasta los altares*”, *no parece que se trate de ningún poeta sino de Pericles, a quien atribuye estas palabras Plutarcho en uno de sus tratados morales*”.

In her version of ES13, Yang, who has largely drawn on Marín's edition of the novel, offers a literal rendition of the source text expression followed by a detailed explanation of the etymological origin of the Latinism in question. By contrast, Liu has taken out the expression *usque ad aras* totally from the target text, and brought in instead a typical Chinese figurative idiom to decode the pragmatic connotation of the figurative use of language in the source text. The Chinese figurative idiom employed is 光明磊落 (bright and candid), which has its origin in 晋书 (History of Jin Dynasty) dating back to the fourth century A. D. in Chinese history. The juxtaposition of two adjectives with visibly positive connotations is used to describe the frankness of one's mind, especially with reference to men's behaviour and attitude towards other people.

ES14 Con todo eso -dijo don Quijote-, mira, Sancho, lo que hablas, porque tantas veces va el cantarillo a la

fuelle..., y no te digo más. (Chapter XXX, I)

EN: "All the same", said Don Quixote, "you had better be careful what you say, Sancho, because the pitcher goes so often to the well....I need say no more".

L: “可也是，”唐吉珂德说，“你看你总是说，桑乔，喋喋不休，难免……行了，我不说了。”

Y: 堂吉珂德道：“可是，桑丘，你说什么话得仔细想想。因为‘水罐儿一次次到井边去’^①，底下我不说了。”（^①西班牙谚语：“水罐儿一次次到井边去，结果就砸碎了”。）

ES14 is selected from Chapter XXX *Que trata del gracioso artificio y orden que se tuvo en sacar a nuestro enamorado caballero de la asperísima penitencia en que se había puesto*. Despite the shared excitement by don Quijote and his squire at the foreseeable material gains promised by Princess Micomicona, the motives behind the acceptance of the commission by don Quijote and Sancho diverge greatly in terms of the alignment of positions between Dulcinea and Princess Micomicona: while don Quijote sees Micomicona as a noble customer worthy of his chivalrous attention but not in the least marrying her. Sancho is overwhelmingly tempted by the prospects of a union between his master and the legendary princess which could bring him fame and riches.

To convince don Quijote of the superior beauty of Princess Micomicona over Dulcinea del Toboso, Sancho laments: “*Es, por dicha, más hermosa mi señora Dulcinea? No, por cierto, ni aun con la mitad, y aun estoy por decir que no llega a su zapato de la que está delante*”. Hearing such blasphemous remarks about his beloved goddess, inexhaustible fountain of spiritual support in his knightly enterprise, don Quijote “*alzando el lanzón, sin hablalle palabra a Sancho y sin decirle esta boca es mía, le dio tales dos palos que dio con él en tierra; y si no fuera porque Dorotea le dio voces que no le diera más, sin duda le quitara allí la vida*”.

However, such serious warnings can hardly take Sancho’s mind away from the material gains to be had from marrying Princess Micomicona. In an attempt to soothe don Quijote’s ruffled feathers or at least to come to an understanding with his master, Sancho suggests that don Quijote should have Dulcinea as his mistress while accepting Princess Micomicona as his legitimate wife: “*cásese vuestra merced una por una con esta reina, ahora que la tenemos aquí como llovida del cielo, y después puede volverse con mi señora Dulcinea; que reyes debe de haber habido en el mundo que hayan sido amancebados*”. Sancho’s impertinence provokes don Quijote to issue a further warning *tantas veces va el cantarillo a la fuente*.... According to Cárcer and De Sobies, in *Las Frases del Quijote* (1916: 214), this is the meaning of the proverb:

Tantas veces va el cantarillo á la fuente ... el ref. Completo es: *Cantarillo que muchas veces va á la fuente, ó deja el asa, ó la fuente*; que advierte que el que frecuentemente se expone á las ocasiones peligra en ellas. (A.) = Otros terminan el refrán: *alguna vez se ha de quebrar*. = Nos advierte las fatales consecuencias que se siguen de no evitar los riesgos y ocasiones peligrosas.

It is clear from the above explanation that the repetitive action of bringing the water jug to the well in the source text implies the potential risk of eventually breaking the said container. Don Quijote

uses this proverb to warn Sancho off with the troubles that he could bring upon himself, if he goes on disparaging Dulcinea and praising Dorotea. In translating ES14, Liu has used the Chinese figurative idiom 喋喋不休 (talk endlessly), which normally annoys the listener and thus conveying a derogatory sense in the target language. On the other hand, Yang has rendered the source text literally as 水罐儿一次次到井边去..’ ①, 底下我不说了 (the water jug goes to the well time and time again... I will say no more). To help with the understanding of the source text, Yang goes on to provide a fuller version of the half-uttered proverb: “it is a Spanish proverb: the water jug goes to the well time and time again and finally cracks”, leaving ample space for the potential readers to infer the pragmatic meaning of the source proverb in the original textual context.

4.5 Liu’s phraseological preference for archaic idioms (AIs)

ES1 (don Quijote) Echemos, Panza amigo, *pelillos a la mar en esto de nuestras pendencias*, y dime ahora, sin tener cuenta con enojo ni rencor alguno... (Chapter XXX, I)

EN: Friend Panza, let us forgive and forget as to our quarrels, and tell me now, dismissing anger and irritation, where, how, and when did you find Dulcinea?

Liu: “桑乔朋友，咱们消释前嫌吧，别再争吵了。”

Yang: “潘沙朋友，咱俩争吵的事，从此撒开手别再计较了。你现在别生气，也别记恨...”

The source text expression appears in Chapter XXX (Part I), where we witness an encounter between the two main characters, but in a context in which the mood of communication between the two has already shifted dramatically from confrontation to apparent desire on the part of the master to engage his squire in conversation. It is again focused upon his idealized *señora* Dulcinea. In a previous scene, don Quijote dispatched Sancho to the village Toboso, where the peasant girl resides, to give her a message from the knight detailing his anxiety and paranoia for not being by her side. Now, on their way to Princess Micomicona’s kingdom, upon Sancho’s speedy return from the village, where the astute squire has actually never been to nor has he ever seen Dulcinea in person, don Quijote, totally unaware of what has happened on Sancho’s way back to the village, is dying to know what messages Dulcinea has asked Sancho to bring him back, as well as the most trivial details of her reaction after reading his love letter.

To translate the source text expression *pelillos a la mar en esto de nuestras pendencias*, whereas Yang has substituted the original metaphor with a simple explanation, Liu has opted for 消释前嫌, which is an idiosyncratic variant of the archaic idiom 不计前嫌 (discard-previous-animosity). An interesting point in the deployment of this Chinese archaic expression consists in that it tends to be used in a target language context where the speaker calls for conciliation between political parties or stakeholders seriously separated by their opposing points of view over fundamental issues. It is commonly seen as a form of diplomatic rhetoric in Chinese formal writings to convey one’s inspiration to the reconstruction of mutual understanding especially in foreign affairs. From there, we can see that the idiomatic expression 不计前嫌 implies the protagonist’s strong intention to convince, which shifted from implicit unease to a genuine attitude towards peace-making.

ES2 (don Quijote) —Pues es verdad —replicó don Quijote— que no acompaña esa grandeza y *la adorna con mil millones y gracias del alma*. Pero no me negarás, Sancho, una cosa: cuando llegaste junto a ella..... (Chapter XXXI, I)

EN: "Well!" said Don Quixote, "and doth she not of a truth accompany and adorn this greatness with a thousand million charms of mind!

Liu: “她其实没有那么高，”唐吉珂德说，“可是她数不尽的美德却使她**楚楚动人**！”

Yang: 堂吉珂德说：“她既有那么高的身材，也就有数不清的才德来配合衬托！”

The source text expression appears in Chapter XXXI *De los sabrosos razonamientos que pasaron entre don Quijote y Sancho Panza, su escudero, con otros sucesos*. A great part of this chapter is devoted to the typical exaltation versus disenchantment dialogue between don Quijote and Sancho. A distinctive divergence on the perspective from which don Quijote and Sancho describe, or rather imagine the qualities that Dulcinea is endowed with is that while Sancho focuses primarily on her physical appearance, which is then inevitably associated with her rough character as a result of constant exposure to heavy work; don Quijote tends to overlook his beloved's outside appearance concentrating instead on the integrity and moral characters that shine through her personality. For example, in describing Dulcinea's height, Sancho exclaims that she is more than a hand taller than him, hinting at her physical strength as somehow contradictory to the fine figure of chivalric damsels admired by his master. In response to Sancho's scoffing at the clumsy look of Dulcinea, whom he knows to be none other than the boorish, peasant girl Aldonsa Lorenzo, don Quijote issues a mild rebuke and swiftly shifts the point of view from the physical to the spiritual level "Pues es verdad que no acompaña esa grandeza y la adorna con mil millones y gracias del alma...".

Yang and Liu have adopted very different treatment to render the tone and manner of Quijote's reaction to the mockery implicit in Sancho's portrayal of Dulcinea. Yang seems to be unaware of the disparaging connotations in Sancho's words, for she has rendered it as (don Quijote) "if she is that tall, she must *also* have countless virtues to accompany her tall figure". Noting the addition of the conjunction *also* suggests that in Yang's translation, being tall has been reckoned as an obvious physical advantage and has been put in juxtaposition with female moral integrity. Such translation implies that don Quijote has accepted Sancho's frivolous portrait of Dulcinea and from there he goes on to build his argument on the spiritual superiority of his idealized lady. Yang's translation of the source text expression thus inevitably weakens the inner tension between the two parties, which I would suggest is the very core of Cervantes's parody of don Quijote's chivalric illusions.

On the other hand, Liu's translation seems to be more prudent in bringing to the front a perceptual shift in don Quijote's response to Sancho's description of Dulcinea, for he has rendered it as "she is not that tall indeed, *but* her countless virtues make her look very beautiful". In his version, Liu has simplified the pompous syntactic structure of the original and retained the formality of don Quijote's speech by introducing a Chinese archaic idiom 楚楚动人, which is often used to describe female delicacy in written Chinese.

ES3 (don Quijote) Por una parte, me acosa y fatiga el deseo de ver a mi señora; por otra, *mi incita y llama* la prometida fe y la gloria que he de alcanzar en esta empresa que ella venga a tener por buena mi *tardanza*, pues verá que todo redunde en aumento de su gloria y fama ... (Chapter XXXI, I)

EN: on the one hand the desire to see my lady pursues and harasses me, on the other my solemn promise and the glory I shall win in this enterprise urge and call me ... to whom I shall make such excuses that she will be led to approve of my delay, for she will see that it entirely tends to increase her glory and fame;

Liu: 一方面，我对我的夫人望眼欲穿；另一方面，我答应的事情和我为此将得到的荣誉又使我 欲罢不能。.... 她会觉得我 姗姗来迟 是对的，因为她发现这增加了她的声誉。

Yang: 我一方面牵肠挂肚要去看看我那位小姐，另一方面又为自己的信义和完成这番事业的光荣振奋得不能罢手.... 她听了我的委婉的解释，就会赞成我，知道我 迟迟不去 是要为她扬名。

The source text expression appears as a follow up on the conversation between the knight and his squire revolving around an imaginary correspondence between don Quijote and Dulcinea during his absence to defeat the giant in the remote territories of Princess Micomicona's kingdom. This is a crucial passage in the first part of the novel which reveals the essence or the rationale behind don Quijote's worship of Dulcinea. It is one of the critical speeches by don Quijote where he rationalizes and justifies his chivalric thoughts and deeds, while the knight gets very emotionally involved.

In this passage, Liu has made full use of Chinese archaism to convey the protagonist's high opinion of the contents of his own speech, as well as the flimsiness of don Quijote's chivalric evocation of Dulcinea. To translate the source text short phrase *mi incita y llama*..., Liu uses the Chinese idiom 欲罢不能 (wanting to stop but cannot help), which is a literary expression often seen in the target language to describe one's attachment or obsession with addictive objects, while Yang has deconstructed the idiom into a common short phrase by translating it as 不能罢手 (cannot stop doing something).

Don Quijote's eagerness to reunite with Dulcinea after the regaining of the Princess' kingdom is well captured by Liu's deployment of the Chinese idiom 姗姗来迟 to translate the single noun *tardanza*. This archaic idiom was first recorded in the Chinese history classic Han Shu (汉书) dating back to the first century AD. The repetition of the adverbial morpheme 姗 is originally used to describe the slow and graceful walking of women and now generally applied to one's delay in arrival. It is normally seen in literary texts and when it is used in colloquial language, the speaker's attitude towards one's delay is more appreciative than derogatory.

The pragmatic function of this archaic idiom within the target language context requires subtle interpretation: on the one hand, in his imaginary personal contact with Dulcinea after a long time engaged in various adventures, don Quijote takes the initiative to blame himself, but his attitude towards his long absence from his beloved damsel is mild rather than overcritical, since he firmly believes that in the last analysis, Dulcinea would be glad to see that "...*todo redunde en aumento de su gloria y fama, pues cuanta yo he alcanzado, alcanzo y alcanzaré por las armas en esta vida, toda me viene del favor que ella me da y de yo suyo*".

4.6 Conclusion

In chapter IV, I have conducted a detailed analysis of the two phraseological patterns which seem to dominate Liu's version of *Don Quijote*, i.e. his relatively higher use of Chinese figurative idioms and archaic idioms. In case after case, it becomes clear that in dealing with source text figurative expressions which are hard to decipher for the target readership, Yang has preferred a more literal approach to the original by translating it word by word with explicit footnotes. Liu's approach, on the other hand, has been more creative and target-text-oriented. He uses Chinese figurative expressions extensively, which sometimes may only match up with source text expressions at a pragmatic or functional rather than semantic level.

Liu's use of Chinese archaic idioms has turned out to be more experimental, more assertive than his use of figurative idioms. As will be shown in the following chapter, archaism is a very complex issue in literary translation and its elusive nature allows the translator a great deal of freedom to follow her or his own understanding or impression of the original text. In the samples studied above, there does not seem to be an explicit link between the archaic feature of the source and the target text in any of the three cases. This highly rhetorical feature of Chinese phraseology has been exploited mainly for the purpose of polishing or improving the language style of the target text while being rather lax with regard to the source text expression.

The very different approaches adopted by Yang and Liu have provided us with valuable first-hand material which helps bridge the huge cultural gap between the two contrasting linguistic systems under discussion, namely, early seventeenth-century Castilian and modern Mandarin Chinese. With their different attitudes they may have shown that despite the formal divergence reflected in the preferred lexical realization by each cultural-linguistic entity, the underlying conceptual relationship between a stereotyped contextual situation and its subsequently established figurative connotation can be seen in many languages. And so it may be feasible, in translation terms, to provide a substitution of ST expressions with TT expressions which can be seen as semantic- and pragmatically equivalent within their own textual environments; and this is exactly what Liu has successfully achieved in his retranslation of *Don Quijote* which has, by the way, proved to be quite a bestseller in China in the 1990s.

At this stage of my research, an interesting question worth asking is that given Liu's seemingly stylistic use of Chinese figurative and archaic idioms, to what extent may his work be seen as independent from Yang's translation and from the original text, i.e. the autonomy of his stylistic use of language in translating *Don Quijote*. As observed from the outset, the comparative study of Yang's and Liu's versions of *Don Quijote* has been designed purposely. Despite the diachronic gap existing between the two, Liu's acknowledged familiarity with Yang's work has added an important dimension to the internal relationship between the three texts under investigation. In an effort to delve into the intriguing relationship between the original and its two target language counterparts, Chapter V will offer a novel quantitative line of research which could reveal the stylistic use of Chinese figurative and

archaic idioms from a subtle and discerning perspective made possible by the application of relevant statistical tests widely used in corpus linguistics.

Chapter V Use of figurative/archaic idioms in the two translations

Through an initial analysis of the distribution of four-character expressions in the two Chinese translations of *Don Quijote* (I), we have been able to identify two important patterns in the use of these phrases as a rhetorical device in Yang's and Liu's works. Whereas Yang's work exhibits a major change of the source text features at a morpho-syntactic level, Liu seems to be in favour of a language style that is characterized by his higher use of Chinese figurative idioms and archaic idioms. Starting from the prominent stylistic features thus identified in the two Chinese translations, we will move on to investigate the rationale behind the use of figurative and archaic idioms in Yang's and Liu's translations of *Don Quijote* (I). This will be achieved through a quantitative analysis of the distribution of figurative and archaic idioms in the two Chinese versions in relation to the occurrence of similar linguistic phenomena in the source text. This quantitative analysis is based on a range of statistical techniques, such as correlation test, linear regression test, etc., which despite their wide application in textual statistics or corpus statistics, seems to have been rarely explored in corpus-based translation studies.

5.1 Use of figurative idioms in the two Chinese translations

5.1.1 Yang's and Liu's translation of ST as figurative

To test whether the use of figurative idioms in the two Chinese translations has actually been stimulated by the figurative nature of the source text, the translation of source text excerpts which are figurative is first highlighted, which amounts to some one hundred samples. The phraseological patterns thus identified in the two Chinese translations will be compared with relevant samples coming from a larger set of comparative corpus data: the rest of the FCEX database, which has been screened out as non-figurative in the first instance and now randomized into a number of 100-sample sets for comparative purposes.

Due to the limited space provided, we shall introduce three 100-sample sets in this part of our study. The inevitable imbalanced ratio of source text samples as figurative versus non-figurative, i.e. 100 (for figurative): 3 x 100 (for non-figurative), is largely due to the fact that the proportion of figurative language in the source text is much smaller than that of its non-figurative counterpart. However, such contrasting study of the phraseological patterns in the two sets of target texts against the backdrop of the source text as shifting from figurative to non-figurative will serve as a general indicator of the different phrasing strategy devised in Yang's and Liu's work, respectively.

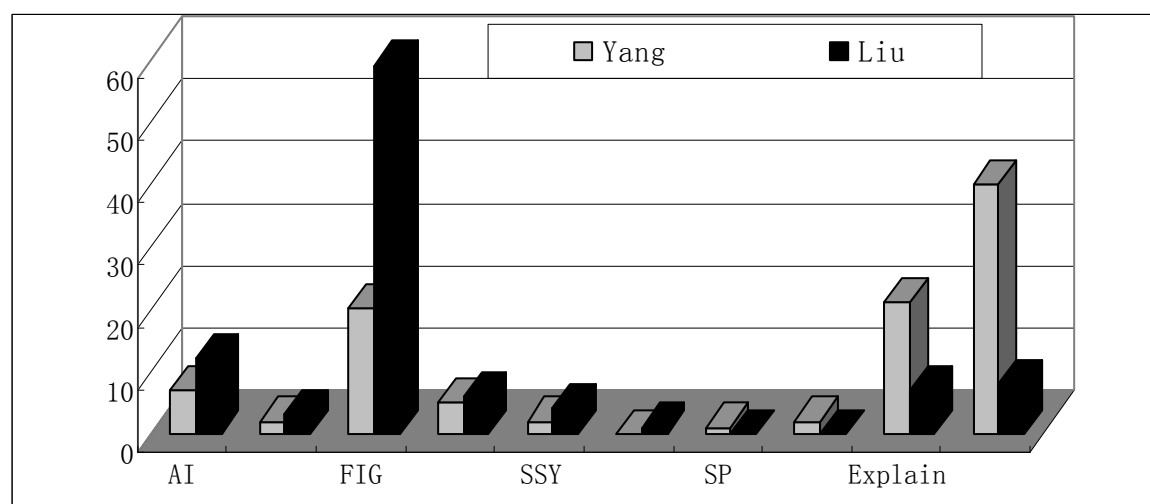
Before proceeding to establish whether there are any similarities or dissimilarities between the two sets of corpus data, we shall have a close look at the phraseological patterns of the two target texts that are translations of the source text figurative speech (see Table VIII). In view of the relatively low occurrence of morphologically modified FCEXs, such as SSY, SB, SP and MP, in the two translations, we shall group them together under a more generalized category, *collocation*, which refers to Chinese

FCEXs with distinctive morphological properties in the current study.

Table VIII Translation of figurative speech in *Don Quijote* by Yang and Liu

NO.	ES	FCEX CATEGORIES		YANG	LIU
1	Figurative	AI		7	12
2	Figurative	I-AI		2	3
3	Figurative	FIG		20	59
4	Figurative	I-FIG		5	6
5	Figurative	Collocation	SSY	2	4
6	Figurative		SB	0	1
7	Figurative		SP	1	0
8.	Figurative		MP	2	0
9	Figurative	Plain Explanation/ Explicitation		21	7
10	Figurative	Literal Translation		40	8
Total	Figurative	FCEX		100	100

Diagram VII Translation of figurative speech in *Don Quijote* by Yang and Liu



To visualize the patterns shown in Table VIII, the figures thus summarized have been converted into a diagram. There are a number of important patterns revealed in the graph plotted above:

(1) Literal translation refers to the process where the translator shows very little intent to modify the phraseological patterns of the source text according to the linguistic habits of the target language, i.e. through adopting a word by word translating strategy. At this level, the statistical difference between Yang's and Liu's work has proved to be indeed very significant: whereas Liu has translated the Spanish text excerpts literally only in eight cases, the relevant figure attributed to Yang is forty, making up nearly half of her translation of the source text. Such difference may be given a fuller account by drawing on the figures shown in Table VIII, which reveals more details of the different phrasing strategies favoured by the two Chinese translators of *Don Quijote*.

(2) Plain explanation or explication should be construed here as the process by which for a number of reasons, among others the absence of similar source language phraseological patterns in the target language, the Chinese translator has opted to paraphrase the source text by providing plain explanations in the target text. It may be seen as a conscious decision made by the translator to encode the source language features within the target language system, which points in turn to a transitional level somewhere in between an overt literal approach and a target-language-oriented normalizing approach to the source text. At this level, the use of plain explanation by Yang again exceeds that of Liu, for the comparative data extracted from Yang's work is twenty-one, as many as three times the figure reported in Liu's work.

(3) The FCEX categories located in the middle of the scale, i.e. SSY, SB, SP and MP, signal the use of phraseological devices as typical of the target language. The difference between Yang and Liu in normalizing the ST at such a morpho-syntactic level has proved to be rather trivial, since the reported data for each text, as shown on the graph, is rather close to each other.

(4) Regarding the manipulation of language register as embodied in the use of archaic idioms or their morphological variants, the figures demonstrate that while in fifteen cases Liu resorts to phraseological devices that may improve the formality of the target language, Yang seems to be rather more conservative in injecting lexical elements that may well signal a marked register of the target language into her first direct translation of *Don Quijote*: she opted for using this kind of phrasal units only on nine occasions.

(5) Lastly, the corpus-based approach adopted in the current study has proved to be a most efficient methodological framework for the search of target text excerpts that exhibit a figurative use of language. Across the two Chinese sub-corpora of CCDQ, some sixty-five target text extracts have been identified from Liu's translation as containing figurative linguistic devices; whilst the relevant figure attributed to Yang's translation abruptly drops to twenty-five.

Such a conspicuous statistical contrast with regard to the use of figurative linguistic devices by Yang and Liu seems to be the most revealing fact emerging from our examination of the phraseological patterns shown above. It suggests that Liu's use of Chinese figurative idioms is more sensitive to the source text features that contain figurative speech. To put such hypothesis to the test, however, we would need to collect more data from the corpora and further contrast the phraseological patterns in the two Chinese translations of the source texts as figurative versus non-figurative.

5.1.2 Yang's and Liu's translation of ST as non-figurative

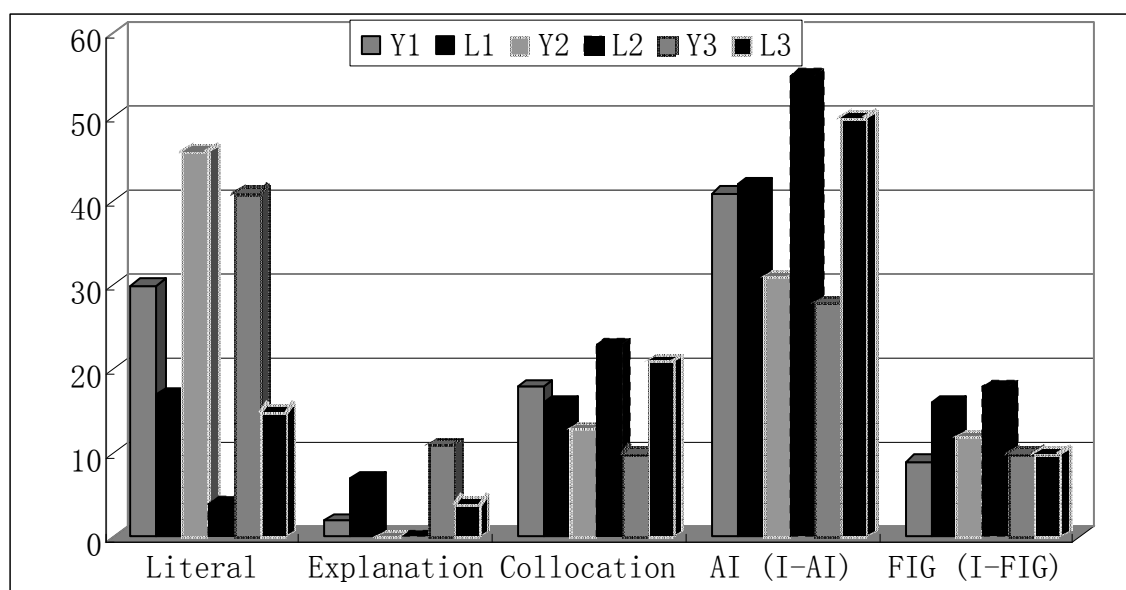
Once the phraseological patterns regarding the transfer of figurative speech in the source text have been established, my next task is to test whether there is any significant change in terms of the use of figurative idioms in the two Chinese translations when the source text shows a semantic shift from figurative speech to non-figurative speech. The selection of samples in this section has largely followed the same procedure of randomization as introduced in the previous section regarding the identification of general phraseological patterns in Yang's and Liu's works.

In this section, again, the term collocation refers to the various types of morpho-syntactic representations of Chinese FCEs: MP, SSY, SS, SB and SP. The adoption of such term is intended to streamline the analysis of possible correlations between the source and target texts, since our attention is primarily focused on the use of figurative and archaic idioms in the two modern Chinese translations of *Don Quijote*.

Table IX Three 100-sample set for comparative purposes (Yang/Liu)

NO.	ES	FCEX CATEGORIES	Y1	L1	Y2	L2	Y3	L3
			First 100		Second 100		Third 100	
I	Non-figurative	Literal translation	30	17	46	4	41	15
II	Non-figurative	Plain explanation	2	7	0	0	11	4
III	Non-figurative	Collocation	18	16	13	23	10	21
IV	Non-figurative	AI (I-AI)	41	42	31	55	28	50
V	Non-figurative	FIG (I-FIG)	9	16	12	18	10	10
Total	Non-figurative	FCEs	100	100	100	100	100	100

Diagram VIII Distribution of FCEs across three sample sets



Based on the data provided in the three test samples, a histogram has been plotted showing the general distribution of FCEXS across the three test sample sets selected from Yang's and Liu's work as shown in Diagram VIII. As we may see from the diagram, despite slight differences in the developing trends of the three testing sample sets, the randomly selected corpus data already begin to show revealing patterns in the distribution of four-character phrases in each translator's work. Liu's use of four-character expressions in translating non-figurative source text excerpts is predominately archaic, which seems to signal an important feature of Liu's version of the novel; that is to say, the enhanced formal style of the target language. On the right, the picture depicted for Yang's phrasing patterns is a bit more complex: (1) the most important feature of the graph is Yang's recurrent use of literal translation in dealing with non-figurative source text expressions; (2) there are also a number of archaic expressions used by Yang in her first translation of the novel, though the proportion taken by archaic idioms in her work remains at a reduced level when compared with Liu's more recent version of *Don Quijote*.

5.1.3 Pattern comparison

Table X Use of FCEXS by Yang in translating source texts as figurative vs non-figurative

NO.	FCEXS	ST AS FIGURATIVE	1 ST SAMPLE SET	2 ND SAMPLE SET	3 RD SAMPLE SET
1	Literal translation	40	30	46	41
2	Plain explanation	21	2	0	11
3	Collocation	5	18	13	10
4	Archaic idioms	9	41	31	28
5	Figurative idioms	25	9	12	10
7	Total	100	100	100	100

Table XI Use of FCEXS by Liu in translating source texts as figurative vs non-figurative

NO.	FCEXS	ST AS FIGURATIVE	1 ST SAMPLE SET	2 ND SAMPLE SET	3 RD SAMPLE SET
1	Literal translation	8	17	4	15
2	Plain explanation	7	7	0	4
3	Collocation	5	16	23	21
4	Archaic idioms	15	42	55	50
5	Figurative idioms	65	16	18	10
	Total	100	100	100	100

Our next task is to test whether the use of figurative expressions in the two Chinese translations have been the result of a shift of the source text feature from non-figurative to figurative.

To this end, the frequencies of main four-character expressions are incorporated into a cross-tabulation which displays the contrastive patterns of the use of four-character expressions in the Chinese translation in relation to the source text as figurative versus non-figurative (see Table X & XI).

Yang's and Liu's use of different FCEX categories shown in the two cross-tabulations have been converted into Diagram IX & Diagram X, which allow us to have a more visualized access to the important patterns hidden behind the figures. As can be seen in Diagram X, the general developing trends of the three gradually-faded bars, which represent the three comparing sample sets, are rather similar to each other, forming a bell-shaped curve. The black bar represents the distribution of four-character expressions in Liu's translation of the source text which contains a figurative use of language. In contrast with the general developing trend of the three samples, the development of the black bar stays at a low level from the first category through to the fourth category, where it sees an abrupt increase and arrives at a peak at the last category.

Diagram IX Contrastive patterns of FCEXs in Yang's translation

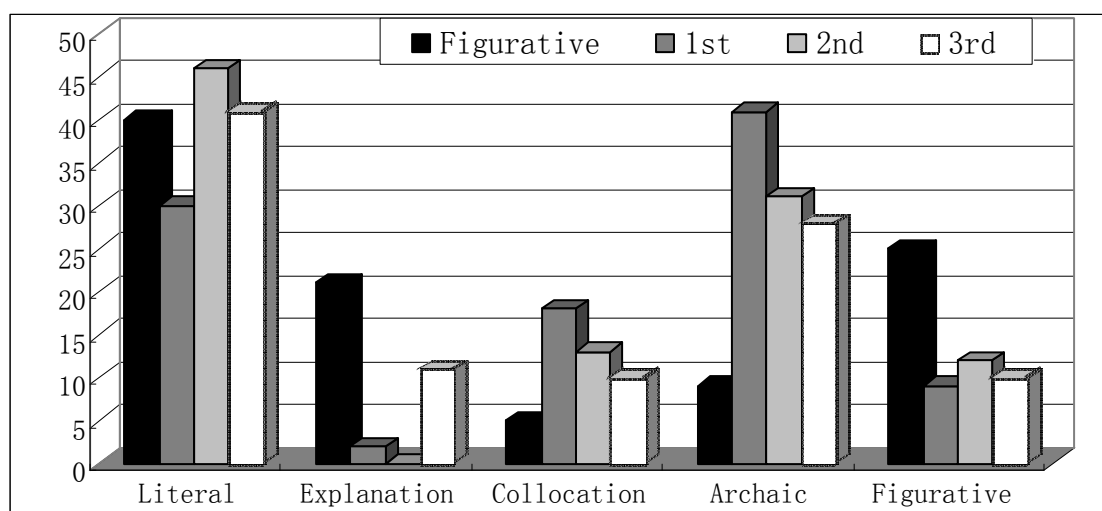
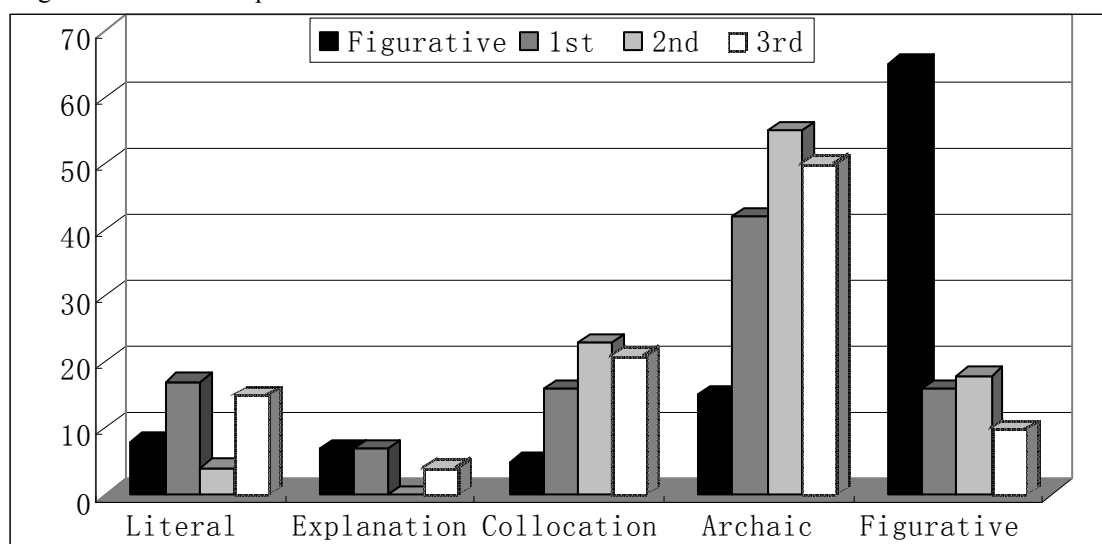


Diagram X Contrastive patterns of FCEXs in Liu's translation



By comparing corpus data this way, important patterns regarding the influence from the source text on the use of four-character expressions in the Chinese target texts become apparent. In the first place, across the four groups of corpus data regardless of their corresponding source text feature, the use of literal translation and plain explanation remains at a low level in Liu's work and does not see any significant change, which may therefore be seen as a minor, though consistent feature of the language used in Liu's translation.

Secondly, the category of collocation in Liu's work appears to have been slightly affected by a change of the source text feature from figurative to non-figurative. The occurrence of collocations found in Liu's work decreases more than twice when the source text shifts from non-figurative to figurative. The use of collocations in Liu's translation appears to be mainly applicable to the translation of source texts as non-figurative rather than being figurative. Such translational phenomenon found in Liu's work indicates that collocations as an important phraseological device used in Liu's work are functioning at a morpho-syntactic rather than semantic level in assimilating the source text elements into the linguistic traits of the target language.

Another phraseological category which turns out to be a major feature of Liu's translation of non-figurative expressions in the source text is archaic idioms. As may be seen from Diagram X, the use of archaic idioms has proved to be a consistent and most prominent feature across the three sample sets when the source text stands as non-figurative. A maximum importance is attributed to archaic idioms in forming the underlying phraseological patterns in Liu's translation of non-figurative source text expressions. However, when the source text feature shifts from non-figurative to figurative, the frequency of occurrence of archaic idioms in Liu's work drops sharply to a level similar to that of the previous three categories, i.e. literary translation, plain explanation and collocation.

Finally, another significant difference between the two sets of data as divided by source text features is the remarkably higher use of figurative idioms in Liu's translation of source texts containing figurative speech. The prevalence of archaic idioms in Liu's work is gradually taken over by figurative idioms as the source text feature shifts from non-figurative to figurative. The alternate use of archaic and figurative idioms in Liu's work as responsive to the presence of figurative speech in the source text seems to suggest that while archaic idioms appears to be a constant feature of Liu's translation, the use of figurative idioms by the Chinese translator is more sensitive to the source language feature as figurative vis-à-vis non-figurative.

Similar conclusions may be drawn for Yang's translation (see Diagram IX). That is, the source text features as figurative versus non-figurative influence the translator's choice of figurative devices in the target language, though the extent of this influence from the original may vary quite visibly from one translator to another. For example, when the source text shifts from non-figurative to figurative, Yang's use of archaic idioms sees a sharp drop, while her use of figurative idioms only increases modestly, with the rest of the source text figurative language being translated by way of plain

expressions. Meanwhile, when the source text changes from non-figurative to figurative, Liu's use of figurative idioms becomes overwhelmingly predominant.

This is an important aspect in the process of establishing Yang's and Liu's distinctive stylistic profiles. On the one hand, the use of figurative idioms appears to be the dominant phrasing strategy in Liu's translation of Castilian figurative expressions. Yang's translation, on the other hand, combines a modest use of Chinese figurative idioms and a profusion of plain explanations. As may be expected, the enhanced use of figures of speech by Liu, when compared to his predecessor, may well give a general impression that the source text contextual settings are more vividly depicted in the more recent version of *Don Quijote*.

In this section, the stylistic difference between Yang and Liu in terms of the use of figurative idioms has been brought to light by comparing the two target texts after contrasting source text circumstances, i.e. figurative versus non-figurative. In the process of comparing parallel texts, the selection of source text samples has been based on a randomized search across the first part of the novel whereby, due to the relatively larger proportion of non-figurative expression in relation to figurative expressions, the ratio of the two groups of sample sets as figurative against non-figurative has been one to three.

This particular type of contrastive study has yielded quite revealing findings: (1) when compared with his predecessor, Liu has greatly increased the use of Chinese figurative idioms in translating source text expressions which are figurative; (2) when the source text is non-figurative, Liu's work has also proved to be more figurative than Yang's work at a general level. Such findings in turn provide important and complementary insights into the general phraseological patterns as initially found in the two modern Chinese translations (see Chapter III).

5.2 Use of archaic idioms in Yang's and Liu's translation of *Don Quijote*

In order to provide a quantitative account of the stylistic differences between the two modern Chinese translations in terms of the use of archaisms, a number of statistical techniques generally applicable in corpus linguistics have been used. Such an approach to textual studies as exemplified by the foregrounding of quantifiable empirical data is characteristic of an emerging field of research, i.e. corpus stylistics. The issue under discussion is archaism as described further on, which may take on very different linguistic representations across languages, especially for broadly unrelated languages, such as Spanish and Chinese. Instead of making an item-by-item, category-by-category contrastive study, I shall approach the topic from a functional and quantitative perspective; that is, I plan to examine the stylistic effects brought about by the use of archaisms in the source and target language systems, separately.

The quantitative methodology applied in this section of the thesis differs quite considerably from the one used in the study of figurative idioms in Yang's and Liu's translation of the source text.

The two fundamental characteristics pertaining to the method used in the previous section are (1) randomization of sample sets and (2) seeking and matching of figurative equivalents in the source and target texts. The rationale behind this traditional comparative approach to the source and target text pairs in translation studies is firmly based on the assumption that the linguistic phenomenon or concept under investigation is omnipresent in both the source and target language systems; and that despite its potentially diverse formal representations in different languages, the nature of the source language, especially figurative language, is both transmittable and traceable when rendered into the target language. For instance, when a Castilian proverb or idiom is translated into modern Chinese, it may well take on the form of a Chinese metaphor, metonymy or hyperbole, or vice versa.

But such observation of translated texts could well be seriously challenged when it comes to the study of translational phenomena that are hard to pin down to a particular uni-dimensional linguistic level, thus requiring a more pragmatic and circumspect attitude rather than an equivalent-seeking approach to the subject matter. Therefore, I propose to single out for special consideration a particular figure of speech, archaism, which is highly dynamic and language-specific, within the specific historical and linguistic background in which it is used. That is, in order to have a proper understanding of the use of archaism in the original text, I would draw on previous studies which tried to define archaisms in *Don Quijote* from the perspective of Cervantes's contemporary readership. Similarly, in the search of archaisms in the target texts, I try to identify idiomatic expressions perceived as archaisms by contemporary readers of Mandarin Chinese.

Archaism is a central theme in the rendition of *Don Quijote*, a contrastive study of the use of archaism in Cervantes's original text and in its two Chinese versions will allow us to have a clear vision of the way the language in *Don Quijote* is handled in the process of being translated into a language as different, both diachronically and diatopically, as modern Mandarin Chinese is from early-seventeenth century Castilian.

5.2.1 Quantification of archaisms in don Quijote's speeches (Part I)

The figure of speech known as archaism may be approached as a linguistic feature reflecting the usage of an earlier period, which may include old-fashioned morpho-syntactic rules or lexical items. Such a rhetorical intent must be based on previous works if it is to achieve a desired effect, and is often the offshoot of a purist attitude towards language, stemming from the belief that language and life in the past were plainer, more democratic and more natural. *Don Quijote* is a seventeenth-century Castilian masterpiece, which depicts the protagonist's hilarious deeds and thoughts in an attempt to revive the value system of medieval chivalry and knighthood.

Mancing (1982: 15) points out that "the most prominent stylistic feature of don Quijote's rhetoric of chivalry is its archaism". In an effort to appreciate the stylistic use of archaic idioms in the Chinese translations, we should first have a fuller understanding of the nature, extent and importance of

archaism in *Don Quijote*. In a previous study Mancing classified the different forms of archaisms in don Quijote's speeches into two major categories, essential and supplementary, functioning at morphological, grammatical, semantic and lexical levels. I should hasten to add that the validation of the phraseological status of source lexical items is never easy, least of all from a diachronic point of view, especially when the source text is so remote. Mancing's list of archaisms may not be complete by any means, but their validity as an old fashioned use of language at the time is supported by a vast range of documents (Mancing, 1982: 19-20).

Table XII Forms of archaism used in don Quijote's speeches (part I)

TYPE	MAIN ARCHAISM		SUPPLEMENTARY ARCHAISMS	
	Archaic form	Modern equivalent	Archaic form	Modern equivalent
Morphology	facér; fermoso ferida	Hacer; hermoso; herida	non; vos; e	no; os; y
Grammar	mostredes; hayades	mostréis; hayáis	levantarse han; prometérsele ha; hobiese; plégaos; pluguiera; rescibió; yoguieren	se levantarán; se lo prometerá; hubiese; plázcaos; placiera; recibió; yacieren
			la vuestra; este su	vuestra; su
Semantics	cautivo	miserable / desdichado		
Lexis	n/a	n/a	See Table XV	

Table XIII Selection of archaic vocabulary in *Don Quijote* (see Mancing, 1982: 19)

ARCHAIC FORMS	MODERN EQUIVALENT	ARCHAIC FORMS	MODERN EQUIVALENT
afincamiento	apremio (constraint)	ál	otra cosa (another thing)
asaz	bastante (enough)	atender	esperar (to await)
ca	porque (because)	catar	mirar (to look at)
desaguisado	agravio (wrong)	fenestra	ventana (window)
guisa	manera (manner)	Infante	niño (young child)
lueñe	lejano (distant)	maguer	aunque (although)
membrarse	recordar (to remember)	prez	premio (reward)
pro	provecho (advantage)	tabla	mesa (table)
talante	voluntad (will)	vegada	vez (time)

Three major types of archaism highlighted are (1) the substitution of *f* for *h* in words like *facér* (hacer), *fermoso* (hermoso), *ferida* (herida); (2) the use of second-person-plural verb forms retaining the *-d* – in words like *mostredes* (mostréis), *hayades* (hayáis), etc. (3) the semantic alteration of *cautivo*

in the sense of *miserable* or *desdichado*. However, for people unaware of the archaic use of *cautivo* in romances of chivalry, it is usually mistranslated as *captive* instead of *miserable* or *unfortunate*.

Other forms or words that have been verified by Mancing as supplementary archaisms are (4) the split future in expressions like *levantarse han* (se levantarán), *prometérsele ha* (se lo prometerá), *hobiese* (hubiese), *plégaos* (plázcaos), *pluguiera* (placiera), *rescibió* (recibió), *yoguieren* (yacieren), etc. (5) the use of the definite article and the demonstrative adjective with an unstressed possessive adjective, e.g. *la vuestra* (vuestra) *este su* (su), etc. (6) three obviously archaic morphological forms from old Spanish, i.e. *non* (no), *vos* (os) and *e* (y). (7) Lastly, there are a number of words which are deemed to carry a modest degree of archaism (see Table XIII).

To facilitate the study of verbal archaisms used by don Quijote, Mancing has proposed to divide the first volume of the novel, which includes 52 chapters, into ten subdivisions based on the main topic of the texts. Such division helps outline the thematic diversity that may give rise to the rate of variation in the occurrence of archaisms in the protagonist's speech. Such a structural modification will serve as the mould on which to proceed with the quantitative analysis of the parallel texts.

Table XIV Ten subdivisions of *Don Quijote* Part I

SUBDIVISION	CHAPTERS	THEME
1.	Chapter I	Alonso Quijano el Bueno
2.	Chapter II-V	The first sally
3.	Chapter VI	The scrutiny of Don Quijote's library
4.	Chapter VII-X	The second sally
5.	Chapter XI-XIV	A pastoral interlude
6.	Chapter XV-XXII	Adventures
7.	Chapter XXIII-XXVIII	In the Sierra Morena
8.	Chapter XXIX-XXXI	Princess Micomicona
9.	Chapter XXXII-XLVI	At the inn
10.	Chapter XLVII-LII	The return home

After a careful identification and classification of the various forms of archaism used in don Quijote's speech throughout the ten subdivisions, Mancing has provided a frequency analysis which traces as precisely as possible the distribution of archaism in the protagonist's speeches (see Table XV). The term *speech* used in Mancing's textual analysis refers to words or phrases quoted directly by the narrator. The content and length of a speech may vary greatly from a simple question like Sancho's "¿Qué gigantes?" to a long monologue like don Quijote's lecture on the Golden Age (Mancing, 1982: 27). An archaic speech is defined as a speech containing any form of archaism as specified in Table XII and XIII. Table XV shows a breakdown of the frequency of don Quijote's archaic speeches throughout the first part of the novel.

Table XV Frequencies of don Quijote's archaic speeches (Part I) (see Mancing, 1982: 31)

SUBDIVISION	NUMBERS OF CHAPS. IN WHICH D.Q. SPEAKS	NUMBER OF SPEECHES	NUMBER OF ARCHAIC SPEECHES
1	1	1	0
2	4	32	11
3	0	0	0
4	4	42	9
5	4	22	2
6	8	145	12
7	4	54	2
8	3	50	5
9	8	50	5
10	5	31	6
Total	41	427	52

5.2.2 Quantification of don Quijote's archaic speeches in the two Chinese translations

5.2.2.1 Strength of correlation

Let us proceed to an examination as to whether the patterns of don Quijote's archaic speeches in the original as shown above bears any resemblance of the occurrence of archaisms in the two Chinese translations.

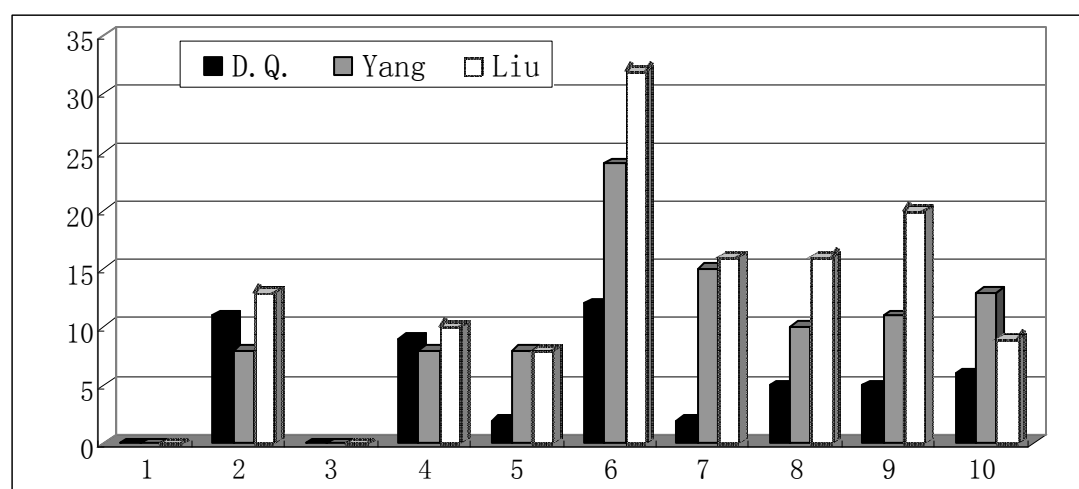
Table XVI Frequencies of don Quijote's archaic speeches in translations

SUBDIVISION	NUMBERS OF CHAPS. IN WHICH D.Q. SPEAKS	NUMBER OF SPEECHES	NUMBER OF ARCHAIC SPEECHES (YANG)	NUMBER OF ARCHAIC SPEECHES (LIU)
1	1	1	0	0
2	4	32	8	13
3	0	0	0	0
4	4	42	8	10
5	4	22	8	8
6	8	145	24	32
7	4	54	15	16
8	3	50	10	16
9	8	50	11	20
10	5	31	13	9
Total	41	427	124	97

To visualize the contrastive patterns hidden behind the abstract data, the numerical information provided in the cross-tabulations is then used to plot a histogram, which gives us some idea of the major similarities or dissimilarities between the source and target texts with regards to archaic speeches uttered by the protagonist. From the graph we can see that up to the fourth subdivision of the novel, the use of archaisms in the target and source text is rather similar to one another; major discrepancies arise from the sixth subdivision onwards, though the general developing trend of three

coloured lines is somehow preserved until the sixth subdivision.

Diagram XI Contrastive patterns of don Quijote's archaic speeches



There are some interesting questions worth asking here as to how the patterns reflected on the graph may be explained in terms of the correlation among the three sets of corpus texts under discussion. For example, despite the huge linguistic differences, to what extent the use of archaism in the original is retained when the source text is rendered into modern Chinese; or, whether the influence coming from the original on Yang's and Liu's use of archaisms differ significantly or not in their respective translations. We may also want to know whether the influence of Yang's first direct translation of the novel from Spanish into Chinese is still perceptible after some twenty years in Liu's recent retranslation. To answer these questions, Pearson's correlation test is first run on the corpus data thus extracted, which yields the important statistical information shown in Table XVII.

Table XVII Pearson's correlation test for Don Quijote's archaic speeches in CCDQ

	Variable Y	Variable X	Variable Y	Variable X	Variable Y	Variable X
Text	Liu	D.Q.	Yang	D.Q.	Liu	Yang
<u>Mean</u>	12.4	5.2	9.7	5.2	12.4	9.7
<u>Variance</u>	81.24	16.96	44.21	16.96	81.24	44.21
<u>Standard Error</u>	9.01	4.12	6.649	4.118	9.01	6.649
<u>Covariance</u>	25.420		17.260		55.02	
<u>Correlation</u>	0.685		0.630		0.918	
<u>Determination</u>	0.469		0.397		0.843	
2-sided p-value	0.02889 (< 0.05)		0.0507 (> 0.05)		0.000178 (< 0.05)	
Df.	8		8		8	
Observations	10		10		10	

N.B. Variable X stands for independent variables; Variable Y refers to dependent variables.

Pearson's correlation test is a basic statistical procedure widely used in quantitative linguistics. It tests the strength of association between variables as indicated by the predefined threshold probability value of 0.05. For example, if the correlation test yields a probability value higher than 0.05, this means that the underlying hypothesis of the test, which assumes null correlation between the variables, is held true; conversely, if the probability value turns out to be lower than 0.05, the null hypothesis is accordingly denied, suggesting a statistically proved association between the variables.

From the statistical results shown in Table XVII, it is easy to see that first, in the row of mean value, while the average prevalence of don Quijote's archaic speeches in the original is as low as 5.2, this figure is almost twice the rate in Yang's translation of the novel, suggesting that don Quijote's archaic speeches were greatly appreciated when the original was first translated into Chinese in the late 1970s. Such a trend of an increased use of archaisms in the target text is further stressed in Liu's recent rendition of the novel, for the average use of archaism in his version of *Don Quijote* is as high as 12.4.

Secondly, in the row of co-variance, while the variation between the original and Yang's translation has shown the lowest value at 17.26, the chance of co-variation between the two translations has shown the highest probability, 55.02. Similar interpretations may be given to the findings displayed in the correlation line: that is to say, where the correlation between the original and Yang's translation has proved to be the lowest, the correlation between the two translations draws close to 1, which means that the rate of use of archaic idioms in the two modern Chinese versions of *Don Quijote* is very similar, if not identical.

In the fourth line counting from the bottom of the table, the value of coefficient determination shows the proportion of variability in the dataset that is accountable by the correlation model. The closer the coefficient of determination to the maximal limit 1.0, the stronger the correlation between the variables as described by the statistical model. For example, the relatively low value of determination coefficient attributed to the pair of the original and Yang's translation suggests that there is less of an association between these two variables when compared with the other two pairs of corpus texts. On the other hand, the high value of determination coefficient ascribed to the two Chinese translations is indicative of a more similar use of archaisms in the protagonist's speeches in the two target texts.

The most important statistical result shown in Table XVII, in my view, is the two-side probability value, which forms the basis of a conclusive argument on the similarities and dissimilarities between the three pairs of corpus texts under investigation. As mentioned earlier on, the commonly accepted threshold probability value in Pearson's correlation test is five per cent. Any value below this predefined figure is interpreted as a statistically verified association between the studied variables; and any value above five per cent suggests that the chance to observe a co-variance in the variables is essentially accidental in statistical terms.

The probability value has thus revealed that (1) the correlation between the two Chinese

translations is undoubtedly high, since the chance to reject the null hypothesis, which assumes the null association between the two sets of texts, is as low as 0.000178, far below 0.005; (2) when construed within the statistical context as delineated by the 0.05 threshold value, there is a proven similarity between the original and Liu's translation in terms of the prevalence of Don Quijote's archaic speeches; however such similarity is denied for the second text pair containing the original and Yang's first translation of the Spanish novel.

Such conclusion seems to indicate that in Liu's recent rendition of *Don Quijote* (Part I), the Chinese translator has greatly drawn on the phraseological strategy as devised by his predecessor (see finding (1) above) when dealing with the Castilian archaism in the original text; however, Liu's work has also made an important contribution to a refined interpretation of the original, since as the statistical test shows, the association between the original and Liu's Chinese version has been somehow corroborated, while Yang's translation has shown to be more distant from the original with a correlation probability value lower than the critical 0.005.

Table XVIII Summary of Pearson's correlation test

	VARIABLE Y	VARIABLE X	VARIABLE Y	VARIABLE X	VARIABLE Y	VARIABLE X
Text	Liu	D.Q.	Yang	D.Q.	Liu	Yang
Observations	10	10	10	10	10	10
Critical value for Pearson's test (two tailed at 5% level)	0.632					
Correlation value (number of archaic speeches)	0.685		0.630		0.918	
Significance (Y/N) (two tailed at 5% level)	Yes		No		Yes	

Table XVIII provides a further interpretation and summary of the Pearson's correlation test with the raw frequencies of don Quijote's archaic speeches identified in the three sets of corpus texts. The critical two-tailed correlation value at five per cent is set at 0.632. As a result, any calculated correlation value higher than this standard signifies a strong association between the variables, with 1.0 as the maximum likelihood value, where the two variables prove to be identical as measured on the highlighted textual features, i.e. the protagonist's archaic speeches in the case of the current study. Similarly, any correlation value below this predefined critical value denies a stable association between the variables in a statistical sense.

Table XVIII shows clearly that (1) among the three sets of corpus texts, Liu's work displays a high similarity with Yang's work in terms of the prevalence of the protagonist's archaic speeches in the two modern Chinese versions of *Don Quijote*, for the correlation value estimated for the two texts is 0.918, much higher than the critical value 0.632; (2) with a marginal surplus, the strength of association between Liu's work and the original is confirmed on a solid empirical grounding; (3) in a similar

fashion, with a minor deficit, the potential analogy between the original and Yang's first translation is rejected by the statistical procedure.

A full consideration on the three findings summarized above bearing in mind my personal impression from a face-to-face interview with Liu in April last year⁴⁴ brings to light the complex and dynamic nature of Liu's recent translation of *Don Quijote* as regards the use of archaisms. It seems that the phraseological strategy developed in Yang's first translation of the Castilian novel has been carefully tailored, with a marked increase in the use of Chinese rhetorical devices, i.e. archaic idioms, in Liu's more recent version.

More importantly, the legacy of Yang's phrasing patterns in Liu's work has been reinforced by her successor through a closer inspection into the textual features of the original. Whether this improvement has been successful, even on a rather limited scale, and would necessitate a full arithmetical enquiry to make it clearly visible. All the same, such findings as I come up with undoubtedly speak for the significance and usefulness, methodologically speaking, of applying relevant statistical methods in text-based translation studies. The subtle, yet qualitatively significant differences between the two Chinese translations would have been easily overlooked had the raw dataset not been subjected to a detailed computer-aided quantitative analysis.

5.2.2.2 Correlation statistical modelling

Given the corroborated correlation obtaining among the three aligned corpus texts, we may want to investigate in more detail how we can make a most reliable prediction on the use of archaisms in the two Chinese translations of *Don Quijote*. This section constitutes a further exploration of the relationship among the three sets of texts, by deploying a more powerful statistical procedure in textual analysis known as linear regression. Regression and correlation tests differ from each other in that while the latter does not assume any causal relationship between the variables under investigation, the former is designed to measure the scale of dependence of observed variables upon the variance in controlled variables.

Before proceeding to test any hypothesis regarding the causal effects or influence of one text upon another, it is necessary to specify the circumstantial contexts in which the relationship between the original and its two Chinese translations is established. Firstly, given the inherent dependence of translation on the source text, the two Chinese translations should be seen as dependent variables in relation to the original; then, on the basis of the data collected from a previous one-to-one interview with Liu, a special tie of the latter translation on the former is hinted at by the Chinese translator when he acknowledged his familiarity and endorsement of Yang's work. In this sense, we may speculate a potential dependence of Liu's work on Yang's work, a conjecture which has actually been somewhat

⁴⁴ In a personal interview with Liu in April 2007, the Chinese translator openly expressed his familiarity and general approval of Yang's treatment of many difficult points in the Castilian text, among others, archaisms.

substantiated by the evidence gained from the results of the correlation test.

The task of testing any cause/effect relationship among the texts will be achieved through a twofold procedure of the statistical modelling of the retrieved corpus data. Firstly, I should look to find out to what extent Yang's use of archaic idioms in don Quijote's speeches depends upon the occurrence of relevant data in the original; secondly, in a relatively more complicated manner, we will build up three linear regression models to see how well may one predict the use of archaic speech in Liu's work, based on the knowledge of the textual features of the original and/or Yang's first translation of the novel.

Linear regression, which may be further divided into simple or multiple regression, is widely used in corpus linguistics to model the influence of one or various independent variables on a single dependent variable. In the current study, to measure the influence of archaism from the original on Yang's use of archaic idioms in the protagonist's speeches, we shall use the simple regression model, for there is only one explanatory variable in this text pair; and to study the rationale behind Liu's use of archaic idioms as a possible result of the textual features of the original and/or Yang's translation, we shall deploy both the simple or multiple regression tests depending on the number of the independent variables featured in the hypothesis.

In linear regression, we can work out a value for R^2 , as the coefficient of simple or multiple determination. It represents the proportion of variance in the dependent variable that may be explained by the independent variables, ranging from zero to one. To examine the influence from the use of archaisms in the original and in Yang's version of the novel on the use of archaisms in Liu's work, three different models of linear regression between the dependent and independent variables have been constructed.

Table XIX Summary of linear regression models

	DV	IDV	DV	IDV(s)	DV	IDV(s)	DV	IDV(s)
Model code	n/a		I		II		III	
Sub-corpus	Yang	D.Q.	Liu	D.Q.	Liu	Yang	Liu	D.Q. + Yang
Mean	9.7	5.2	12.4	5.2	12.4	9.7	12.4	7.45
Observations	10	10	10	10	10	10	10	10
Standard error of estimate	5.7711		7.3434		3.9948		4.0086	
Determination coefficient (R^2)	0.3973		0.4690		0.8429		0.8615	

N.B. DV: dependent variable; IDV: independent variables

The first model may be construed as Liu's use of archaisms as a result of the use of archaisms in the original, which has a R^2 of 0.4690. The second model may be understood as Liu's use of archaism as a result of the use of archaisms in Yang's translation, which yields a much higher R^2 of 0.8429. The third model is built upon the hypothesis of Liu's use of archaism as a result of both the original text and Yang's translation, which has proved to be of the highest determination coefficient, 0.8615. Lastly, the predictive power of the original over Yang's first translation of the novel has reported the least R^2 , as low as 0.3973.

Due to the fact that the original R^2 tends to overrate the strength of association between the dependent and independent variables, we usually work in terms of a slightly different calculation, called adjusted determination coefficient or $R^2_{adj.}$, to assess the goodness of fit of a regression equation. This is calculated as follows:

Formula III Adjusted determination coefficient

$$R^2_{adj.} = 1 - [(1 - R^2) (N - 1) / (N - k - 1)]$$

In this equation, N is the number of observations in the data set, i.e. the ten subdivisions of the first part of the novel, and k the number of independent variables or regressors, which is two in our study referring to both the original text of *Don Quijote* and Yang's first translation of the novel. As a result of the adjustment, the sequence of high predictability is changed from Model III, II and I to Model II, III and I.

Table XX Summary of adjusted R^2 in the linear regression models

Dependent Variable	Independent Factor (s)	Model	R^2 (multiple correlation coefficient)	Adjusted R^2 (to degrees of freedom)
Liu	ES + Yang	III	0.8615	0.8220
	Yang	II	0.8429	0.8232
	D.Q.	I	0.4690	0.4026
Yang	D.Q.	n/a	0.3973	0.3220

To have a closer look at the fitness of the models thus proposed, we have performed the ANOVA test for the linear regression modeling of the corpus data. It may be obtained by using the following formula.

Formula IV ANOVA of linear regression

$$F = [R^2 / (1 - R^2)] [(N - k - 1) / k]$$

As explained in the formula for adjusted R^2 , N represents the number of observations, i.e. the ten subdivisions of the first part of the novel, and k refers to the number of independent variables, which amounts to two in the current study. To interpret the F value thus calculated, we would need two important sets of measurement, i.e. the numerator and denominator degrees of freedom. That is because, F is like some other statistics, e.g. t or χ^2 , in that its significance depends on its degrees of freedom; however, different from the latter two, the interpretation of F statistics is more sophisticated in that it requires the incorporation of two types of degrees of freedom, i.e. Df_1 in numerator ($= k$) and Df_2 in denominator ($= N - k - 1$).

The F statistic is a ratio value of the part of the dependent variable that may be explained by the independent variables over the residual of the model. The larger the F value, the more likely the explainable factors are jointly significant in predicting the behavior of the dependent variable. With regards to the study of Liu's translation, we can see in Table XXI that the F statistic computed for Model II is the largest, which means that in predicting Liu's use of archaic idioms in his work, Model II is the most efficient, while Model I is also somehow reliable when compared with the much lower F value attributed to Model III.

However, the interpretation of F statistics is not as straightforward as R^2 in linear regression, where the larger the R^2 , the stronger the dependence of the observed variables on controlled variables. Similar to the critical probability value introduced in Pearson's correlation test, there is a predefined threshold value, also at five per cent, which serves as the criterion of justification. For instance, if the computed F -value is larger than the critical F -value, the p -value drops much lower than five per cent based on which we have greater confidence in rejecting the null hypothesis, i.e. there is no strong association between the independent variable (s) and the dependent variable.

From the figures shown in Table XXI, we can see that in predicting the use of archaic idioms by Liu, the more secure way is to examine the use of archaic idioms by Yang, since the F statistics attributed to the second model built on the association between the dependent variable (Liu) and the independent variable (Yang) is quite distant from the critical F -value at the two per cent level: 42.884 to 5.32; meanwhile, a mixed consideration on the influence coming from both the original and Yang's work also shows a very positive prediction, for the computed F statistics is much larger than the critical F -value as 21.791 to 4.74. As a result, in both cases, the significance level of the model is confirmed.

In spite of the proven strength of correlation between the original and Liu's work (see the summary of Pearson's correlation test), the F statistics computed for this text pair is lower than the critical value, which suggests the unreliability of Model III. Lastly, the influence of the textual features of the original on Yang's use of archaic idioms in her translation has also proved to the least statistically significant, accordingly, the confidence with which we may make a prediction is greatly reduced, for this time the F statistics is much lower than the critical value as 5.2736 to 11.3.

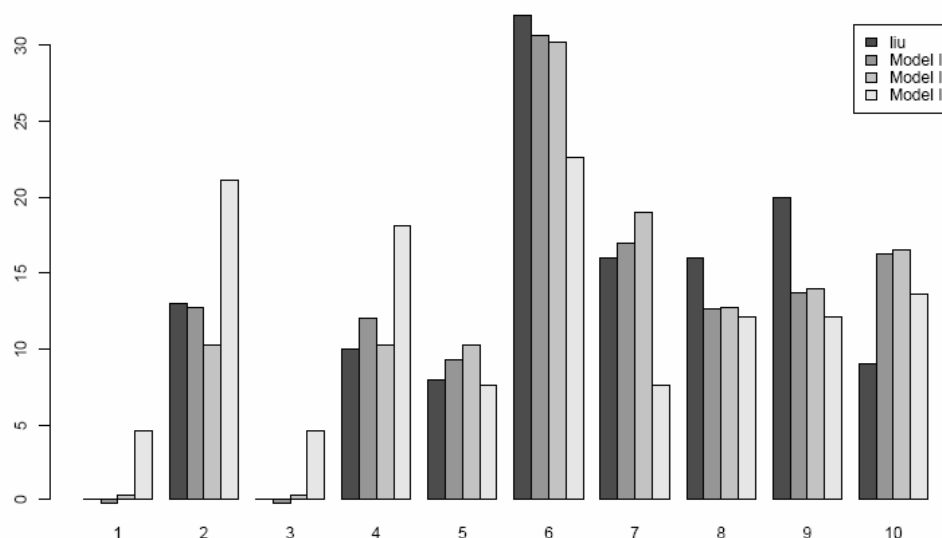
Table XXI ANOVA of the linear regression models

DEPENDENT VARIABLE	INDEPENDENT FACTOR (S)	F STATISTICS
Liu (Model I)	ES + Yang	21.791
	$Df_1 : Df_2$ [Num: De]	2:7
	Critical value of F (two tailed at 2% level)	9.55
	Significance (two tailed at 2% level)	Yes
Liu (Model II)	Yang	42.884
	$Df_1 : Df_2$ [Num: De]	1:8
	Critical value of F (two tailed at 2% level)	11.3
	Significance (two tailed at 2% level)	Yes
Liu (Model III)	<i>Don Quijote</i>	7.063
	$Df_1 : Df_2$ [Num: De]	1:8
	Critical value of F (two tailed at 2% level)	11.3
	Significance (two tailed at 2% level)	No
Yang	<i>Don Quijote</i>	5.2736
	$Df_1 : Df_2$ [Num: De]	1:8
	Critical value of F (two tailed at 2% level)	11.3
	Significance (two tailed at 2% level)	No

To allow us to have a more intuitive view of the fitness of the three models thus built, the observed frequencies of archaic idioms in Liu's translation are then put into comparison with those expected from the hypothetical models on an histogram plotted in the R programming environment (see Diagram XII)⁴⁵. The first set of bars in black represents the observed frequencies of archaic idioms in Liu's work, while the remaining three sets of bars in gradually-faded colors stand for the expected values derived from the statistically modelled linear correlations between the dependent variable, i.e. Liu's translation, and the independent variables, which are both the original and Yang's translation in Model I, Yang's translation in Model II and only the source text in Model III.

⁴⁵ The R statistical programming software may be downloaded at <http://www.r-project.org/>

Diagram XII Comparison of three hypothetical models



In accordance with the findings in Table XXII, which summarizes the results of the ANOVA test of multiple regression of don Quijote's archaic speeches in Liu's work, while the two sets of bars corresponding to the first and second models show greater likeness with the one representing the authentic data, the third set of bars, which embodies Model III, differ to a greater extent from the actual counts of don Quijote's archaic speeches in Liu's work.

A quick comparison of the actual count of the protagonist's archaic speeches and the three sets of hypothesized values leads to the conclusion that on a general scale, Model I and II have proved to be relatively more successful in predicting the prevalence of discursive archaisms in Liu's work, since the predicted frequency that Model I or II has yielded draws quite close to the real breakdown of frequency embodied in the first set of black bars.

However, if we have a closer look at Diagram XII, we may also find out 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, discernible discrepancies between the actual frequency and the hypothesized occurrence of the protagonist's archaic speeches in Liu's work begin to emerge, which is indicative of an idiosyncratic use of language on the part of the Chinese translator when exploring archaism as a rhetorical device in the protagonist's utterance.

To be more specific, while in Subdivision VIII and IX, the rate of don Quijote's archaic speeches seems to overrun significantly its statistically predicted counterpart, in the last subdivision (Chapter XLVII-LII) of the first part of the novel, the frequency of the knight's archaic speeches drops sharply when compared to the standard rate as suggested by the two adjacent bars representing Model I and Model II. This actually brings about the issue of stylistic variation or style shifting within the work

of a single author, which will be the main topic of investigation in the next chapter of the thesis.

5.3 Conclusion

Chapter V provides a quantitative exploration of the stylistic use of figurative and archaic idioms by the two Chinese translators when working on the source text. In studying the use of figurative idioms in the two translations, we have found out that while both Chinese versions of *Don Quijote* seem to be sensitive to the presence or absence of figurative expressions in the source text, Liu has obviously increased his use of figurative idioms in a way that is distinctive from that by his predecessor, and thus showing a more personal style. Through mapping out the distribution of archaic idioms in the three sets of corpus texts, a very interesting phenomenon comes to light; it refers to the experienced variation in Liu's use of archaic idioms in the second half of the novel (Part I). The recognition of stylistic variations in Liu's work would have been impossible without the assistance of statistical tests. The two types of statistical procedures used are correlation and regression tests. Both of the two methods provide a quantitative modelling of the relationship between the corpus texts under investigation and help reveal to what extent the three corpus texts are similar or dissimilar from each other. The statistical test makes it clear that in terms of the use of archaic expressions, while the highest score of strength occurs between the two Chinese translations; Liu's work has proved to be statistically more similar to the original than Yang's version. More importantly, since the scores of correlation attributed to the two text pairs Liu / D.Q. and Yang / D.Q. are located at both ends of the threshold value, we may confirm the existing correlation between Liu's version and the original. But this is not true when comparing the source text and Yang's first direct translation of the novel.

Chapter VI Quantitative Exploration of Stylistic Variation in Liu's Translation (I)

Much of the description of the higher organization of language has remained at the stage of patterns and labels. Little has been done to describe restrictions or to *explain the reasons for the patterns, i.e. to make a proper structural description* Very detailed and careful analysis is required in stylistics, and it is still usually done by hand. We are now in a position to be bold, to *look for testable hypotheses* that a particular description of language can actually provide a set of boxes into which text can be appropriated. We must look for models which help the text to reveal itself to us.

(Sinclair, 2004: 16)

6.1 A context-motivated theory for style shifting in literary translation

My main purpose in this section is to explore the rationale behind stylistic variation, deliberately or otherwise, in literary translations. My approach to the study of style shifting is intended to be interdisciplinary, drawing upon methodologies and techniques from corpus statistics and computational stylometry, and relevant sociolinguistic theories of stylistic variation (Biber & Finegan, 1994). In trying to explain literary translators' idiosyncratic use of language, it sets out to address one of the most difficult issues in this area of research, i.e. the cognitive rationale behind stylistic shifts in literary translation (Riding, 2000: 365-9).

Empirical studies on individuals and their translating styles have generally been criticized for their over-reliance on impressionistic evaluations. The present study aims to show that style variation in literary translation may actually be better explained and consequently further developed through the processing and detailed inspection of quantitative linguistic data extracted from purposely built individual corpora. I believe this practical orientation offers a more appropriate base from which to delve into the cognitive dimension of stylistic variation, i.e. context-motivated or habitual/ unconscious, in literary translation. By establishing a linkage between such linguistic variations and the translator's own or assumed perception of the source text contextual situation, I should help to explain abnormal or unpredictable use of language by a translator.

I should point out at this stage that this section of the thesis differs considerably from the previous chapter in that I intend to move forward from a simple quantification of linguistic data to the search of interesting patterns in the corpus texts. The present chapter focuses primarily on the statistical exploration of co-textual factors that may have possibly triggered stylistic variation or style shifting in the translations under consideration. It thus differs from many current stylistic studies which are invariably concerned with devising and improving techniques that may increase the sharpness or accuracy of authorship attribution tools in textual studies (Hoover, 2003a: 261-86; 2003b: 341-60).

The methodological framework developed here aims to integrate qualitative and quantitative textual analysis to probe into the dual cognitive nature of stylistic variation, i.e. conscious as opposed

to habitual, which has been uncovered in the second half of Liu's translation of *Don Quijote* Part I. That is, instead of pursuing the notion of what constitutes the basic features of an author's literary style, the present paper attempts to establish an empirical line of argument to interpret the variation in stylistic patterns as revealed by corpus statistical techniques.

It is my contention that style shifting in literary translation is a very complex cognitive phenomenon which may sometimes require a quantification of source text co-textual information as it is potentially capable of explaining such an important creative process in literary translation. To mark out the cognitive nature of stylistic variation, a context-motivated theory (CMT) has been formulated and put to the test using primary linguistic data retrieved from the parallel Castilian-Chinese corpus of *Don Quijote*.

The exploratory testing of hypotheses has been significantly aided by the application of and experimentation with statistical techniques, which are widely used in social and behavioural sciences. Their usefulness in corpus-based translation studies has not yet been fully realized. I refer here to two major types of advanced statistical tests, i.e. exploratory and confirmatory modelling techniques, which I am putting into operation to first build a theoretical instrument that should explain style variation in Liu's use of archaism in Subdivision VIII and IX of the novel, and then to validate and test the theoretical model with data collected from the last Subdivision X.

Results obtained so far suggest that in investigating stylistic variation, which is an important representation of the creative nature of literary translation, an integral approach to the subject matter combining the discriminating strength of quantitative statistical analysis with explanatory power of relevant sociolinguistic theories, may well hold the key to a deeper understanding of the cognitive nature of the linguistic phenomenon under discussion.

Style shifting has always been closely associated with the study of context-motivated or proactive speech variation in sociolinguistics (see Labov, 1972; Bell, 2001; Eckert & Rickford 2001), primarily in quantitative terms; while the exploration of stylistic variation in translated texts by empirical methods seems to have been given much less attention. CMT offers an alternative explanation to style shifting and addresses the issue from a perspective that is different from previous views, such as Labov's attention-to-speech model (1972), or Bell's audience design (1984; 1997 & 2001). These past theories basically deal with the interaction between original texts and their reception by the first-language readership; they, however, do not touch upon the intricate relationship between the textual context of the original and its impact on the translator's mind regarding the linguistic organization of the former, and more importantly, the subsequent modification of the source text by the literary translator based on his or her own understanding of the original.

If we consider those previous theories as a social contextualization of textual phenomena, the proposed CMT is firmly based on a co-textualization of target text information within the background

of Translation Studies. The CMT aims to shed some retrospective light on the creative process of literary translation in which translators, when dealing with somehow intangible linguistic issues such as archaism in a totally different language (which may be extremely difficult to pin down from a diachronic point of view), resort to source text contextual cues rather than linguistic cues.

Also, the CMT stands in clear contrast with previous works which would see stylistic variation in literary translation as source text inspired or related (Hypothesis 1); and/or such consequently largely fixed at a linguistic level (Hypothesis 2). But it is my view that holding the first hypothesis to be true does not necessarily mean that Hypothesis 2 may follow. That is, literary translation entails an intellectual activity which can hardly be reduced to a mechanistic transferring or to translating at a word or sentence level; rather, in handling a number of linguistic issues in the source text that do not have a ready answer, or are essentially open to one's own understanding, the translator may well be motivated to develop a different translation strategy to cope with the task, such as taking full cognizance of the contextual features of the source language.

A key set of terms lying at the heart of the proposed CMT is the conceptual pair of dominant versus latent, which has been developed to describe the contextual features that either have a stimulative or alleviative effect on the translator's conceptualization of the source text contextual characteristics, which in turn will play a vital role in the decision process whereby the literary translator considers using one linguistic device as against another, e.g. archaism as against modern language.

For example, in studying stylistic variations in Liu's use of archaism in don Quijote's speech, taking the surrounding audience as a variable of the contextual analysis, the presence of a listening audience in the source text may well trigger off a stimulative urge on the translator's mind to use an archaic idiom, which in turn will enhance the chivalric image of the enchanted knight. On the other hand, the absence of an audience, or the private setting where the communicative act takes place, tends to have the opposite (alleviative) effect on the consciously aware translator, for it is commonsense that in a private setting, it is less likely that the fictional character would take on an ostensibly archaic tone in order to show off his/her knowledge of the particular topic.

In the process of exploring a hypothesis, I first apply Biber's analytical framework for register variation in verbal communications (Biber, 1994: 31-56). As may be expected, Biber's previous study may only be used as guidelines in the current study, given the changed nature of the stylistic variation under discussion. That is to say, while Biber's analytical instrument aims to cover a broad range of contextual factors bearing on register variation in spoken languages, the purpose of the present paper is to identify the principal dimensions of source text situational variables that may help explain style shifting in literary translation, especially with regards to the use of formal linguistic devices such as archaism.

Through computational statistical modeling, I wish to tease out an analytical instrument or a

well-structured theoretical model from the whole spectrum of situational factors included in Biber's previous study. This implies a series of statistical procedures which classify and conflate situational factors that may point to the latent structure of the variables highlighted for special study. This statistically built model may in turn help explain the detected variance in Liu's stylistic use of archaism in the second half of *Don Quijote* (Part I).

6.2 Statistical method applied: CATPCA

The exploratory technique used here is known as Categorical Principal Component Analysis (CATPCA for short). CATPCA is a type of exploratory statistics widely used in social or behaviour sciences to identify the structure of variables that may be used to describe discrete social phenomena. Despite its technical limitations in making causal claims, it has proved to be quite effective in reducing a large number of variables into a handful of statistical dimensions which help reveal the nature and character of the subject matter as quantified by the original variables. CATPCA's mechanism is purported to build a statistical model with limited dimensions, usually two or three, on a wide range of variables. It helps arrange large amounts of original variables in a way that is susceptible to human observation in finding interesting patterns in later qualitative textual analysis.

This is a vital process in the exploration of primary textual data where the reduction of numerous variables into a limited set of statistical dimensions offers a way to better comprehension of the nature of the original dataset; and hence provides important clues to research questions raised around the fundamental structure of textual data. Based on a case by case study on the selected stylistic variations in Liu's version of *Don Quijote*, the present study aims to test the validity and productivity of CATPCA within the context of corpus-based textual analysis, i.e. the establishment of a theoretical model to explain the stylistic variation identified in the work of a single translator.

To sustain the statistical model, the analytical framework of variables that I am using here has been adapted from Biber's famous sociolinguistic study on register variation (Biber, 1994: 40-1). There has been a long debate over the subtle differences in the definition of style and register (e.g. Cheshire & Bell, 2003: 455; Savova, 2005: 898). In the current study, we shall refrain from complicating the issue by treating them as two largely exchangeable working concepts, which broadly refer to the linguistic variation detected in actual textual contexts at a morphological, lexical, phraseological or syntactic level.

But first, since Biber's analytical scheme of situational parameters for register variation has been devised for text/discourse studies generally, I feel it is necessary to adjust the framework to fit the purpose and scope of my current study, which focuses primarily on the use of archaisms as a rhetorical device in literary translations. In Appendix IV, I offer a full version of Biber's analytical framework for register variation, while in Table XXII I provide a modified version of the original framework with the items retained as pertinent to the modelling of source text contextual features in connection with the

use of archaism in the Chinese translation of *Don Quijote*.

Variables retained in the adjusted framework have been intuitively classified into five main categories of situational parameters: (1) communicative characteristics of participants; (2) relationship between addressor and addressee(s); settings of communicative acts; (3) relations of participants to the text; (4) speaker's intent, purpose and goals; (5) degrees of specialization of subject or topic. Under each category, further detailed items of textual features are specified: they vary from one to four. I should add that such an analytical framework is designed so as to facilitate the identification and study of contextual circumstances in the source text that may have a bearing upon Liu's particular use of certain archaisms.

The comprehensiveness of the framework is reflected in the inclusion of four complementary perspectives on the interactive relationship among subjective factors such as the relative status between the speaker and his interlocutors, the speaker's or his listeners' attitude to the speech event and its content, and objective factors like the exposure of the protagonist to a public speech environment and the level of specialty of the speech contents. All these situational variables extracted from the source text are deemed to bear relevance to the deployment of archaic idioms in the target text, which as a whole will furnish a general framework for some initial quantification of source textual information.

To facilitate the use of statistical software, the itemized text features are then assigned either dichotomic or gradient value in accordance with the continuity of measurement of each variable. Due to the non-numeric nature of the variables featured in the adapted analytical scheme, at the first stage of quantifying the framework, the different categories of variables are labelled along an ordinal scale which reflects the sequential rather than the interval relationship among the different values of variables.

Such procedure converts the information conveyed in textual data into quantifiable categorical variables to allow the generation of inferential statistics by using relevant tests such as CATPCA, at a later stage. However, since the variables thus encoded may well have been quantified along different scales of measurements, in order to make them comparable within a single statistical model, the further operation of optimal scaling is deemed to be crucial. In the Statistical Package for Social Sciences (SPSS) 15.0 version⁴⁶, the function of data modification is furnished under Data Reduction and is embedded within the CAPTCA suite.

Table XXII shows a fully encoded version of the adapted analytical framework from Biber's study of register. As can be seen, there are in total seventeen textual features retained in the adapted analytical scheme which, when compared to Biber's original framework featuring some thirty items (see Appendix IV), has been initially streamlined with nearly half of the original variables removed as

⁴⁶ In previous chapters, I have been using different statistical packages which are mostly online (the statistical computing language R is downloadable from its website). Chapter VI is the place where I start to use SPSS.

being of little relevance to the study of archaism in a textual context.

Table XXII Adapted version of Biber's analytical framework for the current study

PARAMETER CATEGORY	SITUATIONAL VARIABLE	VALUE	CODE
I Communicative characteristics of participants	1.1 Addressee's number (AN)	Single	1
		Plural	2
	1.2 Audience (AU)	Absent	1
		Present	2
II Relationship between addressor and addressee	2.1 Social Status between speaker and addressee (ST)	Addressee < Addressor	1
		Mixed	2
		Addressee > Addressor	3
	2.2 Extent of shared specialist knowledge of topic (SSKT)	Low	1
		Mixed	2
		High	3
	2.3 Distance of relationship (DR)	Low	1
		Mixed	2
		High	3
	2.4 Formality of relationship (FR)	Informal	1
		Mixed	2
		Formal	3
III Settings	3.1 Place of communication (PL)	Private	1
		Public	2
	3.2 Domain (DO) (adapted to the current study)	Non-chivalry related	1
		Chivalry related	2
IV Relation of participants to the text	4.1 Addressor's personal evaluation of speech content (ADE)	Non-important	1
		Important	2
	4.2 Addressee's personal evaluation of speech content (AEE)	Non-important	1
		Mixed	2
		Important	3
	4.3 Addressor's attitude to speech content (ADA)	Distanced	1
		Everyday	2
		Emotionally evolved	3
	4.4 Addressor's epistemological stance towards speech content (ADEP)	Against	1
		Dubious / Uncertain	2
		Accept	3
		Belief	4
V Speaker's purpose, intent ad goals	5.1 Speaker's proposition based on levels of factuality (FA)	Imaginative/ Speculative	1
		Mixed	2
		Fact	3
	5.2 Persuasion (PE)	Low	1
		High	2
	5.3 Instruction / transfer of information (IN)	Low	1
		High	2
	5.4 Self revelation (SR)	Low	1
		High	2
VI Subject/topic	6.1 Levels of discussion (LE)	Specialized	1
		General/ everyday	2

Such structural modification of the analytical framework has been based on the researcher's personal judgment and evaluation, and the appropriateness of this adapted analytical system will be further examined through the statistical modelling of textual data retrieved from Subdivision VIII and IX. It should be noted that such initial streamlining does not constitute any structural modeling of the variables; neither has any potential assumption been made regarding the nature of the variable sets, which will be performed by CATPCA after a proper manual encoding of the TL textual data.

6.3 Variable reduction and pattern recognition (Subdivision VIII and IX)

6.3.1 Some notes on manual data encoding

The encoding of situational variables in Table XXII points from the lowest to the highest category, which makes it ready for further statistical modeling at a later stage. An important point to be made here is that the scaling of situational variables listed in the framework must point in the same direction as the object under scrutiny. In other words, given that the analytical framework to be built through principal component analysis is aimed to predict Liu's use of archaisms in the second half of his translation, the quantification of observed variables should all point from the lowest to the highest end of the continuum, along which the odds for the occurrence of a linguistic event, e.g. Liu's use of archaism, tend to increase.

For instance, to encode the first two situational variables labelled under the category of communicative characteristics of participants, i.e. number of addressees and the presence of an audience, contextual situations which do not seem to favour the use of archaism in translating the protagonist's speeches have been assigned with a lower rank of category, e.g. a single addressee or the absence of an audience listening to the protagonist's utterances. On the other hand, features of contextual environment which implicitly encourage one's use of archaism in translation have been marked with a higher rank of category, e.g. the plurality of addressees and the presence of an audience.

The directional uniformity in the categorization of observed variables is essential for a successful statistical modelling of the data at a later stage. Any inconsistency in the data scaling would cause confusion in the computation of factor scores which will be then entered into the hypothetical model testing by using another type of statistical procedures known as logistic regression (see Chapter VII). However, such a manual encoding process may not be straightforward as it first appears to be. In some cases, our intuition or everyday experience somehow fails to make a clear distinction between typical contextual situations that favour the use of archaism, as the values of observed variables change.

For instance, in quantifying the variable Relative Social Status between the speaker and addressees, one may argue that it is likely that don Quijote will use archaic language when talking to someone of an inferior status, a lower social rank, such as Sancho, to show off his knowledge of and familiarity with themes of medieval chivalry. Nevertheless, at the same time, it also makes sense to

assume that the higher social status of the interlocutor, e.g. Princess Micomicona, may also inspire the protagonist to use archaic language, which is actually the case with many of the speech archaisms identified in Liu's translation of Subdivision VIII.

Similarly, in quantifying the situational parameter Addressor's Epistemological Stance of Speech Content, it has been quite difficult to establish a link between the translator's choice of archaism and the most likely attitude the speaker may have towards the speech event. Therefore, in quantifying observed variables like Relative Social Status and ADEP, the proposed methods here for scaling are largely investigative in nature at this stage. The inclusion of such situational variables provides referential information on the source text context where an archaic use of language has been spotted. Further justification for retaining them in the framework, as well as the appropriateness of the direction of phrasing thus suggested, will be left for the statistical modelling of the dataset.

Appendix V exhibits the numerical encoding of speech archaisms identified in Subdivision VIII and IX of Liu's translation of Part I. There are several points I should make here regarding manual categorization of variables, which, despite the great efforts made to reduce the subjectivity involved in data processing, have turned out to be a constant factor in similar research. There is a full version of the encoding results in Appendix V, which is left open for further discussion. Meanwhile, I should like to point out that the quantification of subjective contextual variables (for example, addressor or addressee's evaluation of speech content) is not as straightforward as the measurement of objective factors such as Audience and Setting. This is also the reason why an ordinal rather than interval categorization of these variables is believed to be more accurate and less prone to deviations in the manual classification of the original dataset within the analytic framework of situational parameters.

For instance, given the close relationship between the addressor's personal evaluation of the importance of speech content and the significance of communication as perceived from the addressor's point of view, the majority of archaisms in Subdivision VIII, with the single exception of 23/30, have been marked with the higher ADE value as important, for they all have appeared in contextual environments revolving around one of the central themes in the novel, which is don Quijote's thoughts and deeds in trying to act and behave like a knight in shining armour.

This includes don Quijote's keenness to ascertain who has played the role of messenger, the reaction of his beloved damsel upon receiving his love letter (24/30, 27/30)⁴⁷; don Quijote's obsession with the idealized image of his Dulcinea (36/31, 41/31); the justification of his total dedication to the Lady Dulcinea in the face of Sancho's perplexity (44/31). Although it is arguable that there is some variation among these archaisms as regards the degree of the importance attributed to the protagonist's evaluation of speech contents, they complement each other, in reflecting from different angles don Quijote's enthrallment and purposeful identification with medieval ideals of knightly virtues, honor and

⁴⁷ The n/m expression is composed by (n) the sequential number of the speech and (m) the total number of speech by don Quijote in each subdivision; in Subdivision VIII there are totally 30 speeches by don Quijote, and in Subdivision IX there are 31 speeches by the protagonist in total.

courtly love, and are thus distinctly different from the speech content of 23/30, where don Quijote appears to be fairly impatient of Sancho's idle muttering, which is far too frequent in the setting throughout the novel.

The categorization of ADA is also ordinal and divided into three consecutive levels, ranging from distanced through everyday to emotionally-involved. The lowest level of distancing refers to the speaker's apathetic attitude to the speech content or relevant speech events, which is situated at the negative extreme of the speaker's attitudinal continuum. The intermediate level, everyday, is designated to describe the speaker's attitude which views the speech act as a form of communication containing everyday information.

For example, 23/30 has been assigned with an ADA value at this particular level, for it is easy to perceive that when don Quijote expresses his frustration at Sancho's earthy talk with an euphemism by saying "*cuantas veces va el cantarillo a la fuente...*", his attitude towards the conversation with Sancho has already gone from exasperation and annoyance to leniency and forbearance, accepting Sancho's grim persistence as an everyday element in their binding adventure. On the other hand, with the exception of 23/30, all other speech archaisms in Subdivision VIII have been marked with the highest ADA value as Emotionally Involved. This is due to the fact that the content of the protagonist's utterances is closely related to the main theme of the novel, i.e. Don Quijote's exaltation and enchantment with the idealized image of Dulcinea.

Another important situational variable is the addressor's ostensible desire to persuade. The substantiation of this variable on the CAPTCA model is realized through a binary process of categorization which arranges the variable into low and high level. The positive component loadings of the majority of archaisms on the second dimension as shown on the biplot indicates that they should have a higher score on the variable Addressor's Persuasive Intent, when compared with 23/30, which is somehow distant from the area covered by this variable.

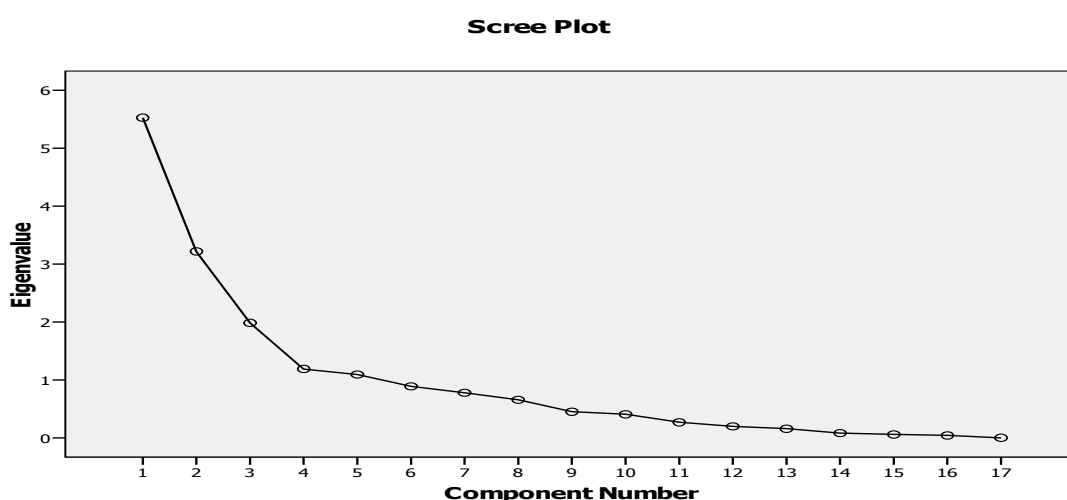
The marked shift of the variable Persuasion within archaisms found in Subdivision VIII is again largely due to a gradual transfer of content in the speech events involved. That is, after a heated debate between the knight and his squire regarding the status and importance of Dulcinea relative to Princess Micomicona, don Quijote first tones down his speech and moderates his language looking for reconciliation between himself and Sancho (23/30). Although it might be difficult to pin down the reason why don Quijote has suddenly calmed down, it may be reasonable to surmise that Sancho's accidental mentioning of his last visit to Dulcinea has turned don Quijote's attention from the dispute itself to enquiring about the outcome of Sancho's visit to his beloved Lady.

24/30 marks an important change in don Quijote's attitude. The knight becomes increasingly keen to find out the impact his love letter had on Dulcinea. Whereas Sancho is unable to provide the kind of response his master desired, partly because he is not very familiar with the traditions of chivalry,

and partly due to the fact that he did not manage to make the trip at all, don Quijote was only too eager to fill in the gaps by dint of his feverish imagination. As a result, don Quijote's role gradually moves away from the inquirer to that of the alternative-answer-provider, and in an attempt to obtain Sancho's endorsement – which by the way don Quijote has so far rarely succeeded in doing–, don Quijote's tone grows increasingly, deliberately persuasive.

6.3.2 Dimension Solution

Diagram XIII Scree plot for CAPTCA model construction (Subdivision VIII and IX)



In using the exploratory technique of principal component analysis, the first issue that needs to be addressed is the so-called dimension solution. Since CAPTCA's main purpose is to reduce a large range of observable variables into a small number of abstract or conceptual components, the first question that comes to mind after applying the statistical test on the encoded data is to decide how many principal components or dimensions should be retained in the CAPTCA model, an interpretive process which may not be so straightforward as it may seem.

There are some widely accepted empirical procedures that may assist the process of dimension solution (Hatcher, 1994: 16-9). First, we could produce a scree plot which associates each dimension with its corresponding eigenvalue⁴⁸ in a consecutive manner. Any abrupt drop in the line segment plot indicates a potentially dividing point between the highest dimension that the CAPTCA model may achieve and the remaining factors that do not contribute significantly to the model. From the diagram shown above, we can see that there is a visible drop of eigenvalue between dimension four and five. From dimension four onwards, the difference between every two components draws increasingly nearer. We may therefore set the highest solution at dimension four as suggested by the scree plot. There is however another rule of thumb in CAPTCA which prefers dimensions with eigenvalues relatively

⁴⁸ In a linear transformation, an eigenvalue is used to measure the length and direction of the transformation of that vector. For example, if the eigenvalue is +1 it means that the eigenvector does not change at all; and +2 means that the eigenvector is now twice longer but there is no change in its direction. For further reference see eigenvalue (2008) in Encyclopedia Britannica <http://www.britannica.com/EBchecked/topic/920628/eigenvector>

higher than 1.0, known as Kaiser Criterion (Bryman and Cramer, 2005:330). According to this empirical rule, the inclusion of dimension four would seem to be problematic, since its eigenvalue is very close to the cutting point 1.0, which in turn suggests its very limited contribution to the CATPCA model.

6.3.2.1 Two dimension model

a) Variance Accounted For (VAF)

To streamline the analysis, we shall adjust the highest dimension from four to three and compute related statistical information for the two and three dimensional model respectively to see which solution may best suit the dataset collected from Liu's translation of Subdivision VIII and IX.

Table XXIII Model Summary of two-dimension solution (1/2)

Dimension	Cronbach's Alpha ⁴⁹	Variance Accounted For	
		Total (Eigenvalue)	% of Variance
1	0.872	5.567	32.748
2	0.793	3.941	23.183
Total	0.951(a)	9.508	55.931

Table XXIII provides a summary of the results obtained by applying the categorical principal component analysis to Subdivision VIII and IX as set at the two dimensional level. An important statistical result provided in the summary is the degree of variance calculated for each dimension, as well as the cumulative total of variance explained by the statistical model as a whole. As shown in the summary, the first dimension thus created has been assigned with a relatively large eigenvalue of 5.567 – noting that the lower limit is 1.0 –, which is then translated into some thirty per cent of the variance in objects as explained by the first principal component.

The second dimension has been given an eigenvalue of 3.941. Consequently, we observe a similarly proportional decrease in the value of Variance Accounted For (VAF) on the second dimension. Apart from some thirty per cent of variance explained by the first dimension, the second dimension, which is statistically uncorrelated with the first component, helps explain another twenty per cent of the variance in contextual variables relating to the use of archaism in Subdivision VIII and IX.

The value of VAF attributed to each dimension as generated by CATPCA is nested, which means that together they may be added up to assess the efficiency of the statistical model thus built. In our study, the two dimensional CATPCA model constructed for Subdivision VIII and IX reports a total VAF value of 55.931, which suggests that the analytical framework adapted from Biber's study on register variation does work, although it may not be the best one available.

⁴⁹ Cronbach's Alpha is an important indicator, which measures the internal consistency of the conceptual scales of a statistical model. It varies from zero to one: the higher the score, the more reliable the generated scale is (Reynaldo and Santos, 1999).

Two possible reasons may account for the relatively low VAF for the two dimension model. Firstly, there could be less efficient variables regarding the distribution of speech archaism in those extracted from Biber's original framework, which has been devised in any case for the study of register variation under more general circumstances. The first condition of constraint may be seen therefore, as the lack of efficiency in variables. Secondly, redundant or less well-defined categories may also add instability to the statistical model. Since my main purpose in using the CATPCA is to identify contextual factors that are most relevant to the use of archaism as opposed to plain language, those with uncertain contributions, for example, equally high, to each dimension of the CATPCA model will also be removed to increase the efficiency of the instrument as a whole. These are indicated by a strikethrough in the table below.

In CATPCA, the statistically built dimension of analysis represents a linear combination of a subgroup of the original variables intended to quantify the object under investigation, e.g. the aspectual features of the context in which Don Quijote's archaic speeches have been found. As very different from the original variables like subject, audience, relative social status between the protagonist and his addressees, etc. which assume largely perceivable properties to allow the gathering and preparation of a database for later textual analysis, the concept of principal component is definitely more abstract, which only makes sense in statistical terms. Consequently, to describe the perceptual quality of a principal component, it is necessary to resort to loadings or weightings of original variables modelled on that particular dimension. This may be done through a breakdown of VAF attributed to each variable on the statistically built principal components.

Table XXIV Component Loadings for two-dimensional solution (1/2)

	Dimension	
	1	2
Addressee number	-.121	.647
Audience	-.256	.590
~~Relative social status~~	~~-.647~~	~~.511~~
~~Shared specific knowledge of topic~~	~~-.345~~	~~.027~~
~~DR~~	~~-.686~~	~~-.651~~
~~EP~~	~~-.686~~	~~-.651~~
Place	-.343	.647
Domain of Subject	-.754	-.318
Ad Evaluation	-.697	-.382
~~AE Evaluation~~	~~-.609~~	~~.502~~
~~Ad Attitude~~	~~-.679~~	~~-.476~~
Addressor's epistemological stance of speech content	-.718	-.410
~~Factuality~~	~~.442~~	~~.460~~
~~Persuasion~~	~~-.550~~	~~-.572~~
~~Instruction~~	~~-.286~~	~~-.321~~
Self-Revelation	-.655	-.043
Levels of Discussion	.692	.329

In order to examine the statistical relevance of the variables included in the framework, we shall proceed to compare individual VAF values of variables incorporated in the two dimensional CATPCA model for Subdivision VIII and IX. The table above presents an overview of the loadings or weightings of variables on each dimension of the two-dimensional CATPCA model. Through scrolling down the cross-tabulation of the statistical results, it becomes clear that two situational parameters have proved to have little VAF value on both dimensions when compared with the remaining items listed in the index. They are SSKT (Shared Specific Knowledge of Topic) with a VAF of 0.345 on Dimension-1 and 0.027 on Dimension-2; and Instruction with a VAF of 0.286 on Dimension-1 and 0.321 on Dimension-2. The removal of these two variables may help to satisfy the first condition of data refinement, i.e. lack of efficiency, as mentioned earlier on.

Another distinctive feature of the two-dimensional model is the equally high loadings of variables on both dimensions, if we set 0.4 or above as high values of variable loadings (Hatcher, 1994: 18). Variables with such statistical properties are Relative Social Status with a VAF of 0.647 on Dimension-1 and 0.511 on Dimension-2; Distance or Formality of Relationship with a VAF of 0.686 on Dimension-1 and 0.651 on Dimension-2; Addressee's Evaluation of Speech Content with a VAF of 0.609 on Dimension-1 and 0.502 on Dimension-2; Addressor's Attitude to Speech Content with a VAF of 0.679 on Dimension-1 and 0.476 on Dimension-2; Factuality with a VAF of 0.442 on Dimension-1 and 0.460 on Dimension-2; and lastly, Persuasion with a VAF of 0.550 on Dimension-1 and 0.573 on Dimension-2. To improve the efficiency of the two dimensional model, these variables with underspecified high loadings on both dimensions will be removed, which will then satisfy the second requirement of data refinement.

b) Reversal of data encoding

An important consideration here refers to the direction of phrasing. The importance of maintaining the same directionality of encoded variables has been mentioned prior to the statistical data modelling (see 6.3.1 Some Notes on the Manual Data Encoding). Furthermore, due to the difficulty observed in establishing an explicit link between the typical values of situational variables and the translator's preferred use of linguistic archaism as opposed to plain language under specific source text circumstances, the initial categorization has been largely tentative. Now, according to the component loadings computed for each of the variables included in the three dimensional CATPCA model, we are consequently in a better position to revise the statements made previously regarding the appropriateness of the encoding scheme.

In Table XXIV we can see that within the two dimensional model, while most of the variables with a meaningful loading on the first scale exhibit a negative score, one variables turns out to have positive value on the same dimension, i.e. Levels of Discussion, which thus require phrase reversal. To reverse the data encoding, there are two basic ways, one automatic, and the other manual. The relatively easier way to perform this task is to create a new variable with the same name by following

the formula shown below (Hatcher, 1994: 39): Variable-name = Constant - variable-name.

In this formula, the quantity of the “constant” is the number of points on the response scale used with the framework item plus one (Ibid). For instance, the variable Levels of Discussion has two categorical values which are Specific (1) and General (2). Since the direction of the original encoding as such has proved to contradict the remaining variables on the first dimension, it is necessary to reverse the original encoding scheme to ensure the genuineness of the reliability score.

Table XXV Model Summary of two-dimension solution (2/2)

Dimension	Cronbach's Alpha	Variance Accounted For	
		Total (Eigenvalue)	% of Variance
1	.794	3.274	40.930
2	.583	2.042	25.519
Total	.928(a)	5.316	66.449

Table XXV shows the result of the reconstruction of the two dimensional model which entails the deletion of variable items with less well defined loadings on both dimensions, and the reversal of coding of the variable Levels of Discussion, which has shown a negative relationship with the rest of variables defining the first dimension. Compared with the original two dimensional model, the efficiency of the streamlined one has improved significantly, from 55.9 to 66.5 per cent. The total increase in the VAF is largely due to the enhancement of the first dimension regarding the quantification of speech content. At the same time, the second dimension still accounts for more than one fifth of the variation in the original dataset.

As for the lower limit of VAF for an acceptable CATPCA model, a generally endorsed figure is 0.7 (Hatcher, 1994: 25-6). However, as it is common in statistical analysis, the interpretation of a rule of thumb, which is the empirically set lower bound of 0.7 in this case, normally requires a fuller contextualization of the generalized rule with the actual dataset under investigation. Both Hatcher (1994: 137) observes that the cutting-point of 0.7 for the total VAF used in dimension decision is often implied in dealing with real-life data in social or behavioural science investigation, which is however a true reflection of the internal complexity or diversity of the constructs being measured. In fact, in some quantitative social studies, the VAF reported have been well below 0.7, even below 0.6. Therefore, in applying rules of thumb in quantitative linguistic research, the researcher ought to adopt a pragmatic attitude in assessing the fitness of a statistical construct built within a theoretical model.

c) **Structural fitness and conceptual scale construction: criteria of interpretability**

Another important issue that we need to make sure of after the structural modification of the two dimensional model is whether with the removal of variables, the original structure is undermined, which may have the following external representations: (1) variables with meaningful loadings on one dimension may now have shifted to another dimension; (2) variables with well-defined contribution to

the two dimensional model may have now been squeezed out due to the distributional uncertainty caused in the course of reconstruction. To examine the effects of the structural modification on the 2D model, we have computed a similar breakdown of the component loadings shown in table below.

Table XXVI Component loadings for two dimensional solution (2/2)

	Dimension	
	1	2
Addressee number	.005	.768
Audience	.171	.865
Place	.170	.746
Domain of subject	.922	-.065
Addressor's evaluation of speech content	.642	-.174
Addressor's epistemological stance of speech content	.814	-.278
Self Revelation	.723	.188
Levels	.876	-.018

To assess the fitness of the streamlined dimensional structure, it becomes necessary to resort to the fourth set of criteria suggested in Hatcher (1994: 26-8), which is known as the interpretability criteria. Compared with the previous three sets of predefined empirical rules, e.g. scree-plot, total VAF and Cronbach's Alpha, the last criteria focuses on the interpretation of the conceptual meaning of retained principal components, a process which is largely model-specific and requires discrete judgement on the part of the researcher.

The interpretability criteria may be further specified through the observation of three essential structural properties of the statistical model which have been summarized below (Ibid):

- (1) Do the variables that load on a given component share the same conceptual meaning?
- (2) Do the variables that load on different components measure different constructs?
- (3) Does the factor pattern demonstrate "simple structure"?

As we may see, the first observation on the structural features of the CATPCA model aims to measure the substantive meaning that the statistical-built latent structure may contribute to our understanding of the object under investigation. In a similar manner, the second rule of the interpretability criteria attempts to gauge the efficiency of the CATPCA model from the perspective of the relative independence of one principal component from another.

In the last evaluation rule, a simple structure is defined by the very high loadings of a group of variables on one dimension and their much lower loadings on other dimensions within a CATPCA model. As mentioned earlier on in the construction of the two dimensional model, the equally high or low loadings of variables on more than one dimension reflect from different perspectives the inefficiency of the model; that is, the observed lack of relevance, redundancy or definitional ambiguity of variables included for statistical analyses.

On the contrary, if the principal component analysis builds up to a statistical model of original variables in which the linear combination of variables or the artificially built dimensions, i.e. principal components, shows a simple structure with original factors neatly classified into one dimension or another, the whole model is markedly improved when compared to models with uncertain variables.

From Table XXVI, we can see that a distinctive feature of cross tabulation is that less well defined or redundant variables, such as those with equally low or high loadings on both dimensions, have all been eliminated from the reconstructed two dimension model. All variables retained in the framework have been better classified and fit into each of the two statistical dimensions thus built. That is, all the variables incorporated into the analytical framework are well discriminated as shown by their separated loadings on each of the two dimensions.

The measurement of speech content is achieved through the quantification of a number of situational factors which move along a continuum of subjective versus objective factors for the assessment of speech content. At one end, we have contextual variables that seem to evaluate from a relatively more objective perspective the formality of the protagonist's utterances, such as Domain of Subject and Levels of Discussion. At the other hand, the high/low nature of speech content is estimated indirectly from the perception by parties involved in the speech event, e.g. the addressor's evaluation of speech content, addressor's epistemological attitude to speech contents, addressor's intent to reveal himself through showing off his knowledge of speech contents. The second dimension is substantiated by three variables measuring the contextual environment in which the speech event takes place, i.e. addressee's number, presence or absence of audience and venue of communication.

A closer look at the distribution of component loadings on the principal components reveals that the reconstructed two dimensional CATPCA model exhibits a statistically cleaner and conceptually simpler structure than the original model does, which consequently meets the last criteria for data modelling, i.e. the interpretability criteria.

d) **Testing of Scale Reliability**

As a final check of the reconstructed 2D CATPCA model, we shall test its scale reliability. In the refined two dimensional model, the first dimension is now mainly substantiated by five variables which at the same time report limited loadings on the second dimension: Domain of Subject, Addressor's Evaluation of Speech Content; Addressor's Epistemological Stance of Speech Content; Addressor's Intent of Self-Revelation and Levels of Discussion. The stability of the first dimension, which focuses on the assessment of **speech content** as partially from the speaker's point of view, is well measured by the statistical indicator of Cronbach's Alpha, which remains as high as 0.794.

On the other hand, the second dimension is now defined by three variables related to **speech**

environment, which are Addressee Number, presence of Audience and Place of Communication. The internal stability of the second dimension has proved to be lower than that of the first dimension, since the variables used to define Dimension-2 have been reduced from five to three. This is also reflected in the relatively lower Cronbach's Alpha value attributed to Dimension-2, which has dropped from 0.793 in the original two dimensional model to the current 0.583.

The progressively deducted Cronbach's Alpha values assigned to dimensions of a higher order are however understandable when taking into consideration the way a CATPCA model is built. In constructing a principal component for CATPCA, the first scale always seeks to include as many measuring variables as possible and thus tends to have the largest Cronbach's Alpha within the whole model. In a similar fashion, the construction of the second scale is based on a maximally optimized combination of variables left out of the first dimension. Consequently, as the variables integrated in dimensions of a higher order drop, the statistical indicator of internal strength, i.e. Cronbach's Alpha, accordingly diminishes.

The visible drop in Cronbach's Alpha seems to suggest reduced reliability of the restructured second dimension. As Hatcher warns, assessing scale reliability should be one of the first tasks a researcher has to take care of. If the scales built in principal component analysis or exploratory factor analysis do not show adequate reliability, there is no point in conducting any further analysis (Hatcher, 1994: 140). Therefore, before proceeding to test the two dimensional model with more textual data collected from Liu's translation, we shall first examine the internal consistency of the two reconstructed dimensions. The testing of scale reliability may be done through the ANALYSIS --> RELIABILITY procedure provided in SPSS, version 15.0.

Table XXVII Reliability Statistics for Dimension 1

Cronbach's Alpha Based on Raw Variables	Cronbach's Alpha Based on Standardized Items	Number of Items
.702	.808	5

The general scale reliability of the refined two dimensional model is shown in Table XXVII. There are two methods to compute the statistical indicator of scale reliability, i.e. the coefficient alpha, which are the Cronbach's Alpha (α) based on raw variables and the one generated based on standardized items. The score based on raw variables assumes less restrictive conditions in measuring internal relationship between observed variables within a scale, while the later follows the classic parallel test assumption (Garson, 2008). In any case, an overall α as above 0.7 indicates that the model is an acceptable one (Hatcher, 1994: 137). In this case, the α value attributed to the speech content scale reflects a reasonable degree of reliability.

Table XXVIII Inter-Item Correlation Matrix for Dimension 1

	Domain	AD Evaluation	Ad Epistemology	Self Revelation	Levels
Domain	1.000	.467	.376	.534	.910
AD Evaluation	.467	1.000	.366	.412	.412
Ad Epistemology	.376	.366	1.000	.316	.316
Self-Revelation	.534	.412	.316	1.000	.468
Levels	.910	.412	.316	.468	1.000

Another important measurement of scale stability is the inter-correlation of items included within a conceptual dimension. That is, if we define the internal consistency of a statistical construct according to the similarity of observed variables belonging to the same dimension, the stronger the correlation among these variables, the higher the stability of the scale. Table XXVIII shows a breakdown of the inter-item co-variance among the variables constituent of the speech content scale. As we can see from the correlation matrix, the correlation score between any two variables ranges from 0.316 to 0.910 (numbers on the orthogonal line are the self-correlation scores). This is a good sign of the scale reliability, for in order to assure of the relevance of the variables classified under the speech content scale, a minimum value of 0.30 as the covariance score is required, which is set as another rule of thumb in explaining correlation-related tests (Field, 2005: 672).

Table XXIX Item-Total Statistics for Dimension 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Domain	8.94	2.511	.700	.852	.597
AD Evaluation	8.92	2.764	.519	.287	.655
Ad Epistemology	7.31	1.533	.423	.199	.822
Self Revelation	8.97	2.656	.516	.326	.646
Levels	8.97	2.542	.616	.830	.615

Table XXIX presents another cluster of statistical information which exhibits the strength of the speech content scale. While the correlation matrix reveals the inter-item relationship among the variables, Table XXIX examines the scale reliability from the perspective of one item in relation with the whole set of variables incorporated into the principal component. This is achieved through the comparison of the overall statistical scores before and after the removal of a certain variable. Any significant change shown in each column should be treated with caution, since this may well be a sign of the abnormal behaviour of the specific item thus highlighted, and thus questioning the suitability of its existence within the whole dimension.

Two most revealing sets of information are shown in the columns of Correlated Item-Total Correlation and Cronbach's Alpha If Item Deleted. The column labelled Correlated Item-Total shows the correlations between each item and the total score from the speech content scale. In a reliable scale,

all items should reasonably correlate with the total. The minimal value prescribed for a valid correlation is 0.3 (Ibid), and as we may see from the middle column Correlated Item-Total Correlation, each single variable demonstrates a relatively high correlation with the rest of the items built in the speech content scale.

The hierarchical distribution of the correlation scores suggests which variables are most representative of the conceptual scale, and which variables play a more supplementary role. That is, since Domain and Levels show higher correlation with the total, it can be said that the conceptual scale of speech content is mainly defined by these two objectively oriented variables; while the other three variables focused on the speaker's point of view over the speech content are more supplementary.

The last column of Cronbach's Alpha If Item Deleted shows the values of the overall α if that item is not included in the calculation. As one may infer from the function of this statistical estimation, any decrease in the value of α when compared to the score measuring the overall scale reliability (see Table XXVII) implies the relevance of that variable to the whole conceptual dimension; whereas any increase in the value of α when compared to the general Cronbach's Alpha means that the deletion of that item would inevitably lead to an improvement of the scale reliability. As may be seen from the Table XXIX, the only item that seems to contradict this stipulation is the variable Addressor's Epistemological Stance of Speech Content. It is shown that if we remove this item from the speech content scale, the overall reliability would increase from .808 to .822. However, such variation is not so dramatic and certain flexibility should be allowed in the assessment of models built with real-life data (Field, 2005: 672).

So far, we have completed the inspection of the reliability of the speech content scale. It has been shown that with the data collected from Subdivision VIII and IX, the first conceptual dimension within the refined two dimensional CATPCA model displays an acceptable level of internal consistency. It may therefore be used as a hypothetical construct in later model testing. The three tables below show the results of reliability test with the second dimension, i.e. speech environment. Since the testing procedure is largely the same, we shall skip a detailed analysis as conducted with the first dimension, and offer only a brief evaluation before going on to consider the fitness of the three dimensional model in the current study.

Table XXX Reliability Statistics for Dimension 2

Cronbach's Alpha Based on Raw Variables	Cronbach's Alpha Based on Standardized Items	N of Items
.734	.730	3

The scale reliability test shows that the second dimension Speech Environment enjoys a similarly high level of internal consistency, since the Cronbach's Alphas computed based both on the raw variables and on the standardized items exceed the empirically stipulated cutting-point of 0.7. To

have a closer look at the inter-item correlations among the three variables encompassed within the second dimension, we produce the inter-correlation matrix as shown in Table XXXI.

Table XXXI Inter-Item Correlation Matrix for Dimension 2

	Addressee number	Audience	Place
Addressee's Number	1.000	.553	.598
Audience	.553	1.000	.272
Place of Communication	.598	.272	1.000

Despite good overall scale reliability, the correlation matrix has spotted an important discrepancy among the three variables: while the correlation levels for the two variable pairs, i.e. Addressee's Number and Audience, Addressee's Number and Place of Communication, are quite high as above 0.5; Audience and Place does not seem to inter-relate with each other very well, for the correlation score computed for this variable pair is merely 0.272. However, at this stage, it is still difficult to decide which variable actually correlates less well with the second dimension as a whole, and for this purpose, we have generated the Item-Total Statistics which is shown in Table XXXII.

Table XXXII Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Addressee's Number	3.08	.536	.722	.522	.427
Audience	2.92	.764	.465	.311	.748
Place	3.33	.686	.505	.362	.708

In the middle column labelled as Corrected Item-Total Correlation (ITC), it becomes clear that the variable that co-relates best with the second dimension is Addressee's Number, for the ITC score attributed for this variable is as high as 0.722, well above the empirically preferred low limit of 0.3. At the other hand, in spite of the slight differences between the two variables, both Audience and Place correlate quite well with the second dimension in general, for the respective ITC scores for these two variables are 0.465 and 0.505, which still show quite a distance from the lower bound.

The last column Cronbach's Alpha (of the scale) If Item Deleted presents similar information regarding the varying relevance of the three variables within the speech environment scale. If we take out Addressee's Number, the total α would drop sharply from 0.730 to 0.427, which suggests the significance of this variable within this conceptual scale. To a lesser extent, the variable Place also contributes positively to the internal consistency of the second scale. The minor increase in the total α with the removal of Audience is permitted here on the ground of the flexibility normally required in dealing with real-life data in quantitative social sciences (Field, 2005: 673).

Now we have come to an end of the construction and evaluation of the two dimensional

principal component models built on Biber's original analytical framework. The refined CATPCA model is composed of two conceptual scales which are Speech Content and Speech Environment. As the scale reliability test shows, both scales are reasonably reliable and the two dimensional CATPCA model may therefore be used as a feasible hypothetical construct or model to predict and explain variation in Liu's use of archaism collected from the target text dataset of a larger size.

6.3.2.2 Three dimension solution

In the initial stages of the construction of the CATPCA model, we noticed from the scree plot (Diagram XIV) that the eigenvalues for both the two and three dimension solutions are above the commonly accepted lower limit of 1.0, a rule of thumb which was initially set in psychological research and is known as the Kaiser's criterion (Kaiser, 1960: 141-51). Therefore, without sufficient empirical evidence, one could not select one model in preference to the other. To assess the suitability and efficiency of the three dimensional model, in this section, we shall use the same exploratory statistical technique trying to build a three dimensional CATPCA model, and then compare the structural optimization between the two models to see which one serves better the purpose of the current study, which is to construct a hypothetical theoretical model for a co-textual explanation of style variation in literary translations.

Table XXXIII Model Summary of three-dimension solution

Dimension	Variance Accounted For		
	Cronbach's Alpha	Total (Eigenvalue)	% of Variance
1	0.875	5.669	33.349
2	0.779	3.755	22.090
3	0.517	1.946	11.445
Total	0.969(a)	11.370	66.883

Table XXXIII presents the results of the construction of a three dimensional CATPCA model. As we can see, the total efficiency of a model of a higher order is roughly the same as the refined two dimension model. At the same time, while the internal stability of the first two dimensions are unquestionably high, the Cronbach's Alpha computed for the third dimension is somehow below the empirically preferred value of 0.7 (Nunnally, 1978). This is a sign of the potential existence of less well-defined variables in this initial three-dimension model. To streamline the framework, we have computed the distribution of component loadings in the three dimension model which is shown in Table XXXIV.

We can see in Table XXXIV above that there are four variables which have proved to be less well defined in the three dimension model, i.e. FR, DR and Addressee's Evaluation of Speech Content and Persuasion. FR, DR and Persuasion have shown similarly high loadings on both Dimension 1 and 2, which in turn have resulted in the conceptual ambiguity of the first two dimensions. In a similar fashion, the variable Addressee's Evaluation of Speech Content loads heavily on both the second and third

dimension.

Table XXXIV Component Loadings of the 3D CATPCA model

Variable	Dimension		
	1	2	3
Addressee number	.071	.662	-.429
Audience	.222	.608	-.495
Status	.706	.381	.514
Shared specific knowledge of topic	.348	.008	.519
~~DR~~	~~.680~~	~~.672~~	~~-.040~~
~~FR~~	~~.680~~	~~.672~~	~~-.040~~
Place	.324	.662	-.152
Domain	.750	-.288	-.361
AD Evaluation	.680	-.365	-.039
~~AE Evaluation~~	~~.651~~	~~.442~~	~~-.415~~
Ad Attitude	.674	-.460	-.079
Ad Epistemology	.744	-.421	-.043
factuality	.511	-.319	.384
~~Persuasion~~	~~-.567~~	~~.590~~	~~-.166~~
Instruction	-.210	.362	.495
Self Revelation	.645	-.015	-.327
Levels	.691	-.287	-.392

To improve the three dimension CATPCA model, we decided to remove these four variables. However, the removal of these noisy elements from the initial framework does not seem to have improved the quality of the three dimension model. On the contrary, the process of model refinement has *squeezed out* another two variables, i.e. Place and Instruction, while bringing the Cronbach's Alpha of the second and third scale further down below the cutting-point 0.7 (see Table XXXV and XXXVI).

Table XXXV 3D Model Refinement (1/3)

Dimension	Cronbach's Alpha	Variance Accounted For
	Total (Eigenvalue)	Total (Eigenvalue)
1	.848	4.596
2	.614	2.307
3	.419	1.630
Total	.956(a)	8.532

Failure in restructuring the initial three dimensional model points to the inherent instability of the CATPCA model of higher dimension solution. Regardless of the already very low Cronbach's Alpha of third dimension, we carry on with the structural refinement, in the hope of achieving a relatively acceptable model in the end.

Table XXXVI Component loadings of refined 3D model (1/3)

	Dimension		
	1	2	3
Addressee number	-.238	.772	-.082
Audience	-.017	.819	.315
Status	.564	-.338	.645
~~Place~~	~~.052~~	~~.618~~	~~.607~~
Domain	.844	.240	-.191
AD Evaluation	.731	.013	-.159
Ad Attitude	.794	-.042	-.206
Ad Epistemology	.816	-.020	-.120
factuality	.626	-.368	.124
~~Instruction~~	~~-.387~~	~~-.203~~	~~-.264~~
Self Revelation	.618	.386	.107
Levels	.814	.246	-.151
Shared specific knowledge of topic	.298	-.311	.709

The equally high loadings of Place on the last two dimensions require its deletion from the framework, since it has shown certain definitional ambiguity. Meanwhile, the low loadings of Instruction on all of three dimensions suggest its irrelevance to this model at this stage of variable reduction. In view of this, one would need to repeat the same procedure of variable removal and data refinement until the whole model shows a statistically simple structure, according to the interpretability criteria.

Table XXXVII Model Refinement (2/3)

Dimension	Cronbach's Alpha	Variance Accounted For	
		% of Variance	Total (Eigenvalue)
1	.853	4.452	40.477
2	.563	2.050	18.637
3	.205	1.229	11.169
Total	.958(a)	7.731	70.282

Table XXXVIII Component loadings of refined 3D model (2/3)

	Dimension		
	1	2	3
Addressee number	-.234	-.818	.226
~~Status~~	~~.580~~	~~.511~~	~~.429~~
Shared specific knowledge of topic	.316	.500	.683
Domain	.840	-.328	-.138
AD Evaluation	.730	-.089	-.120
Ad Attitude	.794	-.051	-.188
Ad Epistemology	.812	-.059	-.112
factuality	.650	.320	.032
Self Revelation	.598	-.310	.103
Levels	.807	-.330	-.032
Audience	-.039	-.664	.658

The result of the second round of model refining turns out to be even less encouraging: with the removal of Place and Instruction, the Cronbach's Alpha has gone down to a level which is not acceptable in any terms, whereas the second dimension shows a similarly fast trend of destabilization. In the last attempt to retain and improve this 3D model, we take out the only variable that heavily loads on all three dimensions which is the Relative Status between the two parties involved in a speech event.

Table XXXIX 3D Model Refinement (3/3)

Dimension	Cronbach's Alpha	Variance Accounted For	
	Total (Eigenvalue)	% of Variance	Total (Eigenvalue)
1	.852	4.123	45.816
2	.371	1.493	16.585
3	.049	1.047	11.628
Total	.956(a)	6.663	74.029

Table XL Component loadings of refined 3D model (3/3)

	Dimension		
	1	2	3
Addressee number	-.152	.757	.097
Shared specific knowledge of topic	-.206	.737	-.096
Domain	.879	.261	-.220
AD Evaluation	.742	-.054	.604
Ad Attitude	.811	-.094	.494
Ad Epistemology	.829	-.038	-.218
factuality	.561	-.375	-.484
Self Revelation	.613	.302	.061
Levels	.836	.252	-.293

Table XXXIX and XL show the result of further model refining. In the final version of the three dimensional CATPCA model, the component loadings of the variables retained in the framework show an improved and cleaner distribution on each of the three dimensions. However, the stability of the last two dimensions has been seriously undermined. Actually the third dimension has been practically eliminated from the three dimensional model, for there does not seem to be any variable with meaningful loadings on this dimension at all. Consequently, we may safely disregard the possibility of building up a three dimensional model, which so far has shown a high degree of instability and cannot therefore be used as a reliable theoretical construct in future model testing.

6. 4 Conclusion

In this chapter I have tried to build a two dimensional CATPCA model of situational parameters for style variation in Liu's use of speech archaisms in Subdivision VIII and IX. As mentioned earlier on, CATPCA is an exploratory statistical technique usually employed to tease out an abstract statistical model of patterns from a large number of variables that may explain as much of the variance in the objects as possible without losing substantial information. In the current study, CATPCA has been deployed with a view to establishing possible contextual patterns in the original text that may help describe stylistic variance identified in the last three subdivisions of Liu's translation of the novel.

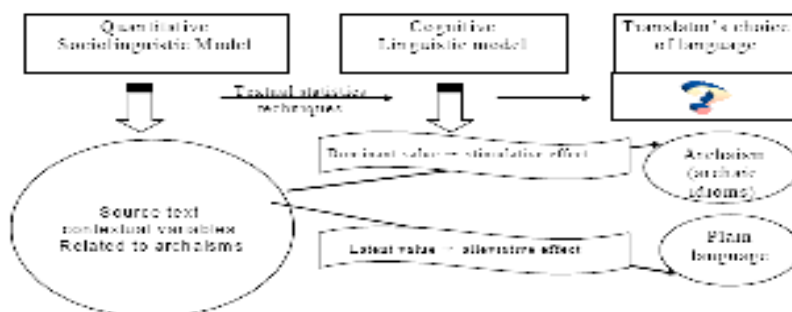
The general strategy of resorting to co-textual features in the original texts, as a way of trying to explain stylistic variation in literary translation, is based upon the following line of reasoning. In a preliminary study, comparing the use of archaic speeches in *Don Quijote* and in Yang's first direct translation of the novel, which by the way Liu has already acknowledged to be very familiar with⁵⁰, I was able to build up a linear regression model which successfully predicted the patterns of the use of archaisms in the first half of Liu's translation. Such finding came to light as a function of the statistical model of variation in don Quijote's archaic speeches found in the original and especially in Yang's version as two explanatory factors. The linear regression model has proved to be less revealing as regards the occurrence of don Quijote's archaic speeches in the second half of Liu's translation, which has then been singled out as stylistic variance within Liu's version of *Don Quijote* (I).

It is my contention, as a working hypothesis may be, that in addition to an inherent dependence of translations upon original texts, translators, or at least those conscious of establishing a distinctive style of their own to achieve certain stylistic effects, would not spare efforts to incorporate their own understanding of the original into their works. This they may well do to take full advantage not only of linguistic cues, but also of situational features (usually overlooked), serving as important contextual guides in the process of manipulating and recreating image of fictional figures within the socio-cultural environment of their own languages.

In view of above, I suggest that a proper explanation of such stylistic variance may hold the key to a better understanding of the nature of literary translation as a special form of creative writing. Within the context of this thesis, the definition of translational creativity may hopefully extend to the exploration of a translator's own understanding of an exotic contextual reality, or the image of fictional characters, to be rendered into a totally different linguistic system.

⁵⁰ According to the records of a personal interview with Liu in April, 2007

Diagram XIV Context-motivated theory and style variation in literary translation



As Diagram XIV shows, I assume, within the purview of a context-motivated theory, that there is always an implicit link between the relevant contextual features obtaining in the original text, in this case *Don Quijote* Part I, and the final outcome as expressed in the particular use of language in the target version, which for the purpose of this thesis consists specifically in the translator's choice between archaic idioms and plain language.

In order to measure or test such hypothetical association between the source text feature and the particular use of language by the translator, each highlighted source text contextual factor or variable has been classified into a taxonomic system in which it assumes either a stimulative or alleviative effect on the translator's cognitive understanding of the source text contextual circumstances which are open to individual's interpretation such as archaism. A crucial procedure in testing the proposed theoretical model consists in the form of quantitative analysis that has been put to work through the statistical technique known as Categorical Principal Component Analysis.

The above-mentioned statistical procedure, i.e. principal component analysis is an advanced exploratory technique which helps identify potential patterns in relevant contextual features in the process of analyzing the rationale behind the stylistic variation detected in Liu's use of archaisms in the last second half of the novel. Such research procedures in my view make a significant step forward in the construction of theoretical models to investigate the contextual and co-textual nature of style variation in literary translation; it should help systematize the actual process of textual analysis, and potentially offers a prolific line of empirical research in corpus-based translation studies.

Chapter VII Testing of Statistical Model for Style Variation (Liu, I)

7.1 Model testing and logistic regression

The purpose of the previous chapter (VI) is to explore the underlying patterns in the source text contextual factors that may help explain stylistic variation in Liu's translation of the second half of Part I of *Don Quijote*. The statistical technique used is known as categorical principal component analysis. As different from hypothesis testing techniques, such as linear regression used in Chapter V, principal component analysis is an advanced exploratory statistical procedure which is used in situations where the researcher has not yet made any assumption regarding the latent structure of the variables selected for special study.

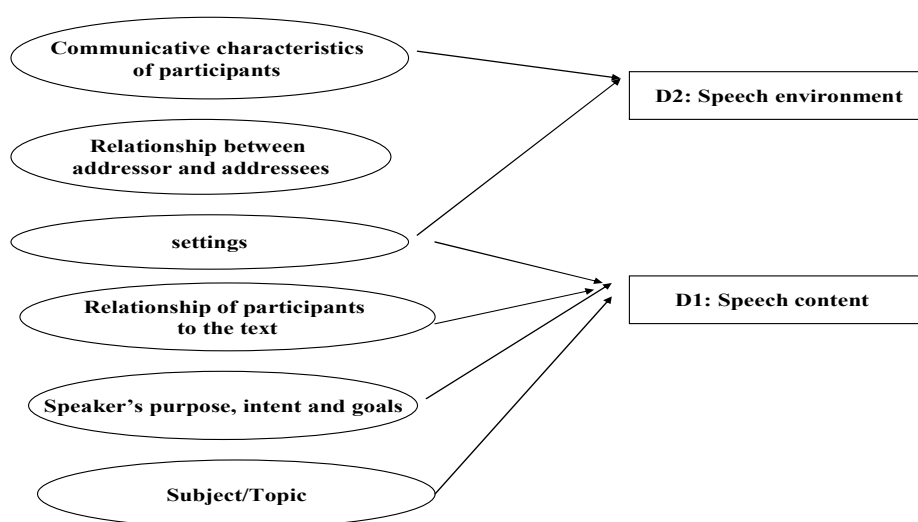
It is an important technique to investigate variable reduction and for subsequent classification of remaining variables into a small number of conceptual constructs. The statistically built conceptual dimensions are meant to be least correlated with each other so as to explain to the maximum the range of variation in the object dataset. In the current study, which sets out to explore the underlying patterns in source text contextual factors, the principal component analytical procedure has successfully managed to build up a two dimensional model which has helped me solve the two main problems presented at the outset of Chapter VI: (1) identification of source text variables most relevant to the variation in speech archaism detected in Liu's work; and (2) construction of a customized hypothetical model for speech archaism (see Table XLI), derived from Biber's general analytical framework for register variation (see Table XXII).

Table XLI 2D CATPCA Model for Stylistic Variation in Speech Archaism

	DIMENSION 1	DIMENSION 2
CONCEPTUAL CONSTRUCT	SPEECH CONTENT	SPEECH ENVIRONMENT
Constituent Variables	Domain of Subject	Addressee's number
	Addressor's evaluation of speech content	Audience
	Addressor's attitude to speech content	Place of communication
	Addressor's epistemological stance of speech content	
	Addressor's proposition based on factuality	
	Addressor's intent to instruct	
	Addressor's intent to reveal himself	
	Levels of discussion	

To be more specific, if we compare the two investigative frameworks designed for variation in the use of speech archaism (VSA) and for general register variation (GRV), we should be able to notice the fundamental structural modifications that principal component methodology brings to Biber's original framework when it is applied here in the course of data modelling.

Diagram XV Statistical structural modification from GRV to VSA



In Diagram XV, the six factors in ovals represent the six major types of situational parameters highlighted in Biber's model for GRV. As different from the statistically built model for VSA in the present study, the GRV model is essentially an intuitive framework, where the constituent elements listed have been proposed largely based on the researcher's accumulated experience in the field, and they may well overlap with each other to a varying extent. On the other hand, the VSA model proposed in the current study has been constructed statistically and is composed by two abstract (or artificial) principal components, which are the optimal combinations of individual factors categorized under the six main types of situational parameters included in the GRV model.

The intersection of items belonging to different parameter categories points to the internal connectivity of Biber's network, especially when it is applied in more narrowly defined subject domains such as speech archaism. The principal components featured in CATPCA also represent underlying conceptual scales which are minimally correlated to each other as a function of the CATPCA's modelling process. In this sense, the VSA model is more parsimonious than the GRV model, though its compositional integrity needs to be developed further and its wide applicability must to be tested.

In this chapter, I propose to test the validity of the VSA model as a way of predicting the frequency of occurrence of speech archaism in the second half of Liu's translation of *Don Quijote* Part I. In building up the VSA model, we have been using data collected from Subdivision VIII and IX; and in order to further assess the predicting capability of the said model, we shall use a different kind of statistical procedure which is called logistic regression. The data used in this case will also include instances of speech archaisms identified in the last subdivision of Liu's translation.

This generic type of logistic regression tests is widely used for predicting the behaviour of

dependent variables, as a follow-up of general variation in a set of predefined independent variables. It is a specific type of regression where the dependent variable takes on dichotomic values between zero and one (Oakes, 1998: 42). It has been selected here on account of the dependent character of the variable under investigation, i.e. the unexpected rise and fall in Liu's use of speech archaism in the second half of the novel (I).

In Chapter V, the use of linear regression technique revealed that, if we look upon Liu's use of archaism in translating the protagonist's utterances as a joint consequence of the use of archaisms in both the original and Yang's first translation of the novel, while the two explanatory factors, i.e. the original text and Yang's version, have successfully predicted Liu's use of archaism in the first seven subdivisions of his translation, some irregularities do come to the surface in the last three subdivisions of his work. That is, whereas the protagonist's speeches feature significantly in many more archaisms than those predicted in Subdivision VIII and IX, the actual frequency of don Quijote's archaic speeches drops sharply at a similar rate in the final subdivision of Liu's translation.

As a result, I shall collate all the instances of speech archaisms occurring in Liu's translation of Subdivision VIII, IX, regardless of whether they may or may not have a functional equivalent in the original or in Yang's version of the novel. With regards to the data gathered from Subdivision X, the situation appears to be slightly more complicated. The number of speech archaisms present in the last six chapters of the novel shows a reversed trend of variation, decreasing rather than increasing. To reflect such change, in preparing the dataset for Subdivision X, we shall gather instances of archaism identified not only in Liu's translation, but also those in the source text and in Yang's translation. The missing slots for archaisms, which have been highlighted through comparing items from Liu's work as well as from the other two sub-corpora, point to the decline in the use of archaism in the last Subdivision (X) of Liu's version of the novel.

Logistic regression is widely used in quantitative social sciences or clinical experiments to predict the binary outcome of the dependent variable, as in the case of 'dead or alive', or, the likelihood of developing a cancer as 'yes or no'. In view of the binary nature of the dependent variable, i.e. archaism versus non-archaism, the regression type known as logistic would appear to be more appropriate for the type of testing required in the current study, which is intended to validate the VSA model as an efficient way of predicting the presence of speech archaisms in the last three subdivisions of Liu's translation of the novel.

A pioneering use of logistic regression in corpus linguistics was reported in Leech, Francis and Xu (1994: 57-76), where the researchers examined and compared the different hypothetical models for the choice made by native speakers between the use of [X's Y] and [the Y of X], which was a good illustration of the so-called phenomenon of gradience. In a more recent publication, Johnson has underlined the usefulness of logistic regression in quantitative sociolinguistics research (Johnson, 2008: 152), as well as its application in modelling syntactic corpora data (Ibid, 236). In the current study, we

shall follow similar procedures in the structural modelling of parallel corpus data and explore the potential of this versatile statistical analysis in the field of empirical comparative literature, especially in corpus-based translation studies.

Table XLII Dependent Variable Encoding (DVE) for Logistic Regression

DVE \ CHA P.	SPEECH SEQUENCE									
VIII	4	11	12	18	19	20	23	24	27	36
DVE	1	1	1	1	1	1	1	1	1	1
VIII	40	41	44	46	48	50				
DVE	1	1	1	1	1	1				
IX	6	8	12	13	14	15	17	18	25	26
DVE	1	1	1	1	1	1	1	1	1	1
IX	33	36	37	40	41	42	43	45	46	50
DVE	1	1	1	1	1	1	1	1	1	1
X	1	4	6	7	9	12	14	15	16	17
DVE	0	1	1	0	1	1	1	1	0	1
X	18	20	24	27	28	29				
DVE	1	1	0	0	0	0				

DVE: 0=Non-archaism; 1=archaism

As mentioned earlier on, in logistic regression analysis, the dependent variable may only take on dichotomic values between zero and one; therefore, to encode the data collected from Liu's work, we shall assign the value zero to speech items which do not contain an archaic use of language, and allocate the value one to the protagonist's speeches which feature an archaic expression in modern Chinese (see Table XLII). I must point out this arrangement of data encoding is not an arbitrary decision. In the SPSS software, the LOGISTIC procedure will predict the "one" category of the dependent variable, while making the "zero" category the reference category. In our study, Liu's increased use of archaism constitutes the bulk of the dataset, we shall therefore focus on the rationale behind the enhanced use of archaism as a predominant trend in Liu's work, while making the decreased use of archaism as the reference category.

7.2 Two Dimension CATPCA Model for Hypothesis Testing

Table XLIV shows the encoded values for the dependent variable to be entered in the logistic regression test, i.e. the presence or absence of speech archaisms in Liu's translation of the last three subdivisions. Another important procedure in the preparation of the data is the computation of factor scores for a predefined set of independent or explanatory factors (Hatcher, 1994: 31). The establishment of an analytical framework for independent factors has been provided by the VSA model which I constructed in the previous chapter. The dataset used in the previous chapter for the construction of the model CATPCA came from Subdivision VIII and IX only. But in this chapter, in order to perform the logistic regression for the purpose of hypothesis testing, I shall use the textual data gathered from Subdivision VIII, IX and X (see Appendix V and VI).

In statistical terms, a factor or component score in principal component analysis represents a

linear combination of optimally weighted observed variables. Within the context of the current study, a factor score is computed to describe in numerical terms the source text contextual environment which corresponds to the use of archaism in Liu's translation of the Castilian novel. The factor scores computed for all the instances of speech archaisms will then be used as values for the independent variables, which are the two principal components, instead of the original observed variables.

To generate the factor scores for the whole dataset to be used in logistic regression, we would need to repeat the principal component analysis with the raw data collected from all of the related chapters in the first part of the novel. The results of the principal component analysis are displayed in Table XLV. The efficiency of the VSA model, when applied to structural modelling for the whole dataset (hereafter refers to all three subdivisions), is somewhat comparable to the one built on the experimental dataset (henceforth referring to Subdivision VIII and IX only).

Table XLIII CATPCA Model Summary (Subdivisions VIII, IX and X) (Liu, I)

Dimension	Cronbach's Alpha	Variance Accounted For	
		Total (Eigenvalue)	% of Variance
1	.802	3.348	41.855
2	.514	1.817	22.707
Total	.922(a)	5.165	64.562

a Total Cronbach's Alpha is based on the total Eigenvalue.

In the first place, we can observe that the percentage of accumulated Variance Accounted For (VAF) remains above sixty, which indicates that the two dimension CATPCA model built upon the experimental dataset fits reasonably well into the whole dataset. Secondly, similar to the VSA model built in the previous chapter, the VAF attributed to each of the two dimensions also shows a similar distribution: while the first dimension explains some forty per cent of the variance in Liu's idiosyncratic use of archaism, the second dimension accounts for slightly more than one fifth of the variation in the dependent variable. Eigenvalue, another important statistical index shown in the model summary, reflects from an alternative perspective the fitness of the model. It clearly argues for the retention of two principal components, for the eigenvalue assigned to each of the two dimensions is well above the lower limit of one, according to Kaiser's Criterion.

The internal consistency of the two dimension CATPCA model is measured by the statistical indicator Cronbach's Alpha, which is generally used in statistics to measure the internal consistency of a model's scales. It varies from zero to one, and the higher the score, the more reliable the generated scale is (Reynaldo and Santos, 1999). With the enlargement of the input dataset, the reliability of the first dimension remains unquestionably high; while the Cronbach's Alpha assigned to Dimension 2 has dropped from 0.583 to 0.514. To have a closer look at the reasons that may have resulted in the decline of scale reliability in the second dimension, we have computed an index of component loadings for the observed variables included in the two dimension CATPCA model (see Table XLIV).

Table XLIV Component Loadings

	Dimension	
	1	2
Addressee number	.241	.682
Audience	.425	.791
Place	.409	.700
Domain	.881	-.250
AD Evaluation	.671	-.263
Ad Epistemology	.707	-.242
Self Revelation	.702	-.064
Levels	.850	-.204

As may be seen in Table XLIV, the distribution of component loadings still shows the desired simple structure (Hatcher, 1994: 30), where each dimension is well defined by observed variables with relatively higher loadings on one dimension, and largely reduced weights on the other. The underlying factor patterns regarding the formation of conceptual scales also remain intact: whereas the first dimension is sustained by five variables measuring the protagonist's speech content; the second dimension feature three contextual variables quantifying speech environment.

Despite the decrease in Cronbach's Alpha, no variables have been singled out because of their failure to register statistically in the process of model reconstruction. As a final check on the internal consistency of the model, we have computed the scale reliability test as well, before we could safely accept the result of the data construction with the whole dataset and subsequently proceed to compute the factor scores to be entered in the test of hypothesis as independent variables.

Table XLV Reliability Statistics for Dimension 1

Cronbach's Alpha Based on Raw Items	Cronbach's Alpha Based on Standardized Items	Number of Items
.698	.827	5

Table XLVI Inter-Item Correlation Matrix for Dimension 1

	Domain	AD Evaluation	Ad Epistemology	Self Revelation	Levels
Domain	1.000	.563	.356	.486	.925
AD Evaluation	.563	1.000	.359	.402	.513
Ad Epistemology	.356	.359	1.000	.545	.304
Self Revelation	.486	.402	.545	1.000	.432
Levels	.925	.513	.304	.432	1.000

Table XLVII Item-Total Statistics for Dimension 1

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Domain	8.88	2.810	.651	.872	.619
Addressor's evaluation	8.87	2.982	.535	.352	.653
Addressor's epistemological stance	7.38	1.379	.489	.324	.825
Self revelation	8.96	2.665	.628	.401	.606
Levels of discussion	8.90	2.834	.583	.856	.631

Table XLV, XLVI and XLVII show the results of the scale reliability analysis as indicated on Speech Content scale. In Table XLVII, both of the overall Cronbach's Alpha based on raw variables and standardized items are reasonably high: 0.698 and 0.827. In Table XLVIII, all the variables correlate well with each other with correlation scores ranging from 0.304 to 0.925. This is a good indication of the internal consistency of the Speech Content scale, since all the variables seem to be pointing to a common underlying construct.

Table XLVII reveals from an alternative perspective the contribution made by individual variables to the conceptual scale. From the middle column Correlated Item-Total Correlation, we may infer that the Speech Content scale is mainly defined by three variables, which correlate highly with the conceptual dimension as a whole. They are domain of subject (0.651), speaker's intent to reveal himself (0.628) and levels of discussion (0.583). The two more subjectively oriented variables, i.e. addressor's evaluation of speech content (0.535) and speaker's epistemological stance of speech content (0.489), turn out to have relatively lower item-total correlation scores, which seem to suggest the supplementary contribution that they make to the second dimension.

The distinction between the two groups is not difficult to comprehend. When considering whether to use an archaic expression or not, the Chinese translator is more likely to turn to contextual factors that are quite conspicuous rather than wasting a great deal of time trying to figure out what the protagonist's personal view over the speech content might be. The varying degrees of awareness or exposure on the part of the translator to the different contextual variables in the source text can be readily observed by focusing on the last column Cronbach's Alpha If Item Deleted.

The score in each cell along the last column represents the total α when the variable concerned is removed from the Speech Content scale. As we may notice from the cross tab of Item-Total Statistics, the variable Self Revelation points to the lowest total α when the item is subtracted from the first dimension. This is suggestive of the fact that the variable Self Revelation plays an important role in maintaining the internal stability of the second scale. If the total α turns out to be larger than the overall one, we may well have something to worry about, because the increase in α means that the deletion of that particular item may well improve the reliability of that scale. However, since all the α values vary

below the Cronbach's Alpha based on standardized item (0.827), we may therefore accept Speech Content as a reliable scale within the two dimensional model.

Table XLVIII Reliability Statistics for Dimension 2

Cronbach's Alpha Based on Raw Items	Cronbach's Alpha Based on Standardized Items	Number of Items
.737	.737	3

Table XLIX Inter-Item Correlation Matrix for Dimension 2

	Addressee number	Audience	Place
Addressee number	1.000	.507	.284
Audience	.507	1.000	.657
Place	.284	.657	1.000

Table L Item-Total Statistics for Dimension 2

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Addressee number	3.42	.680	.442	.261	.789
Audience	3.10	.520	.721	.543	.441
Place	2.98	.686	.544	.435	.673

A quick look at the results of the reliability test of the Speech Environment scale shows that it largely resembles the structural characteristics of the second dimension of the hypothetical CAPTCA model, which has been built on the dataset collected from Subdivision VIII and V: first, the overall α falls within the preferred value domain which is between 0.7 and 0.8. Secondly, the three variables correlate with each other reasonably well. In the cross tab for item-total statistics, each of the three variables correlates well with the rest of the conceptual scale, as their corrected item-total correlation score is above the lower limit of 0.3. As a result, we may accept it as a reliable scale built for the whole dataset.

7.3 Data Preparation: Computation of Factor-Based Scores

7.3.1 Creation of New Variables

PAC (Principal Component Analysis) has come up with a statistical model for the whole dataset which is composed of two conceptual dimensions of variables. The first dimension, which focuses on speech content, encompasses five situational variables, i.e. domain of subject, levels of discussion, speaker's evaluation of speech content, speaker's epistemological stance of speech content and speaker's intent to reveal himself. The second dimension, on the other hand, is substantiated by three variables concentrating on the speech environment: addressee's number, presence of audience and place of communication. To perform the logistic regression test outlined above, we could create two

new variables which could be named SPEECH CONTENT and SPEECH ENVIRONMENT, and compute the factor scores for these two new variables.

Factor scores thus computed are known as factor-based scores (FBS). They are different from true factor score (TFS) in that FBS represent the added values of all the variables that load highly only on a specific dimension, while TFS are the combination of all the object scores. To compute the FBS, one would first need to determine which items have high loadings on each of the principal components constructed within a PA model. Then for a given case, add together the case's responses to each of the highlighted variable on a specific dimension.

The result is the case's FBS on that principal component:

$$\text{FBS: D1} = \text{DO} + \text{ADE} + \text{ADEP} + \text{SR} + \text{LE}$$

$$\text{FBS: D2} = \text{AEN} + \text{AU} + \text{PL}$$

In the SPSS, the creation of a new variable based on a principal component can be done automatically after the construction of a principal component model. Go to TRANSFORM → COMPUTE VARIABLE; type in the name of the new variable in the space labelled as Target Variable, and write in the blank left for Numeric Expression the formula shown above. Also, prior to conducting the PA, one needs to specify the command of saving the object scores in an active dataset.

7.3.2 Variable Selection Method

The factor-based scores generated for the whole dataset are shown in Appendix VII. The distribution of speech archaisms within the two dimensional model is now expressed by the factor-based scores attained on each of the three conceptual scales. This statistical information has been fed into the logistic regression test to see to what extent the CATPCA analytical framework established for the quantification of the source text context may predict Liu's use of archaisms in the last three subdivisions of *Don Quijote* Part I.

This facility (binary logistic regression) is embedded in the SPSS software 15.0 version and classified under the category of Regression. To perform this statistical procedure, we first determine the variable selection method to be used in the validation of a particular hypothesis. There are two major types of algorithms designed for variable selection in logistic regression: forced entry method and stepwise method, both including forward and backward movement. Under the mode of forced entry, all of the input independent variables will enter into a single non-linear regression equation to predict the occurrence of the binary dependent variable (see Formula V).

Formula V Logistic regression equation

$$\text{logit}(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots + b_k X_k$$

In the formula shown above, p stands for the probability of presence of the characteristics of interest, for example, Liu's use of archaism as opposed to plain language. The logit transformation yields an equation in which the significance of each built-in independent variable is expressed through its corresponding coefficient, as $b_1, b_2, b_3, \dots, b_k$. As different from the forced entry method of model construction, the stepwise method allows the independent variables to enter into or exist from the regression model successively, according to the specific direction preset.

This is to obtain the most economical or parsimonious model to describe the variation in the original dataset of the dependent variable. In this sense, while the forced entry method is more suitable for hypothesis testing, the stepwise method lends itself well to data modelling of a more exploratory nature. Also, as Field (Field, 2005: 226) reiterates, within LOGISTIC procedure, the only appropriate method for theory testing is forced entry, while the stepwise techniques are not replicable due to the inherent random variation in data processing.

The aim of this part of the dissertation is to test the efficiency of the three dimensional CATPCA model specifically designed to explain stylistic variation in Liu's use of archaism in the second half of the novel. That is, the purpose for using logistic regression in this section is to test a particular hypothesis rather than to explore the underlying data structure. Therefore, in designing the statistical test, we shall specify in the LOGISTIC procedure that the forced entry method is the one preferred. The selection of different methods for model construction ends up generating different sets of statistical data, which in turn require different interpretations of the statistical information thus computed for a proper understanding of the fitness of the model for the dataset under investigation.

7.4 Interpretation of Statistical Results

As we have indicated from the outset of the quantitative part of the thesis, the ultimate purpose for establishing a theoretical construct through statistical analyses is to explain the rationale behind the kind of stylistic variation we have observed in the second half of Liu's translation of *Don Quixote* Part I. The theoretical assumptions or value judgments underlying the whole procedure of model construction and testing may be expressed this way: given the stylistic variation identified in Liu's distinctive use of language, which is somewhat different from the original text and from Yang's version, in particular, the most likely reason (certainly worth exploring further) to explain such stylistic variation is the way in which source text contextual factors have impacted on the translator's understanding of the original, as well as on his subsequent attitude to language style in his own work.

This line of reasoning seems to be especially compelling when it comes to explain style variation in the use of linguistic devices or textual structures which can be rather difficult to gauge from a diachronic point of view, such as archaism, which are therefore largely subject to translators' personal interpretation of the original text. One could say that, such a theoretical perspective might be

able to probe into the cognitive dimension of style variation in literary translation, which to the best of my knowledge, has not yet been properly explored, at least not with the help of solid empirical methodology (Boase-Beier, 2006: 75).

In the process of data modelling, the deployment of statistical tests has been quite instrumental in devising a hypothetical model composed of two principal components, i.e. Speech Content and Speech Environment, which describe the two most important contextual features of the source text that may have had a direct impact upon the use of archaism as a deliberate stylistic device in the target text. Is there a direct (causal) relationship between the way Liu read the source text in the light of the two principal components abovementioned and his own translating style? Such is the main line of enquiry as I approach the end of this study.

As to the outcome of this project I would suggest there are two possible endings likely to come up from this cognitive line of analysis:

(1) There could well be a clear link between the highlighted contextual features in the source text (*Don Quixote* Part I) and the observed stylistic variation in the target text (Liu's version). In other words, the identified stylistic variation is very likely to be a conscious and deliberate response to the source text on the part of the translator. If that be the case, the hypothetical model deployed in this chapter should be able to corroborate such relationship by being able to predict to an acceptable degree Liu's choice between archaism and plain language.

(2) On the other hand, it may well be the case that, as in many literary translations, any form of stylistic variation is mainly due to the translator's internally consistent or *free* use of language. Under such circumstances, it would be difficult for the theoretical construct to *perceive* or to establish a solid link between the highlighted source text contextual variables and the variation in the translator's use of certain linguistic devices. In the process of hypothesis testing, this will be reflected in the statistically insignificant contributions made by certain contextual variables in predicting the translator's use of language.

As may be expected, either of the two outcomes may occur within the interpretative framework constructed in the current study, which will help in any case to further our understanding of the cognitive nature of style variation in literary translation. In Table LI, we can see the results of the predictive analysis carried out with the LOGISTIC procedure. The two independent or explanatory variables entered into the hypothesis trial are Speech Content and Speech Environment, whose values are the factor-based scores derived from the two dimensional principal component model.

Table LI Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1(a)	Environment	-.309	.453	.467	1	.494	.734	.302	1.783
	Content	.237	.429	.305	1	.581	1.267	.547	2.935
	Constant	1.914	.427	20.096	1	.000	6.779		

a Variable(s) entered on step 1: environment, content.

The most revealing statistical information is shown in the last three columns. Since the null hypothesis of the logistic analysis is that the variables included in the equation (see Formula V) do not contribute significantly to the prediction of the values of the dependent variable, i.e. archaism versus non-archaism, any significant score above 0.05 indicates that the variable is not needed in the predictive analysis; and the 0.000 score attributed to the constant means that one may still have a good guess at the dependent variable without those explanatory variables.

Similarly, in the column labelled Exp (B), any score below 1.0 indicates a reverse trend in the increase and decrease of the independent and dependent variables. For example, with an increase in the factor score computed for the Speech Environment scale, there will be a decrease in the probability for the occurrence of archaism in Liu's work. On the contrary, if the factor score computed for the Speech Content scale increases, there will be accordingly an increase in the odds for the appearance of archaism. Lastly, in the column which shows the lower and upper confidence levels for the odds ratio, if the low-high range contains the value one, then the inclusion of that variable in the logistic regression equation will not make significant differences on the odds of the dependent, when compared to being in the reference category for that variable.

The predictive analysis reveals that the stylistic variation detected in the second half of Liu's translation actually points towards the translator's creative or experimental style of writing in dealing with linguistic issues that can be rather difficult to pin down from a diachronic point of view, such as archaism. The stylistic variation therefore may be approached as a growing style of writing, which the translator himself may not be well aware of, but has nevertheless quite clearly developed along the second half of the novel. That is, although in the first half of Liu's translation one may see that he is more derivative, quite ready to accept what he finds in Yang's work and in the source text, in the second half there is a perceivable shift where the particular linguistic feature that we are considering, i.e. archaism, becomes more frequent, and much more freely rendered.

7.5 Conclusion

Chapter VII is the last part of the quantitative analysis designed specifically to help understand the rationale behind stylistic variations in Liu's version of *Don Quijote* Part I. It is based upon the statistical results obtained in Chapter VI, and proceeds in a fashion which may generally be described as replicate (categorical principal component analysis) → testing of model reliability (scale reliability test) → testing of model efficiency through the so-called predictive analysis (logistic regression). The specific design of data modelling is chosen here in order to ensure that the statistical model, conceived in the light of assumptions implied in the context-motivated theory (CMT), comes up with substantial and reliable information which will later serve as the basis for further qualitative analysis. Furthermore, this methodological framework, which has so far proved to be highly replicable and quite fruitful in bringing to the surface interesting linguistic patterns, could be further improved in future research and quite possibly help to open up new possibilities of a more empirical approach from a quantitative perspective to corpus-based translation studies in general. .

Despite the relatively small size of linguistic data that I have been using in modelling stylistic variation regarding the use of archaism in literary translation, statistical analysis has, I hope, sharpened my understanding of the issue by bringing me deeper into the cognitive dimension of translating. The results obtained in the predictive analysis seem to suggest that the detected changes in Liu's use of archaisms in the second half of his translation is more likely to be a reflection of the translator's freer or more personal use of language for, in statistical terms at least, the logistic regression test does not seem to point to a strong link between the source text contextual features and the presence or absence of archaism in the target text. It must be pointed out that the negative outcome, which has been predicted right at the beginning of the chapter as one of the two tangible results, does not damage at all the significance and usefulness of the quantitative methodology. In fact, the logistic regression has successfully tested the hypothesis that Liu has mainly followed his own impression of the source text in using Chinese archaisms. In some other cases, if the translator has indeed been quite conscious about the use of the target language as dependent on the source text features, the statistical result may well turn out to be positive. It all depends on the specific nature of a translated text.

If the source text contextual features in *Don Quijote* Part I do not appear to be so influential in determining Liu's particular approach to the translation of this particular text, one must look at the writer/translator's socio-cultural background in search of other considerations of a social or cultural nature which might have had more of an impact in helping Liu evolve his own distinctive style in rendering *Don Quijote* Part I into contemporary Mandarin Chinese. The changing situation in China's publishing markets is obviously an aspect which must be looked at. When compared with his predecessor, Liu certainly belongs to a new generation of writers/translators in the 1990s who were fully aware of a fast-growing competitive publishing market seeking to satisfy the ever increasing demands of a potentially huge readership.

No doubt there is a very wide range of ability and grades of literacy among Chinese potential readers, but the writer/translator aiming to reach the highest possible number of readers will most probably seek to secure a large share of the market for foreign literature by adopting a language style which is as popular, natural, idiomatic, i.e. Chinese-sounding, as one could possibly make it. When it comes to a particular linguistic feature (figure of speech) to signify and to represent idiomaticity, such as archaism, it may not be surprising to discover that the translator appears to be guided by a sense of his own stylistic development as a writer, rather than subscribing too faithfully to the notion of rendering the original text with all its contextual and co-textual circumstances. So far, then, we could argue that the evidence gathered by statistical means with regards to Liu's personal use of archaism would seem to point to a more creative urge on the part of Liu, who as a contemporary Chinese writer, is perhaps more interested in delivering a text which is going to appeal to the literate as well as the less literate Chinese general public.

Chapter VIII Conclusion

“...[t]he corpus-based approach is evolving, through theoretical elaboration and empirical realization, into a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description and the practice of translation.”

(Laviosa, 1998: 474-9)

8.1 Summary of the main purpose of this thesis

The central aim of this PhD dissertation is to compare two modern Chinese versions of Cervantes's *Don Quijote de La Mancha* Part 1: Yang Jiang's in 1978 and Liu Jingsheng's in 1995. The selection of these two specific versions of *Don Quijote* has been based on the following considerations. First, the importance of Yang's translation has always been connected with the fact that her 1978 version is generally believed to be the first direct translation from Castilian into contemporary Mandarin Chinese and is widely used as the standard version of *Don Quijote* in secondary education in China. Secondly, Liu's is a more recent version which may be singled out for its popularity and huge commercial success. It is reported that by the end of the year 2008 there will be at least seven conventional publishing houses and a number of e-publishers in China that have opted for Liu's translation of *Don Quijote* among several existing versions of the novel: Li River Publisher (1995); Yi Li People's Publisher (2000); Beijing Yan Mountain Publisher (2001); Farland Publisher (2002); Yan Bian Publisher (2003); China Opera Publisher (2005); China Bookshop Publisher (2007), etc⁵¹. Such conspicuous publishing record is very rare among Liu's contemporaries, and it is indeed very indicative of the far-reaching impact of Liu's version in the re-introduction *Don Quijote* to contemporary Chinese readers of all descriptions.

Based on such observations, it is my view that a comparative study of these two Chinese versions of *Don Quijote*, from a stylistic point of view, will help to explain the nature and extent to which the language used by the two translators, separated by approximately two decades, has evolved in the context of changing economic, social and political circumstances. To the best of my knowledge, the model of analytical framework that I have set up specifically for the purpose of this enquiry is quite original, at least where this particular combination is concerned. Stylistics has usually been criticized for its cursory and sporadic treatment of empirical textual evidence, which is usually in short supply. John Sinclair (2004: 16), the pioneering British linguist, outlines the way to improve the situation regarding both quantitative and qualitative analysis of style, made feasible by the new corpus technology:

[1]⁵² Very detailed and careful analysis is required in stylistics, and it is still usually done by hand.

⁵¹ Last accessed on 12th June, 2008 http://www.soobbb.com/Destination_Wiki_3668_2.html

⁵² Numbers in squared brackets are mine

We are now in a position to be bold, to [2] look for testable hypotheses that a particular description of language can actually provide a set of boxes into which text can be appropriated. We must [3] look for models which help the text to reveal itself to us.

In at least three aspects, Sinclair's observations are indicative of innovative and potentially prolific future research in the field: (1) how to alleviate the labour-intensive nature of manual analysis in stylistics; (2) how to establish appropriate descriptive frameworks for systematic textual analysis; (3) how to establish testable hypotheses or replicable models that may reveal the nature of the texts under investigation. It appears to me that all three recommendations may actually be pursued in an integral and more efficient manner, i.e. through the deployment of methodologies and selection of data specially adapted from corpus linguistics. As a matter of fact, the fast growing use of corpora tools and methodologies in language and literary studies has already given rise in some people's minds to corpus stylistics as a discrete discipline engaging a large number of scholars all over the world who make it their main field of research⁵³.

My main contribution to this field of enquiry, I believe, consists in the way we put to the test the scope and capacity of corpus stylistics' tools and methodologies in the study of translation. To be more specific about my contribution I propose to focus on Sinclair's recommendations: (1) Aiming to keep the balance between manual and automatic textual analysis. (2) Testing the systematicity and replicability of quantitative models to process literary textual information. Quantitative methodologies have indeed proved to be very prolific in CL and corpus-based literary studies, and they have been explored mainly to expand the field from an empirical point of view. (3) Putting together theoretical models or positing hypotheses which will help to reveal the nature and character of translated texts.

8.2 Balancing manual and automatic textual analysis

The first two chapters of the thesis deal with the technical problems encountered in the process of building a parallel Castilian-Chinese corpus of *Don Quijote* Part I, retrieving corpus data automatically, and subsequently annotating linguistic data thus extracted. The technical problems first encountered in the construction of *Don Quijote's* parallel corpora come to the fore the moment when one attempts to segment Chinese natural language automatically, trying then to align parallel texts largely manually. Chinese text segmentation is of necessity a primary task in natural language processing: just as tokenization and lemmatization is for inflectional languages. The topic has been extensively discussed in computational linguistics, mostly from a highly technical point of view. But neither the general issue of Chinese text segmentation, nor the serious implications for corpus-based

⁵³ In fact, in the last few months there have been quite a number of meetings dealing with this particular area of research, to name a few, *Annual Meeting of the International Association of Empirical Studies of Literature*, July, Memphis, USA; *Biennial Congress of the International Association of Empirical Aesthetics*, August, 2008, Chicago, USA; the *Annual Conference of British Association of Applied Linguistics*, September, Swansea, UK (Ji, M. 2008d); the *Annual Meeting of the Poetics and Linguistics Association*, September, Sheffield (Ji, M. 2008e); *New Trends in Language Teaching and Translation Studies*, September, Granada, Spain; *Corpus Linguistics in Practical Application*, Oct., Viterbo, Italy.

language research in Chinese appears to have attracted a great deal of attention in applied linguistics in general and translation studies in particular. As a result, technical problems related to text segmentation in the construction of Chinese corpora may sound rather perplexing to less experienced corpus builders.

In the opening chapter of the thesis, I gave a brief account of the Chinese segmentation tool available at the time and provided some details of its performance on the two Chinese target texts; inevitably this had a significant impact on the way my dissertation was to develop later. Unlike Chinese corpora built for commercial purposes or by collective efforts in academic institutions, the segmented Chinese target texts in the parallel corpus of *Don Quixote* are raw corpus without the benefit of post-editing or any other human intervention. The automatically segmented texts thus reflect the actual level of accuracy that an appropriate segmentation tool may achieve with a variety of text genres. No need to add how advantageous this could be for individual corpus builders with limited time and most probably short on financial resources

Through a detailed classification of the segmentation results, obtained initially from a morphological perspective and then from the point of view of character n-gram, I have been able to observe that while most four-character expressions have been correctly segmented by the Chinese segmentation tool, mis-segmentation does occur frequently in a variety of two-character segments also produced by the natural language processing tool. The rate of mis-segmentation goes up visibly where foreign names and cultural-linguistic events are concerned. I have taken good care to note any deviant behaviour on the part of the segmentation tool. It is obviously a basic prerequisite for corpus-based language studies that linguistic data collected automatically or manually from corpora must be validated. A corpus database which potentially contains a large amount of faulty linguistic information would clearly impair any conclusion one might make regarding the translated texts under investigation.

From there on, my attention moves to the use of four-character expressions in the two target texts as generated by the Chinese segmentation tool, which exhibit a very high level of accuracy when compared to any other types of Chinese phraseological categories. In Chapter II, the automatically retrieved four-character expressions have been further classified and annotated, mainly manually, according to a pilot typological framework of Chinese four-character-gram phrasemes proposed on the basis of previous studies of Chinese idioms, and more importantly, observations on the behaviour of four-character words or phrases found in the parallel corpus texts. I may refer to the annotation scheme I am using here as a problem-oriented kind of annotation. I would like to suggest in this regard that in testing the feasibility of extracting data automatically prior to annotating the data; this problem-oriented scheme is more experimental and more circumspect in dealing with individual corpora than other generally established annotation schemes. I have contributed significantly to the development of this highly practical mode of research.

In conclusion, in the first two chapters of the dissertation, my two main tasks were: (1) the

construction of a Castilian-Chinese corpus of *Don Quijote* Part I, solving technical problems encountered in the process of Chinese text segmentation; and (2) the establishment of a flexible annotation scheme for the processing of automatically retrieved linguistic data from the parallel corpus. It may be said that if the originality of the investigation presented in the first chapter is largely technical, the value of the discussions in Chapter II is more methodological. These two initial chapters form an essential and coherent section of the current project, one of whose main objectives is to further our understanding and development of the instrumental possibilities that corpus material and methodology could bring to empirical translation studies. The successful construction of the parallel corpus and the subsequent corpus data annotation complement each other and facilitate the great potential benefits brought about by corpora, especially with regards to the balancing of automatic and manual analysis of translated texts.

8.3 Systematicity and replicability of methods and procedures for textual information

Chapter III, IV and V centre upon quantitative and qualitative analysis of the underlying phraseological patterns detected in Yang's and Liu's translation of *Don Quijote* Part I. My purpose was to ensure that any observations of a qualitative kind on the two translated texts are firmly grounded on a systematic and replicable statistical analysis of the database. I should make the point in this connection that many past translation projects which aimed at exploring corpora quantitatively have shown to be rather restricted to the computation of basic descriptive information derived from parallel or translated texts: i.e. frequency of tokens or word types, keywords indexing, node word collocation patterns, etc. Such general statistical procedures are no doubt critical in corpus-based translation studies, but have a limited application. There must be a substantial space in this area that should allow for further experimentation with methods, procedures and techniques borrowed from more established research areas like computational linguistics, textual statistics or textmetrics, stylistics or authorship attribution.

Despite current efforts made by corpus-oriented translation researchers in an attempt to get the most out of the limited statistical information available for the study of parallel or translated texts, there seems to be a methodological bottleneck. Further processing and/or elaboration of quantitative corpus data are deemed to be constrained by an invisible hand. The influence of such an invisible hand may be partly explained by the fact that research methodology is indeed a weak point in the education of translation students, since quantitative analysis is not yet a normal component in translation studies curricula. Secondly, and probably more significantly, what holds back the application of an open quantitative approach to the study of translated texts is the general feeling that language and literature, and in particular questions of style in the art of writing (the aesthetic side of it, that is) are never going to be easily computable or quantifiable.

Computing linguistic data from literary texts may well sound mechanistic and chancy to many

literary scholars. With some justification they may well be inclined to believe that the creative powers of a writer/translator reside at the heart of 'good' literature, and a good writing style cannot be reduced to a numbers game. Style is very much a matter of taste. My own view about literary translation as to the quality of a particular rendering vis-à-vis the source text or a different translation of the same original text appears to have evolved, with the help of some experimentation with techniques and procedures adapted from related disciplines as mentioned above, from a rather simplistic outlook to start with to the development of quite a complex theoretical construct which, fortunately, is proving to be quite productive. I daresay the linguistic data so far extracted from translational corpora in the course of this study would seem to show that criticism levelled against a quantitative treatment of linguistic data is unwarranted.

The use of quantitative methods need not distort in any way an in-depth study of literary translation. On the contrary, an informed use of statistical methods does greatly help to broaden our outlook on the art of writing or translating with a claim to literary status, and further deepen our understanding of hidden linguistic phenomena which may be hard to detect by means of statistical observation of translational corpora only or by closely reading the texts alone. Secondly, as for the suitability or compatibility of certain research methods borrowed and adapted from the social and linguistic sciences with corpus-based translation studies, I would also argue against a dismissive attitude, and in favour of a cautious and experimental approach towards the validity and usefulness of quantitative analysis in the study of literary translation.

As the case may be with data collected in the social or natural sciences, so it would seem that linguistic information retrieved from translational corpora has its own characteristics and will thus require different treatment when subjected to quantitative analysis. In expanding any subject area, researchers experiment with methods that are likely to be useful and productive in their own field of enquiry. Mistakes are made, inevitably, from time to time, but one can also learn a great deal from methods and procedures well tried in other fields. For my part, I am well aware of the risks that may obstruct or even impair the design and development of an adequate statistical framework for corpus stylistics or corpus-based literary studies. I have therefore always been minded to explain every step of the way when choosing statistical tests, or in interpreting the results obtained from such tests. In this pioneering project, suggestions and statistical advices from colleagues or anonymous reviewers of top academic journals working in different areas around the world have indeed been vital in securing the methodological suitability of the models that I have tried to establish for the purpose of my research (see Acknowledgement).

The typological framework set up in chapter II for the specific purpose of analysing four-character expressions automatically extracted from the parallel corpus of *Don Quijote* Part I, provides the basic instrument for further processing and structuring data. Furthermore, it affords the sort of empirical ground on which to build on later. But it causes some problems when it comes to the task of data annotation by manual means. In order to minimize the burden of manual data categorization and to

avoid the kind of inconsistency that might creep in any lengthy process, I have sampled at random some three hundred instances of validated four-character expressions from Yang's and Liu's translation, respectively.

The two sets of samples selected from the two Chinese translations are then compared by means of the statistical procedure known as the chi-square test. This procedure has been, and still is widely used in corpus linguistics to ascertain whether two sample populations differ significantly from one another. This particular test shows that despite the fact that Liu acknowledged his familiarity with Yang's work and that, presumably, he regularly went back to Yang's work in search of inspiration when working out his own version of *Don Quijote*, there are still significant differences between the two Chinese versions as regards the pattern of distribution of different types of four-character expressions.

In this connection I suggest two external (socio-cultural) factors which may have contributed to the different distributional patterning of four-character expressions mentioned in the previous paragraph. In the first place, Mandarin Chinese itself must have been undergoing some change over time. The seventeen years between the publication of the two Chinese translations under investigation (1978 and 1995, respectively) were a crucial period in modern Chinese history in which Deng Xiaoping's famous open-market policy and related economic reforms came very dramatically to the fore. It must have seriously impacted the whole of the country, and the literate classes, in particular: the language could not have remained unaffected either. Secondly, the generational gap between the two translators must have also conditioned their different ways of writing and translating. Prof. Yang (1911-) was already in her mid sixties when she embarked on the translation of *Don Quijote*; whereas Liu (1956-) was barely forty when his version of *Don Quijote* was published.

The use of corpora once again plays a crucial role in bringing these diachronic factors to bear upon the stylistic differences between the two Chinese translations. As described in some detail in Chapter II, four-character expressions represent a phraseological category very typical of Mandarin Chinese, signalling a very complex and heterogeneous component of the language system. Selecting and highlighting these four-character expressions, a process largely conditioned by my own research, do not necessarily add up to a cluster of conventional tagsets which could be used in the annotation of Chinese corpora. That is because the discourse function assumed by four-character expressions as a whole is more rhetorical or idiomatic than purely linguistic. Nevertheless, since the subtype of four-character expressions which separates Liu's work from that of Yang's is basically Chinese idioms, for the purpose of tracing any possible linguistic variations by means of monolingual corpora I have focused my attention primarily on this specific tagging category.

The selection of corpora as a basis for comparison has not been that straightforward either. First and foremost, in order to make a valid comparison between corpora, there must be some similarity between the different sampling frameworks used in the corpus construction. In search of contemporary

Chinese corpora that may satisfy this prerequisite, only two stand out as suitable candidates: the Lancaster Corpus of Mandarin Chinese (LCMC) (data gathered in the 1990s) and the UCLA Corpus of Chinese (data gathered in 2000s). The different time spans of the texts collected in the two corpora would seem to suggest that given the corpus resources available at the time we would probably get just a glimpse of the kind of change with regard to the distribution of idiomatic expressions that Mandarin Chinese as a linguistic system may have undergone. Therefore, to put to the test my main working hypothesis regarding the different ways Yang and Liu use idioms as against comparable (chronologically, at least) corpora, we should require a much larger and broader database than is currently available.

On the basis of the results obtained from comparing Yang's and Liu's translated texts, I look closely in chapter V to a possible triangular relationship between the three sets of texts, i.e. the original Castilian novel and the two modern Chinese versions. Firstly, I compare the two target texts in relation to the use of figurative versus non-figurative language in the source text. From this separate exercise it would seem as though in general Liu's use of figurative idioms is much higher than Yang's. Furthermore, such distinctiveness as one can detect in Liu's higher use of idioms, in my view may well be due to a closer connection with the contextual circumstances of the original work. In fact, when the source text appears to shift from non-figurative to figurative use of the language, Liu is normally more responsive to such variation by using similar figurative idioms than Yang is.

The use of several statistical procedures to compare the three sets of texts as regards the distribution of archaic idioms brings to light other intriguing findings. The Pearson's moment-product correlation test, for example, spotlights the following: (1) there is a proven strong correlation between Yang's and Liu's use of Chinese archaic idioms, which Liu himself has acknowledged in a personal interview (mentioned in Chapter I: 1.2); (2) while the similarity between the original and Liu's version can be corroborated statistically, the pattern and frequency of distribution of archaic idioms in Yang's work turns out to be less contingent (dependent) upon the source text. Further to the Pearson's moment-product correlation test as outlined above there is also another statistical procedure known as linear regression test which is quite similar to the correlation procedure, but can also help to predict the behaviour of dependent variables in the course of observing variations in controlled variables. Based on the findings elicited by Pearson's correlation test, four possible models of causal connection between the three corpus texts have been tested here as a working hypothesis. The first three models predict the frequency of occurrence of archaic idioms in Liu's translation on the basis of a calculation of the number and frequency of archaisms in Yang's translation and in the original text (Model I), archaic idioms in Yang's version (Model II), and the distribution of archaisms in the original text (Model III), respectively.

The linear regression test further confirms that Models I and II are quite successful in anticipating Liu's use of archaic idioms in the first half of his translation of *Don Quijote* Part I. On the other hand, Model III, which was intended to establish some level of dependence of Yang's translation

on the original, in so far as archaism is concerned, has been discarded as not so reliable. The limited usefulness of Model III seems to suggest that the rendition of archaisms present in *Don Quijote* Part I is a rather challenging issue for any Chinese translator. Chinese translators may well be perfectly aware of the importance of archaisms as a means of conveying the full flavour of the original, but their knowledge and understanding of the source text may operate at a merely conceptual level, not sufficiently clear perhaps to be able to perform the desired linguistic shift competently and accurately.

Going back now to the question of how to explain the success achieved by the first two models in predicting Liu's use of archaic idioms in translating *Don Quijote* Part I, while it is reasonable to suppose that Liu has largely followed Yang's example in the use of Chinese archaic idioms in the first half of his own translation, it is rather puzzling to notice that he has significantly increased his use of archaisms in the second half of the novel. I believe stylistic variation is not a stylistic feature unique or exclusive to Liu's work, but rather a typical feature in literary translation where the creative urge on the part of the translator from within the target context is not merely an option, but more of a necessity in view of the linguistic complexity and peculiarity of archaisms.

8.4 Construction of models to explore stylistic variations in literary translation

If we consider the textual investigation carried out so far to be simply descriptive, the kind of quantitative analysis that follows in chapters VI and VII was intended to look more closely into stylistic variation in literary translation; that is to say, it is more experimental and inductive. As Sinclair has observed, there is an imminent need to look for testable hypotheses or working models in literary stylistics that may prod the text into revealing itself to us. In this part of the dissertation, I begin by trying to explain stylistic variation in Liu's use of archaism with the help of the context motivated theory which I have adapted from Biber's famous sociolinguistic study on register variation. As mentioned before, in my thesis, register and style are treated as two largely exchangeable concepts which broadly refer to the linguistic variation detected in actual textual contexts at a morphological, lexical, phraseological or syntactic level.

Then through various statistical procedures I have been able to set up an analytical instrument with two dimensions from a variety of source text contextual factors, i.e. speech environment and speech content. This has been made possible by the use of dimension reduction techniques, i.e. principal component analysis, which can extract a small but highly efficient group of variables from a large amount of information. The analytical instrument thus constructed can then be used to anticipate the occurrence of archaic idioms in Liu's translation through the computation and modelling of contextual data collected from the source text.

Going further into the intriguing triangular relationship between the three texts as explained earlier on, the first thing to notice is that, in certain parts of his translation Liu has largely followed the phraseological patterns characterizing Yang's use of Chinese archaic idioms. Secondly, it is interesting

to see that while there is strong link between the two target texts, Liu's link with the original can be corroborated statistically, while Yang's version turns out to be more distant from the source text as regards the use of archaism. Finally, one could suggest that Liu's approach to the original text, vis-à-vis Yang's, grows to a more independent and assertive attitude than Yang's by the time he came to translate the second half of *Don Quijote* Part 1 (see Chapter VI).

It is my contention therefore that there is a direct relationship between the contextual circumstances of Cervantes's text and the translator's particular use of archaisms, as he sees them, in his own version, -- pointing inevitably to a tendency or inclination on Liu's part to use this particular feature of the language quite deliberately. But there still remains the question of how the writer/translator, Liu in this case, deliberately or otherwise, chooses to change or modify his style in the course of this work. Or perhaps the supposed style variation acting upon this particular text is simply the expression of a general desire to stamp his own stylistic mark on the basis of his own general impression of the original text. Liu would have been aware that Chinese readers, generally, would value highly a translation that fed seamlessly into the linguistic habits of the target language, consistent within the socio-cultural environment at a given time. The huge commercial success of his version of *Don Quijote* in the last decade or so would seem to vindicate Liu's choice of translating style, well in keeping with the expectations of his readers and the nature and character of Mandarin Chinese at a given moment in time. Such an assumption regarding the stylistic behaviour at an individual level as somehow conditioned by a broader socio-cultural background and the general expectation of foreign literary translation in contemporary China would be very interesting topics for future research.

8.5 How the methods and procedures developed in my thesis will help with future research

In conclusion, the main contribution of this thesis resides, I feel, in the way I have managed to conduct and test theoretical models for the study of literary translation. It is the first work of its kind in introducing and experimenting with essential quantitative methods in corpus-based translation studies, especially in the field of Hispano-Chinese literary translation. Despite the growing interest in the use of corpus material and methodologies in translation research (Baker, 1995, 2000, 2004; Olohan, 2004), there is an obvious lack of systematic description of the various statistical tests that may be adapted from corpus linguistic or textual statistics for the purpose of translation research. As a result, many studies pursued in line with corpus-based translation studies are largely tentative in terms of the deployment of quantitative methods in the processing of linguistic data retrieved from translational corpora. It is generally felt that the lack in the development of quantitative analytical methodologies in corpus-based translation studies has been severely limiting the construction and testing of theoretical models for literary translations, which in turn has been in the way of further expansion of the field as a whole. This thesis has made the first attempt at testing the feasibility and productivity of basic statistical methods with data collected from corpora of translated texts. It may therefore be of great help and interest to research students or scholars interested in exploring the quantitative profile of literary

translations which may in turn provide us with valuable insights into the nature of literary translation as a form of creative writing in its own right.

The linguistic data used for the purpose of explanation have been extracted from the first parallel Castilian-Chinese corpus of *Don Quijote* (Part I) and the use of statistical tests has proved to be instrumental in the whole process of textual analysis: (1) they have helped distinguish the differences in the phraseological patterns underlying the two translations; (2) they have been instrumental in revealing the intriguing triangular relationship between the three texts in terms of the influence of one text upon another. All these findings are not easily perceptible to human observation and the insights offered by quantitative analysis have helped bring our understanding of literary translation to a higher level of theoretical investigation. Another integral part of the thesis is the construction of contextual models for stylistic variation identified in literary translation, which again has been made possible through the exploration of statistical procedures adapted from corpus linguistics and textual statistics. The proposed context-motivated theory, which is built upon Biber's multidimensional model for register variation, has successfully allowed me to enter into the contextual dimension of stylistic variation as a complex textual phenomenon reflecting the translators' manipulation of language. In sum, this thesis, I sincerely believe, the first of its kind, certainly with regards to the combination of Spanish and Chinese. It will provoke much debate amongst corpus translation scholars regarding the methodological direction that Translation Studies as a discrete academic discipline should take for the sake of its future development.

References

- Allen, J. (1979) "Traduttori traditori: Don Quijote in English", in *Crítica Hispánica*, 1: 1-3
- Andrist, D. (1983) "Male versus female friendship in Don Quijote", in *Cervantes: Bulletin of the Cervantes Society of America*, 3.2: 149-59
- Baker, M. (1993) "Corpus linguistics and translation studies: implications and applications", in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: in Honour of John Sinclair*, Amsterdam: Benjamins, pp. 233-352
- Baker, M. (1995) "Corpora in Translation Studies: an overall view and some suggestions for future research", in *Target*, no. 7, pp. 223-43
- Baker, M. (1999) "The role of corpora in investigating the linguistic behaviour of professional translators", in *International Journal of Corpus Linguistics*, 4 (2): 281-298
- Baker, M. (2000) "Towards a methodology for investigating the style of a literary translator", in *Target*, Volume 12, Number 2, 1, pp. 241-266
- Baker, M. (2004) "The Treatment of Variation in Corpus-based Translation Studies", in Karin Aijmer and Hilde Hasselgård (eds) *Translation and Corpora* (Göteborg Studies in English), Göteborg University, Sweden, pp. 7-17
- Banchs, R. et al (2006) "Chinese Spanish statistical machine translation experiments", in *IV Jornadas de Tecnología del Habla*, November, pp. 55-7
- Bannard, C. (2007) "A measure of syntactic flexibility for automatically identifying multiword expressions in corpora", in Rayson et al. (eds.) *Special Issue of the International Journal of Language Resources and Evaluation on Multiword Expressions*, pp. 1-8
- Bara, W. (1952) "The Refranes of Marqués de Santillana", in *The Modern Language Journal*, Vol. 36, No.3, pp. 140-141
- Barlow, M. (1995a) *A Guide to ParaConc*, Houston: Athelston
- Barlow, M. (1995b) *ParaConc: Multilingual Concordance Software for Multilingual Parallel Corpora* at <http://www.athel.com/parapaper.html>
- Baroni, M. & Bernardini, S. (2006) "A new approach to the study of translationese: machine-learning the difference between original and translated text", in *Literary and Linguistic Computing*, vol.21, no.3, pp. 259-274
- Bell, A. (1984) "Language style as audience design", in *Language in Society* 13(2): 145-204
- Bell, A. (1997) "Language style as audience design", in N. Coupland and A. Jaworski (eds.) *Sociolinguistics: A reader*, pp. 240-50
- Bell, A. (2001) "Back in style: reworking audience design", in: P. Eckert and J.R. Rickford (eds.) *Style and Sociolinguistic Variation*, Cambridge University Press, pp. 139-169.
- Biber, D. & Finegan, E. (1994) (eds.) *Sociolinguistic Perspectives on Register*, Oxford University Press
- Biber, D. (1995) *Dimensions of Register Variation: A Cross-Linguistic Comparison*, Cambridge: Cambridge University Press
- Boase-Beier, J. (2006) *Stylistic Approaches to Translation*, in Hermans (ed.) *Translation Approaches Explored*, Manchester: St. Jerome

- Brown, P. F., et al. (1991) "Aligning sentences in parallel corpora", pp. 169-76, in *Proceedings of the Annual Meeting of the Association of Computational Linguistics*
- Bryman, A. and Cramer, D. (2005) *Quantitative data analysis with SPSS 12 and 13: a Guide for Social Scientists*, London: Routledge
- Cárcer, E. D. and De Sobies (1916) *Las frases del Quijote*, Barcelona: Subirana, Editor y Librero
- Cervantes, M. S. (1952) *Don Quijote I*, F. Marín (ed.) Madrid: Clásicos Castillanos
- Cervantes, M. S. (1974) *Don Quijote Part I*, E. García (ed.), La Habana: La Editorial Edaf
- Cervantes, M. S. (1977) *Don Quijote Part I*, J. Avalor-Arce (ed.) Madrid: Alhambra
- Cervantes, M. S. (1998) *Don Quijote I*, F. Rico et al. (eds.) Barcelona: Instituto de Cervantes
- Chang, C. and Chen, C. (1993) "HMM-based part-of-speech tagging for Chinese corpora", pp. 40-7, in *Proceedings of the Workshop on Very Large Corpora (WVLC-1)*, Ohio State University
- Chen, G. (2002) "Don Quijote and the twentieth-century Chinese literature", in *Foreign Literature Studies*, No.3, Beijing, p. 123-6
- Chen, J. (2003) "Investigation on Twentieth-Century Chinese Rhetoric", in *Journal of Bi Jie Teachers College*, vol.21, no.1, Wu Xi: Bi Jie Teachers College
- Cheshire, J. & Bell, A. (2003) "Register and style", in *International Encyclopedia of Linguistics*, vol.3, W. Frawley (ed.) Oxford: Oxford University Press
- Culpeper, J., Louw, B. and Hoover, D. (2008) *Approaches to Corpus Stylistics*, London: Routledge
- Dawson, B. & Trapp, G. (2004) (4th edition) *Basic and Clinical Statistics*, New York: Lange Medical Books-McGraw Hill
- Debra, D. (1983) "Male and Female Friendship in *Don Quijote*", in *Cervantes: Bulletin of the Cervantes Society of America*, vol. 3, no.2, pp. 149-159
- Deng, Y., et al. (2006) "Segmentation and alignment of parallel text for statistical machine translation", pp. 1-26, in *Natural Language Processing*, Cambridge University Press
- Di Salvo, J. (1989) "Spanish guides to princes and political theories in *Don Quijote*", in *Cervantes: Bulletin of the Cervantes Society of America*, vol. 9 (2), pp. 43-60
- Dirk, S. (2001) "Multi-word discourse markers in translation: a corpus-based investigation into restrictors", in *Lebende Sprachen*, no.3, pp. 97-107
- Dudley, E. (1979) "Boccaccio and Cervantes: Novella as Novella", in *Hispano-Italic Studies*, vol. 1, no.3, pp. 23-40
- Eckert, P. & Rickford, J.R. (2001) *Style and Sociolinguistic Variation*, Cambridge University Press
- El Saffar, R. (1981) "On beyond conflict", in *Cervantes: Bulletin of the Cervantes Society of America*, 3.2, pp. 149-59
- Eisenberg, D. (1983) "On editing *Don Quijote*", in *Cervantes: Bulletin of the Cervantes Society of America* 3.1, pp. 3-34
- Fajardo, J. (1994) "Closure in *Don Quixote I*: A Reader's Canon's", in *Cervantes: Bulletin of the Cervantes Society of America*, 14.1, pp. 41-60
- Farwell, D., Helmreich, S. and Casper, M. (1995) "SPOST: a Spanish part-of-speech tagger", in *Procesamiento del Lenguaje Natural*, vol.17, pp. 42-56
- Feng, Z. (2001) "Hybrid approach for automatic segmentation and annotation of Chinese text corpus",

- in *International Journal of Corpus Linguistics*, vol.6, pp. 35-42
- Field, A. (2005) *Discovering Statistics Using SPSS* (2nd edition), London and Thousand Oaks: Sage Publication
- Forbes, S. (1978) *Dynamism in the English Translations of Don Quixote: a Stylistic Analysis and Evaluation*, Ph. D dissertation, the Catholic University of America, p.925
- Foulché-Delbosc, R. (1901) “El sastre del cantillo”, in *Revue Hispanique*, VIII, pp. 332-7
- Fu, Q. et al (2005) “A hidden Markov model based segmentation and recognition algorithms for Chinese handwritten address character strings”, pp. 590-4, in *Proceedings of the 8th International Conference on Document Analysis and Recognition*, IEEE Computer Society
- Frantzi, K. (2008) “Computing stylistics and corpus linguistics for author attribution”, in *International Journal of Interdisciplinary Social Sciences*, vol.1, issue. 3, pp. 69-78
- Gale, W.A. and Church, K.W. (1991) “A program for aligning sentences in bilingual corpora”, pp. 177-84, in the *Proceedings of the Annual Meeting of the Association of Computational Linguistics*
- Gao, J. F. and Li, M. (2005) Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach, in *Computational Linguistics*, 4, vol.31, MIT: Cambridge, USA pp. 531-574
- Garside, R. and Smith, N (1997) “A hybrid grammatical tagger: CLAWS4”, pp. 102-21, in R Garside, G. Leech and A.McEnery (eds.) *Corpus Annotation*, London: Longman
- Garson, G. (2008) “Reliability Analysis”, in *Statnotes: Topics in Multivariate Analysis*, retrieved on 24 May, 2008 from <http://www2.chass.ncsu.edu/garson/pa765/statnote.htm>
- Gilman, S. (1989) *The Novel according to Cervantes*, Berkely: University of California Press
- Goyal, S. and Chateerjee, N. (2006) “Parsing aligned parallel corpus by projecting syntactic relations from annotated source corpus”, pp. 301-8, in *Proceedings of the COLING/ACL Meeting*, Sydney
- Hatcher, L. (1994) *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*, North Carolina: SAS Publishing
- Hoover, D. (2001) Statistical Stylistics and Authorship Attribution: an Empirical Investigation, in *Literary and Linguist Computing*, vol.16, pp.421 – 444
- Hoover, D. (2003a) “Frequent collocations and Authorial Style”, in *Literary and Linguistic Computing*, 18 (3), pp. 261-86
- Hoover, D.L. (2003b) “Multivariant analysis and the study of style variation”, in *Literary and Linguistic Computing*, 18 (4): 341-360, Oxford University Press
- Hoover, D., Louw, B. and Culpeper, J. (2008) *Approaches to Corpus Stylistics: the Corpus, the Computer and the Study of Literature*, in T. McEnery and M. Hoey (eds.) *Routledge Advances in Corpus Linguistics*, London: Routledge
- Hu, Z. (2004) “Interview with Mrs. Yang on translating Don Quijote”, in *Digest of Science and Technology*, vol. 6 pp. 41-55
- Hunston, S. (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press
- Jarman, B. G. (2003) *Gran Diccionario Español (Español-Inglés; Inglés-Español)*, London: Oxford University Press (3rd Edition)
- Ji, M. (2007a) “Corpus encoding and corpus-based textual analysis”, in *International Journal of Translation Studies*, vol.19, no.1, pp. 93-109, New Delhi: Bahri Publications

- Ji, M. (2007b) "Phraseology in translation studies", in Lewandowska-Tomaszczyk (ed.) *Linguistics, Computer Tools and Applications: State of the Art*, Frankfurt and Main: Peter Lang AG
- Ji, M. (2007c) "What is the starting point? In search of a working definition of Chinese idioms", in *African and Asian Studies*, vol.6: 1-2, Leiden, Netherlands: Brill, pp. 1-11
- Ji, M. (2008a) "Problem-oriented annotation in parallel corpus exploration", paper presented at the *Annual Meeting of the Special Interest Group of Corpus Linguistics of the British Association of Applied Linguistics*, Edinburgh
- Ji, M. (2008b) "A corpus-based study of structural variants of Chinese idioms in naturally-occurring contexts", in *Journal of Chinese Language and Computing: an International Journal of Chinese and Oriental Languages Information Processing Society*, Singapore, 17 (2): 67-82
- Ji, M. (2008c) "A hybrid statistical-linguistic model of style shifting in literary translation", pp. 10-13, in Kibble & Rauchs (eds.) *Style in Text — Creative Generation and Identification of Authorship*, symposium held in conjunction with the Society for the Study of Artificial Intelligence and the Simulation of Behaviour Annual Meeting, Aberdeen, Scotland
- Ji, M. (2008d) "From frequency counts to statistical prediction: expanding quantitative research in corpus-based translation studies", paper presented at the *Annual Conference of British Association of Applied Linguistics: Taking the Measure of Applied Linguistics*, Swansea, Wales
- Ji, M. (2008e) "Phraseology and translator's style: from a diachronic perspective", paper presented at the *Annual Meeting of the Poetics and Linguistics Association*, Sheffield, UK
- Johnson, K. (2008) *Quantitative Methods in Linguistics*, London and New York: Blackwell
- Kaiser, H.F. (1960) "The Application of Electronic Computers to Factor Analysis", in *Educational and Psychological Measurement*, vol. 20, pp. 141-151
- Kenny, D. (2000) "Lexical Hide-and-Seek: looking for creativity in a parallel corpus", pp. 93-104, in M. Olohan (ed.) *Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects*, Manchester: St. Jerome
- Kenny, D. (2001) *Lexis and Creativity in Translation: A Corpus-Based Study*, Manchester: St. Jerome
- Labov, W. (1972) *Sociolinguistic Patterns*, Philadelphia: University of Pennsylvania Press
- Lakoff, G. (1993) "The contemporary theory of metaphor", in A Ortony (ed.) *Metaphor and Thought* (2nd edition) Cambridge: Cambridge University Press, pp. 205-251
- Lakoff, G. (1995) "Metaphor, morality and politics or, why conservatives have left liberals in the dust", in *Social Research*, 62: 2, pp. 177-213
- Laviosa, S. (1997a) "How comparable can comparable corpora be?", in *Target*, vol.9, pp. 289-319
- Laviosa, S. (1997b) "Investigating Simplification in an English Comparable Corpus of Newspaper Articles", 531-540, in K. Klaudy and J. Kohn (eds.) *Transferring Necessity Est. Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting*, Budapest, Hungary. Budapest: Scholastica
- Laviosa, S. (1998) "The corpus-based approach: a new paradigm in translation studies", in *Meta* 43 (4), pp. 474-9
- Leech, G., Francis, B. and Xu, X. (1994) "The use of computer corpora in the textual demonstrability of gradience in linguistic categories", in Fuchs, C. and Vitorri, B. (eds.) *Continuity in Linguistic*

- Semantics*, Amsterdam and Philadelphia: John Benjamins, pp. 57-76
- Leech, G., McEnery, T. and Wynne, M. (1997) "Further levels of annotation", pp. 85-101, in R. Garside, Leech, G and A. McEnery (eds.) *Corpus Annotation*, London: Longman
- Leech, G. and Weisser, M. (2003) "Generic speech act annotation for task-oriented dialogues", in D.Archer, et al. (eds.) *Proceedings of Corpus Linguistics 2003*, Lancaster University, pp. 441-6
- Leuven-Zwart, M.van (1989) "Translation and Original: Similarities and Dissimilarities", in *Target*, 1:2, pp.151-181, Amsterdam: John Benjamin
- Liu, J. (1995) *Don Quijote de La Mancha* (Tang Ji He Te), Li River Publisher
- Liu, J. (2000) *Don Quijote de La Mancha* (Tang Ji He Te), Yi Li People's Publisher
- Liu, J. (2001) *Don Quijote de La Mancha* (Tang Ji He Te), Beijing Yan Mountain Publisher
- Liu, J. (2002) *Don Quijote de La Mancha* (Tang Ji He Te), Farland Publisher
- Liu, J. (2003) *Don Quijote de La Mancha* (Tang Ji He Te), Yan Bian Publisher
- Liu, J. (2005) *Don Quijote de La Mancha* (Tang Ji He Te), China Opera Publisher
- Liu, J. (2007) *Don Quijote de La Mancha* (Tang Ji He Te), China Bookshop Publisher
- Mac Cormac, E. (1985) *A Cognitive Theory of Metaphor*, Cambridge: MIT Press
- Malmkjær, K. (1998) "Love thy neighbour: will parallel corpora endear linguists to translators?" in *Meta*, 43/4, pp. 534-541
- Mancing, H. (1982) *The Chivalric World of Don Quijote. Style, Structure, and Narrative Technique*, Missouri: University of Missouri Press
- Mariscal, G. (1994) "Woman and Other Metaphors in Cervantes's 'Comedia famosa de la entretenida' ", in *Theatre Journal*, Vol. 46, No. 2, pp. 213-230
- Martín, L. (1991) *Cervantes and the Burlesque Sonnet*, Berkeley and Los Angeles: University of California Press
- McEnery, A. & Xiao, Z. (2003) "Aspect Marking in English and Chinese: Using the Lancaster Corpus of Mandarin Chinese for Contrastive Language Study", in *Literary and Linguistic Computing*, vol.18, no.4, pp. 361-78
- McEnery, A. and Xiao, Z. (2002) "Domains, text types, aspect marking and English-Chinese translation", in *Languages in Contrast*, vol.2, no.2, pp. 211-29
- McEnery, A., Wilson, A. et al. (1997) "Multilingual resources for European languages: contributions to the Crater project", in *Literary and Linguistic Computing*, vol. 12, no.4, pp. 219-26
- McEnery, T & Wilson, A (1993) "Corpora and Translation: Uses and Future Prospects", in <http://ucrel.lancs.ac.uk/papers/techpaper/vol2.pdf>
- McIntyre, D. (2006) *Point of View in Plays: a Cognitive Stylistic Approach to Viewpoint in Drama and Other Text-Types*, Amsterdam: John Benjamins
- McIntyre, D. et al. (2003) "The construction of a corpus to investigate the presentation of speech, thought and writing in written and spoken British English", pp. 513-22, in D. Archer, et al. (eds.) *Proceedings of Corpus Linguistics 2003*, Lancaster University
- Mikhailov, M. and Villikka, M. (2001) "Is there such a thing as a translator's style?" pp. 378-385, in *Proceedings of Corpus Linguistics 2001*, Birmingham
- Milic, L. (1991) "Progress in stylistics: theory, statistics, computers", in *Computers and Humanities*,

- vol. 25, pp. 393-400
- Montoto, S. and Rautenstrauch, L. (1911) *Personajes, personas y personillas que corren por las tierras de ambas Castillas*, Sevilla, 3 vols
- Moon, R. (1998) *Fixed Expressions and Idioms in English: a Corpus-based Approach*, (Oxford Studies in Lexicography and Lexicology), Oxford University Press
- Moon, R. (2005) "Corpus analysis of idioms", in K. Brown et al. (eds.) *Encyclopedia of Language and Linguistics*, Elsevier
- Moon, R. (2007) "Phraseology in general monolingual dictionaries", in N. Norrick et al. (eds.) *International Handbook of Phraseology*, Berlin and New York: Mouton de Gruyter
- Munday, J. (1998) 'A Computer-assisted Approach to the Analysis of Translations Shifts', pp. 542-556 in, L'approche basée sur le corpus/ The Corpus-based approach, special issue of *Meta*, XLIII (4), guest edited by S. Laviosa
- Nuccorini, S. (ed.) (2002) *Phrases and Phraseology: Data and Description*, New York and Oxford: Peter Lang AG
- Nunberg, G., I. Sag, & Wasow T. (1994) "Idioms", in *Language* 70/3, pp. 491-538
- Nunnally, J. (1978) *Psychometric Theory*, New York: McGraw-Hill Publication
- Oakes, M. (1998) *Statistics for Corpus Linguistics*, Edinburgh University Press
- Olohan, M. (2004) *Introducing Corpora in Translation Studies*, London: Routledge
- Omsby, J. (1885) *Don Quijote*, available at the Project Gutenberg <http://www.gutenberg.org/etext/996>
- Ooi, Y.B. (1998) *Computer Corpus Lexicography*, Edinburgh: Edinburgh University Press.
- Oppenheim, R. (1988) "The mathematical analysis of style: A correlation-based approach", in *Computers and Humanities*, vol. 22, no.4, pp. 241-252
- Pápai, V. (2004) "Explicitation, a universal of translated texts?" pp. 143-65, in Mauranen & Kujamäki (eds.) *Translation Universals: Do They Exist?* Manchester: John Benjamins
- People's Daily*, March 1, 1999
- Piao, S. (2002) "Word alignment in English-Chinese Parallel Corpora", in *Literary and Linguistic Computing*, 17 (2), Oxford University Press, pp. 207-30
- Power, J. (1967) *The Voices of Don Quixote: a Study of Style through Translation*, Ph. D dissertation, University of Wisconsin
- Preacher, K. J. (2001) *Calculation for the Chi-Square Test* [Computer software]. Available from <http://www.quantpsy.org>
- Puurtinen, T. (2004) "Explicitation of clausal relations: a corpus-based analysis of clause connectives in translated and non-translated Finnish children's literature", in Mauranen & Kujamäki (eds.) *Translation Universals: Do They Exist?* Manchester: John Benjamins, pp. 165-76
- Raffel, B. (1996) *Don Quijote*, New York: W W Norton & Co Ltd
- Raffel, B. (1993) "Translating Cervantes: *Una vez más*", in *Cervantes: Bulletin of the Cervantes Society of America*, Vol. 13, No.1, pp. 5-30
- Reynaldo, J. and Santos, A. (1999) "Cronbach's Alpha, a tool for assessing the reliability of scales", in *Journal of Extension*, vol.37, no.2, URL: <http://www.joe.org/joe/1999april/tt3.html>
- Rickford, J. and McNair-Knox, F. (1994) "Addressee- and topic-influenced style shift: A quantitative

- sociolinguistic study", in D. Biber and E. Finegan (eds.) *Sociolinguistic Perspectives on Register*, Oxford: Oxford University Press
- Riding, R. (2000) "Cognitive Style: A Strategic Approach for Advancement", in *International Perspectives on Individual Differences: Cognitive Styles*, vol.1, Stamford: Greenwood Publishing
- Saldanha, G (2005) *Style of Translation: an Exploration of Stylistic Patterns in the Translations of Margaret Jull Costa and Peter Bush*, unpublished PhD dissertation, Dublin City University
- Saldanha, G. (2007) "The use of italics as stylistic devices marking information focus in English translations", paper presented at the *Corpus Linguistics Conference 2007*, Birmingham University
- Savova, L. (2005). "Register", in *New Encyclopedia of Linguistics*, volume 2, P. Strazny (ed.) New York: Taylor & Francis, pp. 898-899
- Semino, E. and Culpeper, J. (2002) *Cognitive Stylistic: Language and Cognition in Textual Analysis*, London & Manchester: John Benjamins
- Semino, E. and Short, M. (2004) *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*, London: Routledge
- Shi, S. (1979) *Studies on Chinese Idioms*, Si Chuan People's Publisher
- Sinclair, J. (2002) "Phraseogonomy", in S. Nuccorini (ed.) *Phrases and Phraseology: Data and Descriptions*, New York and Oxford: Peter Lang
- Sinclair, J. (2004) *Trust the Text: Language, Discourse, Corpus and Discourse*, London: Routledge
- Storrer, A. and Schwall, U. (1995) "Description and acquisition of multiword lexemes", pp. 35-50, in P. Steffens (ed.) *Machine Translation and the Lexicon*, Berlin and Heidelberg: Springer
- Stubbs, M. (1995) "Collocations and semantic profiles: on the cause of the trouble", in *Functions of Language*, 2/1, pp. 23-55
- Stubbs, M. (1996) *Text and Corpus Analysis*, Oxford: Blackwell
- Stubbs, M. (2001) *Words and Phrases, Corpus Studies of Lexical Semantics*, Oxford: Blackwell
- Stubbs, M. (2002) "Two quantitative methods of studying phraseology in English", in *International Journal of Corpus Linguistics*, 7/2, pp. 215-44
- Tao, Y. (2002) "Studies on the types and explanatory modes of Chinese four-character expressions", in *Academic Investigations*, no. 9, pp. 130-7
- Teubert, W. (1996) "Corpora or parallel corpora?" in *International Journal of Lexicography*, vol.9, no.3, pp. 238-64, Oxford: Oxford University Press
- Thomas, J. & Short, M. (eds.) (1996) *Using Corpora for Language Research*, London: Longman
- Thomas, J. and Wilson, A. (1996) "Methodologies for studying a corpus of doctor-patient interaction", pp. 92-109, in J. Thomas and M. Short (eds.) *Using Corpora for Language Research*, London: Longman
- Véronis, J. (2000) *Parallel Texts Processing: Alignment and Use of Translation Corpora*, Kluwer Academic Publishers
- Williamson, E. (1982) "Romance and realism in the interpolated stories of the Quixote", in *Cervantes: Bulletin of the Cervantes Society of America*, 2:1, pp. 43-67
- Wong, P. & Chan, C (1996) "Chinese segmentation based on maximum matching and word binding force", in *Proceedings of Computational Linguistics 96*, pp. 200-3, also available online at

- <http://citeseer.ist.psu.edu/wong96chinese.html>
- Wu, Z. & White, J. D (1990) "Computer Processing of Chinese characters: an overview of two decades' research and development", in *Information Processing & Management*, vol. 26, pp. 681-692, Elsevier Science
- Wynne, M. (2005) "Stylistics: Corpus Approaches", pp. 59-64, in *Encyclopaedia of Linguistics*, Elsevier Science
- Wynne, M. (2005) *Stylistics: Corpus Approaches*, Oxford University Eprints, available at http://eprints.ouls.ox.ac.uk/archive/00001003/01/Corpora_and_stylistics.pdf
- Wynne, M. (2008) "Corpus Approaches to the Language of Literature", paper presented at *the International Conference of Corpus Linguistics in Practical Applications*, Viterbo, Italy
- Xaira or its older version SARA may be downloaded from <http://www.natcorp.ox.ac.uk> or <http://www.xaira.org>
- Xiang, G. (1979) "Relationships between Chinese Idioms, Natural Environment, Cultural Traditions, and Linguistic Characteristics", in *Chinese Language*, vol. 2, pp. 112-121
- Xu, Sheng H. (2006) "Proximity and complementation: studies on the formation mechanism of Chinese four-character idioms from cognitive point of view", in *Journal of Si Chuan International Studies University*, vol.22, no.2, pp. 107-111
- Yang, J. (1978) *Don Quijote de La Mancha* (Tang Ji He De), Beijing: People's Literature Publisher
- Zelle, J. and Ng, H. T. (1997) "Corpus-based approaches to semantic interpretation in NLP", in *Artificial Intelligence Magazine*, Special issue of *Empirical Natural Language Processing*, vol.18, no.4, pp. 45-64
- Zhang, H and Liu, Q (2002) "Model of Chinese words rough segmentation based on N-shortest-paths method", in *Journal of Chinese Information Processing*, vol.16, no.5, pp. 1-7
- Zhang, Ling J. (1987) "Don Quijote is not a satirical or chivalric novel", in *Academic Journal of Luo Yang Institute*, vol. 16, pp. 43-50
- Zhou, Q. & Duan, H.M. (1994) "Segmentation and tagging in Chinese corpus processing", in *Chinese Computing*, no. 21, pp. 85-87
- Zou, T. (2006) "Discussion on Yang's translation of Don Quijote from an ideological perspective", in *Data of Culture and Education*, no. 11, pp. 17-25

Appendix I Random selection of test sample sets from Liu’s work

Research Randomizer Results:

3 Sets of 100 Unique Numbers Per Set

Range: From 1 to 2247 -- Sorted

Set 1			Set 2		Set 3			
8	[[闻名遐迩]]	AI	2	[[肌肉干瘪]]	SP	24	[[想入非非]]	FIG
14	[[受之无愧]]	SS	22	[[通宵达旦]]	AI	94	[[煞有介事]]	AI
18	[[琢磨其意]]	SP	36	[[心地善良]]	SP	105	[[一无所知]]	AI
22	[[通宵达旦]]	AI	52	[[全副武装]]	FIG	193	[[招摇撞骗]]	FIG
44	[[绝代骑士]]	I-FIG	77	[[可想而知]]	AI	199	[[大获全胜]]	AI
58	[[啼声宛转]]	SP	117	[[蹄下生风]]	FIG	282	[[食不近桌]]	SS
73	[[迷惑不解]]	SB	124	[[生死关头]]	FIG	316	[[情同手足]]	FIG
78	[[谈吐文雅]]	I-AI	163	[[气息奄奄]]	FIG	322	[[直抒心怀]]	I-AI
94	[[煞有介事]]	AI	201	[[目瞪口呆]]	SB	338	[[殷殷关切]]	AI
95	[[垂暮之时]]	I-AI	215	[[前所未闻]]	AI	403	[[洋洋自得]]	FIG
128	[[如数照付]]	I-AI	302	[[不知所云]]	AI	418	[[妖魔鬼怪]]	SB
180	[[不可思议]]	AI	303	[[彬彬有礼]]	AI	437	[[明眸黯然]]	I-AI
188	[[稀世珍宝]]	FIG	332	[[深山老林]]	SB	445	[[气喘吁吁]]	MP
237	[[惊魂失魄]]	I-FIG	334	[[若明若暗]]	MP	496	[[醋意大发]]	FIG
284	[[一一照做]]	SS	376	[[艰难险阻]]	SB	502	[[放荡不羁]]	AI
309	[[高谈阔论]]	FIG	379	[[真心实意]]	SB	534	[[青史留名]]	FIG
332	[[深山老林]]	SB	415	[[心灰意冷]]	SB	539	[[迎面而至]]	I-AI
336	[[义无反顾]]	AI	416	[[信以为真]]	SP	541	[[扶弱济贫]]	AI
342	[[指天为誓]]	I-FIG	481	[[荒唐至极]]	SP	560	[[恍然大悟]]	AI
405	[[两厢情愿]]	FIG	486	[[光彩夺目]]	FIG	626	[[若无其事]]	AI
419	[[妖魔鬼怪]]	SB	510	[[满怀信心]]	FIG	668	[[远道而来]]	FIG
429	[[暗无天日]]	FIG	512	[[寿终正寝]]	FIG	676	[[雄姿英发]]	I-FIG
500	[[动弹不得]]	SP	527	[[迫不得已]]	AI	700	[[乳臭未干]]	FIG
539	[[迎面而至]]	I-AI	540	[[一口咬定]]	FIG	706	[[痛痛快快]]	MP
581	[[除暴安良]]	AI	569	[[饥肠辘辘]]	FIG	717	[[闻此狂言]]	SP
606	[[黯然失色]]	I-FIG	585	[[不尽相同]]	I-AI	723	[[镇静自若]]	AI
610	[[远离人烟]]	I-FIG	653	[[五脏六腑]]	FIG	740	[[引以为戒]]	AI
613	[[无可奈何]]	AI	689	[[谈情说爱]]	SSY	771	[[来去匆匆]]	MP
620	[[咬紧牙关]]	FIG	701	[[举棋不定]]	FIG	786	[[赤身裸体]]	SB
632	[[彻夜不安]]	I-AI	703	[[两鬓斑白]]	SP	815	[[风情难捺]]	SP
648	[[闪闪发光]]	MP	705	[[催人纵欲]]	I-AI	865	[[载入史册]]	FIG
692	[[忧心忡忡]]	AI	727	[[知恩图报]]	AI	869	[[忘恩负义]]	AI
733	[[勃然大怒]]	AI	784	[[千谢万谢]]	FIG	887	[[矫揉造作]]	AI
763	[[破破烂烂]]	MP	810	[[山盟海誓]]	FIG	888	[[光明正大]]	FIG
769	[[原封不动]]	FIG	820	[[目瞪口呆]]	SB	937	[[可想而知]]	AI
807	[[急不可耐]]	I-AI	833	[[诸如此类]]	AI	945	[[悠然自得]]	I-AI
871	[[四面环山]]	I-AI	860	[[忍辱负重]]	AI	985	[[雍容华贵]]	SB
872	[[孤峰独立]]	I-AI	878	[[形影相吊]]	FIG	990	[[所作所为]]	MP

890	[[一无所知]]	AI	900	[[异想天开]]	FIG	995	[[胆小怯懦]]	SB
909	[[一时冲动]]	SS	909	[[一时冲动]]	SS	1011	[[胡言乱语]]	SS
912	[[不得要领]]	FIG	953	[[难以自拔]]	I-FIG	1017	[[滔滔不绝]]	FIG
916	[[哭天号地]]	SSY	966	[[背信弃义]]	AI	1058	[[热热闹闹]]	MP
929	[[专心致志]]	AI	973	[[满腹狐疑]]	FIG	1134	[[面黄肌瘦]]	SB
932	[[转眼之间]]	I-FIG	979	[[背信弃义]]	AI	1186	[[早有耳闻]]	SP
933	[[价值连城]]	FIG	986	[[相得益彰]]	AI	1192	[[毫不留情]]	SS
956	[[品貌双全]]	I-AI	991	[[三言两语]]	FIG	1202	[[尽人皆知]]	AI
972	[[纤细白皙]]	SB	1003	[[忘恩负义]]	AI	1206	[[千恩万谢]]	FIG
998	[[有气无力]]	SSY	1041	[[赞叹不已]]	AI	1207	[[天涯海角]]	FIG
1012	[[疯疯癫癫]]	MP	1048	[[奉若神明]]	FIG	1217	[[受封晋爵]]	SSY
1015	[[神志不清]]	SP	1062	[[恰恰相反]]	MP	1238	[[苦心修行]]	I-AI
1063	[[满不在乎]]	AI	1073	[[美丽无比]]	I-FIG	1251	[[百般殷勤]]	I-FIG
1101	[[渺茫黯淡]]	SB	1081	[[泪流满面]]	SP	1281	[[国泰民安]]	AI
1104	[[严辞拒绝]]	I-AI	1138	[[以防万一]]	FIG	1300	[[欢天喜地]]	FIG
1116	[[破衣烂衫]]	SB	1148	[[英勇强悍]]	SB	1305	[[深居简出]]	AI
1130	[[名正言顺]]	AI	1151	[[受苦受难]]	SSY	1328	[[兼而有之]]	AI
1145	[[尽力而为]]	AI	1155	[[远道而来]]	FIG	1364	[[甜言蜜语]]	SB
1183	[[所知甚少]]	I-AI	1171	[[各种各样]]	MP	1375	[[欲言又止]]	AI
1193	[[蒙在鼓里]]	FIG	1295	[[苦口婆心]]	FIG	1414	[[嫉妒之火]]	I-FIG
1212	[[事实确凿]]	SP	1301	[[莫名其妙]]	AI	1433	[[无耻之徒]]	SS
1308	[[堂堂正正]]	MP	1338	[[无缘无故]]	MP	1444	[[不知廉耻]]	I-AI
1312	[[全神贯注]]	AI	1382	[[迷惑不解]]	SB	1448	[[信誓旦旦]]	AI
1328	[[兼而有之]]	AI	1395	[[装聋作哑]]	FIG	1458	[[和盘托出]]	FIG
1374	[[正襟危坐]]	AI	1406	[[胡思乱想]]	SS	1476	[[怒气冲冲]]	MP
1378	[[情意绵绵]]	FIG	1420	[[默默无语]]	AI	1491	[[长吁短叹]]	SSY
1394	[[唉声叹气]]	SSY	1448	[[信誓旦旦]]	AI	1494	[[后悔不迭]]	I-AI
1413	[[信以为真]]	SP	1468	[[慌慌张张]]	MP	1505	[[呆若木鸡]]	FIG
1457	[[事实真相]]	SP	1515	[[以身相许]]	FIG	1513	[[深居闺阁]]	I-AI
1497	[[一言不发]]	SS	1529	[[动弹不得]]	SP	1528	[[命途多舛]]	AI
1503	[[动弹不得]]	SS	1545	[[言辞得体]]	SP	1549	[[不容分说]]	AI
1533	[[结成眷属]]	I-FIG	1610	[[一贫如洗]]	FIG	1582	[[心慈手软]]	SB
1569	[[一本正经]]	FIG	1630	[[大手大脚]]	FIG	1616	[[身不由己]]	AI
1587	[[舞文弄墨]]	FIG	1631	[[挥霍无度]]	AI	1628	[[大手大脚]]	FIG
1648	[[英名永存]]	I-FIG	1690	[[层出不穷]]	FIG	1684	[[痛哭流涕]]	AI
1656	[[精疲力竭]]	AI	1714	[[神不知鬼]]	FIG	1707	[[绝代佳人]]	FIG
1671	[[别有用心]]	AI	1750	[[无可比拟]]	I-AI	1709	[[忐忑不安]]	AI
1741	[[一帆风顺]]	FIG	1758	[[立足之地]]	FIG	1719	[[迫不得已]]	AI
1761	[[世上罕见]]	I-AI	1761	[[世上罕见]]	I-AI	1739	[[泪如泉涌]]	FIG
1765	[[狭窄简陋]]	SB	1763	[[堂堂正正]]	MP	1749	[[心惊胆战]]	FIG
1766	[[莫名其妙]]	AI	1776	[[一举一动]]	MP	1774	[[几经周折]]	I-AI
1838	[[恶语相讥]]	AI	1822	[[胡说八道]]	SS	1799	[[当之无愧]]	AI
1858	[[确凿无疑]]	SB	1838	[[恶语相讥]]	AI	1816	[[胆战心惊]]	FIG
1862	[[弥天大谎]]	FIG	1841	[[无动于衷]]	AI	1822	[[胡说八道]]	SS

1871	[[一口咬定]]	FIG	1880	[[一片狼藉]]	FIG	1856	[[能攻善守]]	SSY
1891	[[勃然大怒]]	AI	1891	[[勃然大怒]]	AI	1897	[[高高兴兴]]	MP
1910	[[扶弱济贫]]	AI	1899	[[牵肠挂肚]]	FIG	1919	[[慌慌张张]]	MP
1930	[[低三下四]]	FIG	1920	[[急急忙忙]]	MP	1920	[[急急忙忙]]	MP
1932	[[一五一十]]	FIG	1936	[[全然不知]]	I-AI	1934	[[有血有肉]]	SSY
1941	[[毛骨悚然]]	FIG	1965	[[乐善好施]]	AI	1953	[[耿耿于怀]]	AI
1946	[[世世代代]]	MP	2059	[[由此可见]]	AI	1962	[[黯然失色]]	FIG
1961	[[默默无闻]]	AI	2064	[[混淆是非]]	I-AI	1963	[[用心险恶]]	I-AI
2008	[[愚昧无知]]	SB	2080	[[秀色可餐]]	FIG	1972	[[美貌绝伦]]	I-FIG
2038	[[是非不分]]	I-AI	2088	[[叹为观止]]	AI	2000	[[乌七八糟]]	SS
2042	[[稀奇古怪]]	SB	2104	[[胡言乱语]]	SS	2057	[[千真万确]]	FIG
2097	[[豪爽大方]]	SB	2126	[[谈情说爱]]	SSY	2073	[[其乐无穷]]	AI
2112	[[洁身自好]]	FIG	2128	[[整装待命]]	FIG	2091	[[香气扑鼻]]	SP
2121	[[目瞪口呆]]	SB	2133	[[情有可原]]	AI	2117	[[又薄又花]]	MP
2126	[[谈情说爱]]	SSY	2153	[[兴高采烈]]	AI	2182	[[公诸于众]]	AI
2131	[[鬼迷心窍]]	FIG	2198	[[物造其类]]	AI	2229	[[风雅文士]]	I-AI
2169	[[一往情深]]	AI	2225	[[饱经世故]]	I-FIG	2234	[[应有尽有]]	AI
2210	[[冥思苦索]]	SSY	2245	[[胡编滥造]]	SSY	2242	[[转忧为笑]]	SS

Appendix II Random selection of test sample sets from Yang's work

Set 1		Set 2		Set 3	
5. [[理亏气短]]	SB	37. [[丰功伟绩]]	AI	18. [[名传千古]]	FIG
29. [[兴高采烈]]	AIbv	41. [[自言自语]]	MP	29. [[兴高采烈]]	AI
35. [[丰功伟绩]]	AI	45. [[一模一样]]	MP	41. [[自言自语]]	MP
51. [[斯文和悦]]	SB	57. [[贼皮贼骨]]	MP	59. [[长夜清醒]]	SP
54. [[哈哈大笑]]	MP	62. [[殷勤周至]]	SB	79. [[清清楚楚]]	MP
55. [[莫名其妙]]	AI	68. [[相貌堂堂]]	MP	91. [[提心吊胆]]	FIG
62. [[殷勤周至]]	SB	85. [[愚蠢粗暴]]	SB	93. [[嘟嘟囔囔]]	MP
76. [[专心一志]]	I-AI	104. [[死心塌地]]	SB	153. [[扭扭捏捏]]	MP
94. [[百战百胜]]	MP	117. [[狂妄自大]]	SB	165. [[离奇有趣]]	SB
115. [[独一无二]]	AI	119. [[勃然大怒]]	AI	200. [[糊涂放肆]]	SB
120. [[怒气冲天]]	I-FIG	120. [[怒气冲天]]	I-FIG	202. [[支离破碎]]	SB
123. [[七死八活]]	SS	152. [[子子孙孙]]	MP	228. [[枝枝节节]]	MP
149. [[兴妖作怪]]	SS	228. [[枝枝节节]]	MP	284. [[几次三番]]	SSY
155. [[实事求是]]	AI	235. [[哆哆嗦嗦]]	MP	298. [[平平静静]]	MP
160. [[趣味无穷]]	I-FIG	282. [[一味偏袒]]	I-AI	302. [[辛勤劳苦]]	SB
164. [[千真万确]]	FIG	328. [[一一奉告]]	MP	335. [[三脚两步]]	SB
169. [[显显身手]]	MP	360. [[自由散漫]]	SB	347. [[唯命是听]]	AI
177. [[天花乱坠]]	FIG	370. [[不堪入耳]]	AI	348. [[不由自主]]	AI
190. [[提神醒睡]]	SSY	388. [[矢志不二]]	AI	358. [[幽娴贞静]]	SB
193. [[妖魔鬼怪]]	SB	459. [[绝世无双]]	SB	359. [[洁身自好]]	FIG
200. [[糊涂放肆]]	SB	473. [[不相上下]]	AI	368. [[痛痛快快]]	MP
215. [[无从寻觅]]	SP	520. [[萧萧马嘶]]	MP	374. [[宛转悲啼]]	SP
235. [[哆哆嗦嗦]]	MP	537. [[天晴风定]]	SB	389. [[咎由自取]]	AI
246. [[千千万万]]	MP	597. [[奇事险遇]]	SB	395. [[平静清澈]]	SB
248. [[诚诚恳恳]]	MP	598. [[丰功伟绩]]	AI	399. [[气喘吁吁]]	MP
259. [[吃饱喝足]]	SSY	601. [[聪明绝顶]]	SP	401. [[有气无力]]	SSY
307. [[殷勤周至]]	SB	615. [[想入非非]]	FIG	437. [[不以为耻]]	SP
314. [[艰苦卓绝]]	SB	617. [[糊里糊涂]]	MP	522. [[咚咚鼓响]]	MP
318. [[忍饥耐渴]]	SSY	627. [[心满意足]]	SB	553. [[打抱不平]]	AI
323. [[喃喃求告]]	MP	646. [[游街示众]]	SSY	561. [[妖魔鬼怪]]	SB
356. [[无情残酷]]	SB	656. [[出言不逊]]	AI	585. [[追根究底]]	FIG
362. [[一清二楚]]	FIG	664. [[无影无踪]]	MP	622. [[单枪匹马]]	FIG
366. [[坚如铁石]]	FIG	736. [[日日夜夜]]	MP	666. [[荒谬绝伦]]	SB
369. [[寸寸断肠]]	FIG	773. [[年轻漂亮]]	SB	698. [[咬牙切齿]]	SSY
371. [[怨慕凄怆]]	SB	775. [[放诞风流]]	SB	714. [[忍无可忍]]	AI
440. [[追根究底]]	FIG	826. [[何德何能]]	MP	717. [[不以为然]]	AI
444. [[千遍万遍]]	FIG	843. [[眼泪汪汪]]	MP	730. [[赤手空拳]]	FIG
453. [[尊贵美丽]]	SB	862. [[心慌意乱]]	SB	744. [[捕风捉影]]	FIG
546. [[妖魔鬼怪]]	SB	877. [[信以为真]]	SP	748. [[十全十美]]	FIG
571. [[心惊胆颤]]	FIG	878. [[失魂落魄]]	SS	763. [[青葱悦目]]	SP
609. [[少见多怪]]	SSY	902. [[口角玲珑]]	FIG	769. [[规规矩矩]]	MP

688. [[钻头觅缝]]	FIG	925. [[回心转意]]	SSY	778. [[美貌贞静]]	SB
699. [[万恶之窝]]	I-FIG	980. [[名不虚传]]	AI	786. [[疯疯癫癫]]	MP
718. [[哑口无言]]	FIG	1007. [[自取灭亡]]	AI	827. [[与世隔绝]]	FIG
719. [[慷慨多情]]	SB	1009. [[身强力壮]]	SB	830. [[头头是道]]	FIG
736. [[日日夜夜]]	MP	1017. [[推三阻四]]	SSY	836. [[野心勃勃]]	FIG
749. [[独一无二]]	AI	1031. [[奇奇怪怪]]	MP	872. [[理直义正]]	SB
751. [[智勇兼备]]	I-AI	1040. [[诺诺连声]]	FIG	889. [[立志高尚]]	SP
753. [[意味深长]]	SP	1042. [[平心静气]]	SB	899. [[又惊又慌]]	MP
759. [[疯疯癫癫]]	MP	1056. [[邪门外道]]	SB	905. [[十全十美]]	FIG
762. [[荒唐无稽]]	AI	1059. [[丰功伟业]]	AI	907. [[应有尽有]]	AI
777. [[古往今来]]	SSY	1117. [[粗心大意]]	SB	917. [[寻欢取乐]]	SSY
808. [[哭哭啼啼]]	MP	1120. [[幽娴贞静]]	SB	921. [[坚贞不移]]	AI
819. [[荒山野地]]	SB	1128. [[不堪入口]]	SS	928. [[没头没脑]]	MP
820. [[乔装打扮]]	FIG	1139. [[倾吐衷情]]	I-AI	938. [[自告奋勇]]	AI
845. [[街坊亲友]]	SB	1157. [[追根究底]]	FIG	950. [[是死是活]]	MP
857. [[真心诚意]]	SB	1160. [[光明正大]]	FIG	961. [[衣不蔽体]]	FIG
869. [[总而言之]]	AI	1171. [[畅所欲言]]	AI	965. [[千真万确]]	FIG
911. [[严严实密]]	MP	1193. [[计上心来]]	FIG	986. [[糊里糊涂]]	MP
943. [[想来想去]]	MP	1209. [[歪歪倒倒]]	MP	998. [[不甘落后]]	SP
951. [[又怕又累]]	MP	1280. [[有始无终]]	SSY	1011. [[攒眉怒目]]	FIG
966. [[明明白白]]	MP	1297. [[金玉良言]]	SB	1017. [[推三阻四]]	SS
1002. [[灵心妙舌]]	SB	1300. [[多福多寿]]	MP	1021. [[胡说乱道]]	SS
1006. [[多高多大]]	MP	1303. [[淌眼抹泪]]	SSY	1094. [[适可而止]]	AI
1046. [[昏头昏脑]]	MP	1315. [[东拼西凑]]	SS	1130. [[敷衍了事]]	AI
1071. [[谈情说爱]]	SSY	1323. [[盖世英雄]]	FIG	1134. [[称心满意]]	SB
1073. [[神魂颠倒]]	FIG	1348. [[各执己见]]	AI	1157. [[追根究底]]	FIG
1117. [[粗心大意]]	SB	1349. [[横行霸道]]	SSY	1170. [[凭空捏造]]	AI
1156. [[一鼻子灰]]	FIG	1353. [[出人头地]]	FIG	1198. [[七上八下]]	FIG
1200. [[胡作非为]]	AI	1358. [[毫无畏惧]]	SS	1201. [[一时一刻]]	MP
1204. [[拖拖拉拉]]	MP	1364. [[识见高明]]	SP	1205. [[肆无忌惮]]	AI
1238. [[追根究底]]	FIG	1370. [[深思熟虑]]	SB	1241. [[怒不可遏]]	AI
1300. [[多福多寿]]	MP	1397. [[出乎意外]]	AI	1256. [[可想而知]]	SP
1343. [[荒野露宿]]	SP	1398. [[心痒难熬]]	FIG	1300. [[多福多寿]]	MP
1346. [[千载难逢]]	FIG	1424. [[心碎肠断]]	I-FIG	1313. [[七凹八凸]]	SS
1353. [[出人头地]]	FIG	1475. [[奄奄一息]]	FIG	1325. [[千稳万妥]]	FIG
1357. [[难分难解]]	MP	1498. [[身分体面]]	SB	1354. [[性命交关]]	I-AI
1363. [[雄心壮志]]	SB	1500. [[疯疯癫癫]]	MP	1358. [[毫无畏惧]]	SS
1386. [[精疲力穷]]	SS	1515. [[近在咫尺]]	FIG	1374. [[欣喜欢乐]]	SB
1433. [[伤心叹气]]	SSY	1527. [[冒冒失失]]	MP	1386. [[精疲力穷]]	SS
1455. [[登山越岭]]	SSY	1559. [[太平盛世]]	FIG	1424. [[心碎肠断]]	I-FIG
1480. [[称心快意]]	I-AI	1573. [[下流无耻]]	SB	1434. [[泰然自若]]	AI
1489. [[百折不挠]]	FIG	1588. [[奇形怪状]]	SB	1435. [[连声长叹]]	SP
1534. [[一口咬定]]	FIG	1616. [[胡说八道]]	SS	1442. [[出乎意外]]	AI
1535. [[莫名其妙]]	AI	1624. [[总而言之]]	AI	1476. [[吉凶甘苦]]	SB

1564. [[称心享福]]	SSY	1632. [[德貌兼备]]	I-AI	1494. [[又惊又喜]]	MP
1568. [[无言可对]]	SP	1637. [[善良忠诚]]	SB	1593. [[恭恭敬敬]]	AI
1579. [[低声下气]]	FIG	1643. [[胡言乱语]]	SS	1598. [[忙忙叨叨]]	MP
1597. [[妖魔鬼怪]]	SB	1651. [[荒谬绝伦]]	SB	1639. [[足智多谋]]	AI
1613. [[连一接二]]	SSY	1656. [[名满天下]]	FIG	1666. [[心安理得]]	AI
1618. [[素不相识]]	AI	1662. [[独一无二]]	AI	1689. [[千真万确]]	FIG
1680. [[信以为真]]	SP	1668. [[死心塌地]]	SB	1694. [[津津有味]]	FIG
1693. [[各种各样]]	MP	1687. [[威震天下]]	FIG	1706. [[一门心思]]	FIG
1709. [[实实在在]]	MP	1691. [[信以为真]]	SP	1735. [[皮开肉绽]]	SB
1723. [[一口咬定]]	FIG	1699. [[巧夺天工]]	FIG	1742. [[面黄肌瘦]]	SB
1725. [[杨花水性]]	FIG	1719. [[顺顺当当]]	MP	1751. [[威力所及]]	SP
1727. [[口是心非]]	AI	1753. [[一往情深]]	AI	1763. [[狰狞可怕]]	SB
1739. [[为非作歹]]	AI	1765. [[大功告成]]	AI	1783. [[束手无策]]	AI
1756. [[芳草芊芊]]	MP	1766. [[物生其类]]	AI	1788. [[千稳万妥]]	SS
1792. [[奇思异想]]	SB	1793. [[不乱不涩]]	MP	1790. [[朴实无文]]	SB

Appendix III Translations by Yang and Liu of figurative source text expressions

Don Quijote, F. Marín (ed.) (1912), Madrid: Real Academia Española, Ediciones de La Lectura

The identification of proverbs in the Spanish texts by using *ParaConc* should be seen as accidental rather than exhaustive or systematic, since it has taken advantage of the fact that in translating or explaining the Spanish proverbs, from time to time the two Chinese translators have chosen to use similar idiomatic expressions in Chinese, which may well be FCEs; or as marked below with double brackets, there are also FCEs highlighted by *ParaConc* which have served as important clues for the retrieval of adjacent texts which are the Chinese translations of the Spanish proverbs in D.Q.

1. villano ruin, que ha de haber lugar siempre para ponerme la mano en la horcajadura, y que todo ha de ser errar vos y perdonaros yo?

(Chapter XXX, **Martín Riquer**, p. 325, locución que significa “tener demasiada familiaridad con una persona”; *horcajadura*, el ángulo que forman las piernas en su nacimiento)

(**Francisco Marín**, p. 120, poner a uno la mano en la horcajadura, frase figurada y familiar que falta en el léxico de la Academia, es como dice Clemencín, “acción propia de quien que coge a otra persona para arrojarla lejos, como pelota o cosa semejante, e indica la superioridad de quien lo ejecuta y el desprecio y vilipendio de quien lo sufre”).)

L: “可恶的乡巴佬，”唐吉珂德过了一会儿又说，“你以为我总让你这么放肆吗？总让你办了错事再饶你吗 (plain explanation)

Y: 你别打错了主意，你这个无法无天的混蛋！你分明就是这么个混蛋，因为你竟敢毁谤天下无双的杜尔西内娅！（FIG）

1. Mas la honesta presencia de Camila, la gravedad de su rostro, la compostura de su persona era tanta, que ponía freno a la lengua de Lotario.

L: 然而卡米拉正襟危坐，表情严肃，举止稳重，使得洛塔里奥欲言又止。（I-AI）

Y: 可是卡蜜拉非常贞静，脸色端庄，举止安详，使罗塔琉不敢轻易开口。（Plain explanation）

2. Escuchóla don Fernando sin replicalle palabra, hasta que ella dio fin a las suyas y principio a tantos sollozos y suspiros, que bien había de ser corazón de bronce el que con muestras de tanto dolor no se enterneciera.

L: 费尔南多一言不发地听多罗特亚说话。多罗特亚说完后不禁哀声饮泣，心肠再硬的人也不会无动于衷。（Plain explanation）

Y: 堂费南铎只是听着，一句话不说。她讲完了又流泪叹气；除非铁石心肠的人，才能冷眼瞧她那悲苦样儿。（FIG）

3. sosiegue vuestra merced el pecho, que si el señor licenciado supiera que por ese invicto brazo habían sido librados los galeotes, él se diera tres puntos en la boca, y aun se mordiera tres veces la lengua, antes que haber dicho palabra que en despecho de vuestra merced redundara.

L: 请您消消气，假如神甫大人知道是您放了那些苦役犯，他就是再忍不住，也会守口如瓶，不至于说出那些有损您尊严的话来的。（FIG）

Y: 您别生气吧。如果硕士先生早知道那队囚犯是您这条天下无敌的胳膊放走的，他宁愿嘴

上缝三针，甚至把舌头咬三下，也决不说出冒犯您的话来。(Literal translation)

4. si ya vuestra merced no gusta que su padre y mi señor la dé al otro mundo, porque no se puede esperar otra cosa de la pena con que queda por vuestra ausencia.

L: 假如你不愿意让你的父亲即我的主人到极乐世界去的话。你的出走给你父亲带来的痛苦已经让他悲痛欲绝了。(FIG)

Y: 除非你愿意把你的爸爸、我们的主人赶出人世；他为你出走，伤心得只有死路一条了。(SP)

5. En tanto que el agradecido y nuevo correo esto me decía, estaba yo colgado de sus palabras, temblándome las piernas de manera que apenas podía sostenerme.

L: “我听这位值得我感激的临时信使说话时，心一直悬着，两腿不住地打哆嗦，几乎要站不住了。(Plain explanation)

Y: 那位满心感激的临时信差在讲话的时候，我全神贯注地听着，两腿索索地抖个不停，几乎连站都站不住。(AI)

6. y fue tanto el asco que tomó, que, revolviéndosele el estómago, vomitó las tripas sobre su mismo señor, y quedaron entrambos como de perlas.

L: 不禁一阵恶心，胃里的东西全翻出来，又吐到了主人身上，弄得两个人都湿漉漉的。(Plain explanation)

Y: 他恶心得很，一阵反胃，把肚里的东西全吐在他主人身上；两个人都淋漓尽致。(AI)

7. y, pues yo fui el fabricante de mi deshonra, no hay para qué...

L: 是我自己制造了我的耻辱，没有理由... (Literal translation)

Y: 我的耻辱是咎由自取，何必... (AI)

8. así los que llaman de naturaleza como los de fortuna, no puedo yo corresponder con agradecimiento que llegue al bien recibido, y sobre al que me hizo en darme a ti por amigo y a Camila por mujer propia: dos prendas que las estimo, si no en el grado que debo, en el que puedo.

L: 人们称我为天生富贵命，我就会感恩不尽吧。其实，我还有你做我的朋友，有卡米拉做我的妻子。这两样宝贝我也十分看重。(Literal translation)

Y: 上帝赏赐的福气，正感激不尽。我有这样的父母，天赋的才和人间的财都不薄；而且锦上添花，还有你做朋友、卡蜜拉做妻子。(FIG)

9. Pero quizá la caballería y los encantos destos nuestros tiempos deben de seguir otro camino que siguieron los antiguos.

L: 不过，也可能是我们这个时代的骑士和魔法都不同以往了。(Plain explanation)

Y: 大概今非昔比，骑士道和魔法都换了样了。(AI)

10. —No podré hacer eso —respondió Sancho—, porque en apartándome de vuestra merced, luego es conmigo el miedo, que me asalta con mil géneros de sobresaltos y visiones.

L: “我不能去，”桑乔说，“我只要离开您就害怕，觉得危机四伏”。(FIG)

Y: 桑丘说：“这可不行，我一离开您就心惊肉跳，见神见鬼的。(MP)

11. —Para conmigo no hay palabras blandas, que ya yo os conozco, fementida canalla —dijo don Quijote.

L: “花言巧语对我不起作用。我认识你们这些卑鄙的家伙。”(FIG)

Y: “我不吃这套花言巧语！我看破你们是撒谎的混蛋！”(FIG)

12. Así que, si tú tienes ceñido mi cuerpo con tus brazos, yo tengo atada mi alma con mis buenos deseos, que son tan diferentes de los tuyos como lo verás si con hacerme fuerza quisieres pasar adelante en ellos.

L: 所以，尽管你已经把我的身子搂在你怀里，我仍然坐怀不乱。如果你想强迫我再走下去，你就会看到你我的想法有多么不同。(FIG)

Y: 你尽管抱住我的身体，我的心却是坚贞不移的。你如果遂着自己的心蛮来硬做，我就会叫你瞧瞧，你的心远不是我的心。(AI)

13. es yo, hermana -replicó Camila-, qué tengo de saber, que no me atreveré a forjar ni sustentar una mentira, si me fuese en ello la vida? Y si es que no hemos de saber dar salida a esto, mejor será decirle la verdad desnuda, que no que nos alcance en mentirosa cuenta.

L: “可是妹妹啊，”卡米拉说，“我怎么知道该怎么说呢？就是要了我的命，我也不会说呀。如果咱们想不出什么办法来，最好还是把事情和盘托出吧，免得不能自圆其说。”(FIG)

Y: 卡蜜拉说：妹妹啊，我怎么会撒谎呢？我即使性命交关，也不敢撒谎，连帮腔都不敢。咱们打不破这重难关，还是和盘托出吧，别说了谎给捉出来(FIG)

14. y, por salir presto de la agonía que le causaba aquella suspensión, le dijo que hacía notorio agravio a su mucha amistad en andar buscando rodeos para decirle sus más encubiertos pensamientos,

L: 洛塔里奥不愿意再绞尽脑汁猜测了，对安塞尔莫说，这样转弯抹角地说自己的内心秘密是对他们之间深厚友谊的公然侮辱。(I-FIG)

Y: 罗塔琉免得空着急，就怪安塞尔莫不推心置腹，这样拐弯抹角，对不起他们深挚的友谊。(FIG)

15. El ferido de punta de ausencia y el llagado de las telas del corazón, dulcísima Dulcinea del Toboso, te envía la salud que él no tiene.

L: 亲爱的托博索的杜尔西内亚，诚致问候。离别的刺痛，心灵的隐伤，已使我心力交瘁。(Literal translation)

Y: 尊贵无比的小姐：一别至今，肝肠寸断。我身不安，心不宁，但愿最甜蜜的杜尔西内娅台尔托波索身心安宁。(FIG)

16. Porque, cuando es tiempo de la siega, se recogen aquí, las fiestas, muchos segadores, y siempre hay algunos que saben leer, el cual coge uno destos libros en las manos, y rodeámonos dél más de treinta, y estámosle escuchando con tanto gusto que nos quita mil canas

L: 他手里拿着一本这样的书，有三十多人围着他。我们都认真地听他念，仿佛觉得自己也年轻了。(Plain explanation)

Y: 中间总有个把识字的，就拿一本来读，我们三十多人都围着他，听得津津有味，简直都返老还童了。(FIG)

17. al cual, si aquí le hallo, y que habla en otra lengua que la suya, no le guardaré respeto alguno, pero, si habla en su idioma, le pondré sobre mi cabeza.

L: 如果我在这儿碰到他，他竟对我讲他母语之外的其他语言，我就对他不客气；他要是讲自己的语言，我就把他奉若上宾。”(I-FIG)

Y: 如果他跑来不说本国话，我对他并不佩服；如果他说本国话，我对他顶礼膜拜④。”(FIG)

18. antes haced, dando la vida a estos papeles, que la tenga siempre la crueldad de Marcela, para que sirva de ejemplo, en los tiempos que están por venir, a los vivientes, para que se aparten y huyan de caer en semejantes despeñaderos;

L: 你还是把这些手稿保留起来，让人们永远记得马塞拉的冷酷吧，把它作为例证，避免活着的人们今后重蹈覆辙。(FIG)

Y: 我劝你倒是留着这些稿子，让后世见到玛赛拉的残酷而有所鉴戒，免得一失足遗恨千古。(FIG)

19. Y sírvale esto que digo de aviso, para que de aquí adelante no me aparte un dedo de su presencia.

L: 我告诉您，从现在起，我要一直守在您身边，寸步不离。(FIG)

Y: 我跟您说开了吧，从今以后，我就是寸步不离地紧跟着您。(FG)

20. Y no me hubieron bien visto cuando, según él dijo después, quedó tan preso de mis amores cuanto lo dieron bien a entender sus demostraciones.

L: “他后来对我说，他还没认清我的模样就已坠入了情网，他后来的所作所为也证明了这一点。(Literal translation)

Y: “据他后来对我说，他一看见我，就颠倒得不由自主。这从他的行为上都看得出来。(AI)

21. con uno dellos comenzó a dar a nuestro don Quijote tantos palos, que, a despecho y pesar de sus armas, le molió como cibera.

L: 虽然唐吉诃德身着甲冑，可还是被打得遍体鳞伤。(FIG)

Y: 堂吉诃德虽然披着一身铠甲，也打得象碾过的麦子一样。(Literal translation)

22. No sólo no me pagó -respondió el muchacho-, pero, así como vuestra merced traspuso del bosque y quedamos solos, me volvió a atar a la misma encina, y me dio de nuevo tantos azotes que quedé hecho un San Bartolomé desollado;

L: 只剩下我们两人时，他就重新把我捆在那棵树上，又打起我来，把我打得遍体鳞伤。

(FIG)

Y: 孩子答道：“不但没付工钱；您一走，树林里只剩了我和他两个，他就重新把我绑在那棵橡树上，又从头把我鞭打一顿，打得我成了揭掉皮的圣巴多罗美。(Literal translation)

23. y él fue diciendo: Esto digo, señor, porque, si al cabo de haber andado caminos y carreras, y pasado malas noches y peores días, ha de venir a coger el fruto de nuestros trabajos el que se está holgando en esta venta,

L: 桑乔又说：“我是说，大人，咱们走大路绕小道，白天黑夜都不得安生，可换来的却是让这些在客店里逍遥自在的人坐享其成。(FIG)

Y: 先生，我这话有个道理。咱们东奔西走，黑夜受罪，白天更吃苦，劳累了一场，如果到头来却让这个客店里作乐的人去享现成，(plain explanation)

24. pues, sin hacer alguno que bueno fuese, ni aun razonable, sin más ni más, antes que Anselmo se levantara, impaciente y ciego de la celosa rabia que las entrañas le roía, muriendo por vengarse de Camila, que en ninguna cosa le había ofendido, se fue a Anselmo y le dijo:

L: 他把自己以前的谨慎置于脑后，没有认真合理地考虑一下，就按捺不住胸中的嫉妒之火，一心要报复卡米拉。(Literal translation)

Y: 头脑全糊涂了，谨慎的考虑都抛开了，尽管卡蜜拉没丝毫对不起他，他却妒火中烧，按捺不住，拼命要对她报复。(FIG)

25. Así, noramala alcanzaré yo el condado que espero, si vuestra merced se anda a pedir cotufas en el golfo. (chapter XXX)

(M.R. p. 324, pedir cotufas en el golfo, pedir cosas imposibles de lograr; cotufa, tubérculo parecido a la chufa)

L: 如果您还心存疑虑，我想当个伯爵也就没什么指望了。(Plain explanation)

Y: 您要往海底捞针去，我一心想封伯爵的希望就完蛋了。(FIG)

26. No, por cierto, ni aun con la mitad, y aun estoy por decir que no llega a su zapato de la que está delante

L: 不比她漂亮，一半都不如。我甚至敢说，比起现在您面前的这位公主来，她简直望尘莫及。(FIG)

Y: 当然不如！连一半儿都比不上！我竟可以说，她给咱们跟前的这一位拾鞋还不配呢！(Literal translation)

27. a muchos hemos visto que, habiendo pasado por estas Sirtes y por estas Scilas y Caribdis,

(Chapter XXXVII, M.R. p. 410, Escila y Caribdis, cabo y torbellino del estrecho de Mesina, de peligrosa navegación, cuya mención constituye un tópico desde los autores clásicos)

(F.M. p. 324, Scilas, como dice la edición príncipe, y no Escilas, como mal leyeron algunos editores, entre ellos Clemencín y Cortejón)

L: 我们看到过很多这种情况，他们含辛茹苦 (FIG)

Y: 咱们看见许多人历尽艰难困苦 (plain explanation)

28. como llevados en vuelo de la favorable fortuna, digo que los hemos visto mandar y gobernar el mundo desde una silla,

L: 一旦达到了目标，就好像插上了时来运转的翅膀，开始坐在椅子上统治世界，(literal

translation)

Y: 到了这一步，就飞黄腾达了；咱们看见他们坐在安乐椅里统治世界，(FIG)

29. No permitas, con dejarme y desampararme, que se hagan y junten corrillos en mi deshonra;

L: 你不要抛弃我，让我成为街头巷尾被人们羞辱的话题。(FIG)

Y: 你别抛弃我，让街头巷尾把我的耻辱当作话柄；(FIG)

30. Así es y así será -dijo don Fernando-; por lo cual debe vuestra merced, señor don Quijote, perdonalle y reducirle al gremio de su gracia, sicut erat in principio, antes que las tales visiones le sacasen de juicio. (chapter XLVI)

L: “所以您，唐吉珂德大人，应该原谅他，与他和好如初，别让那些幻觉使他丧失了理智。” (I-AI)

Y: 所以，堂吉珂德先生啊，您得原谅他，和他言归于好，‘依然如故’^①，别让那些幻象迷糊了他。”（^①原文是拉丁文。）(I-AI)

31. que, puesto que su bondad y valor podía poner freno a toda maldiciente lengua, todavía no quería poner en duda su crédito ni el de su amigo,

L: 虽然他的人品可以让那些恶意的中伤不攻自破，可他还是不想让人们对他自己以及他朋友的信誉产生怀疑。(FIG)

Y: 尽管她的贤德封得住恶毒的口舌，他却不愿意自己和朋友遭人家议论。(Literal translation)

32. abrió las puertas de su recato y te entregó las llaves de su libertad

L: 向你敞开了我贞洁的大门，把我的自由的钥匙交给了你，以身相许(FIG)

Y: 才败坏了操守，把身心交付给你 (literal translation)

33. pero las mías, criadas entre sinabafas y holandas, claro está que sentirán más el dolor desta desgracia.

L: 可是我的背却弱不禁风，这回挨打，自然会疼得很厉害。(FIG)

Y: 我的肩膀却是裹着软布细纱娇养惯的，这番遭了殃，痛得就更厉害。(Literal traslation)

34. Si estas que ahora se usan, así las imaginadas como las de historia, todas o las más son conocidos disparates y cosas que no llevan pies ni cabeza.

L: 现在风靡于世的都是这种戏剧，它们无论出于虚构还是根据历史改编的，都是彻头彻尾的胡编乱造。(AI)

Y: 现在风行的戏，情节无论出于虚构、或有历史根据，几乎全都是没头没尾的胡言乱语(MP)

35. Pero el provecho que las muchas virtudes de Camila hicieron, poniendo silencio en la lengua de Lotario, redundó más en daño de los dos,

L: 卡米拉的端庄举止使洛塔里奥沉默不语，但也给两人带来了不利的后果(I-AI)

Y: 卡蜜拉的美德使罗塔琉箝舌无言，可是这对他们两人却更有害。(I-FIG)

36. y, ardiendo en ira y en honroso celo, de los cuerpos la fuerza ejercitastes, que en propia y sangre ajena colorastes, el mar vecino y arenoso suelo;

L: 你们 义愤填膺，热情满腹，奋力拼搏，驰骋沙场，以自己和他人的鲜血，染红了邻海疆土 (AI)

Y: 你们 燃炽着满腔热忱和义愤，英勇苦战直到精疲力穷，把海水和沙岸染成一片殷红。

(Literal translation)

37. Anselmo le replicó que aquél era su gusto, y que no tenía más que hacer que bajar la cabeza y obedecelle.

L: 安塞尔莫说他愿意这样，她应该做的就是 俯首听命。(FIG)

Y: 安塞尔莫说，他爱这么安排，她 只消依顺就行。(Plain explanation)

38. ha de poner su nombre en el templo de la inmortalidad para que sirva de ejemplo y dechado en los venideros siglos, donde los caballeros andantes vean los pasos que han de seguir, si quisieren llegar a la cumbre y alteza honrosa de las armas.

L: 他们的 英名依然会长存于庙宇，供后人仿效的那种骑士。(Literal translation)

Y: 却是 世世传名、人人效法的模范骑士 (MP)

39. -No sé nada -respondió Sancho-; sólo sé que vendré a ser tan desdichado que, por no hallar esta cabeza, se me ha de deshacer mi condado como la sal en el agua.

L: 桑乔说，“我只知道若是找不到这个脑袋，我就会倒霉透顶，我的伯爵称号就会 化为乌有。” (AI)

Y: 桑丘道：什么也不理会，只知道那个脑袋要是找不到，我那份伯爵的封地就 好比盐着了水全化掉了，我就倒霉透顶了。(Literal translation)

40. y tendidos sobre la verde yerba, con la salsa de su hambre, almorzaron, comieron, merendaron y cenaron a un mismo punto,

L: 饥不择食地把早饭、午饭、点心和晚饭合成一顿 (AI)

Y: 他们 饿得胃口正好，就躺在草地上把早饭、午饭、点心、晚饭都并作一顿吃 (plain explanation)

41. y adonde no me da pena alguna el pensar si fue afrenta o no lo de los estacazos, como me la da el dolor de los golpes, que me han de quedar tan impresos en la memoria como en las espaldas.

L: 我伤心的倒不是这顿棒打算不算羞辱，而是肩上背上被打的疼痛劲儿，那真是 刻骨铭心啊。(FIG)

Y: 至于挨了这顿桩子算不算侮辱，我是满不在乎的；苦的是给揍得疼痛，肩膀上、心眼里都痛得撒不开。(Literal translation)

42. Sabían nuestros padres nuestros intentos, y no les pesaba dello, porque bien veían que, cuando pasaran adelante, no podían tener otro fin que el de casarnos, cosa que casi la concertaba la igualdad de nuestro linaje y riquezas.

L: 我们的父母知道我们的意思。他们并不担心，知道发展下去，最后无非是让我们结婚。这

简直是 门当户对 的安排。(FIG)

Y: 我们父母知道我们的心，可是并不担忧，因为他们很明白，到我们爱情更深厚的时候，无非让我们结婚就完了；彼此 门户相当，家道相称，简直就是天生的配偶。(Literal translation)

43. de la cual lamentable historia se puede sacar cuánto haya sido la crueldad de Marcela, el amor de Grisóstomo, la fe de la amistad vuestra, con el paradero que tienen los que a rienda suelta corren por la senda que el desvariado amor delante de los ojos les pone.

L: 从这个可悲的故事里，可以了解到马塞拉的残酷、克里索斯托莫的痴心、你们之间友谊的真诚以及在爱情的迷途上 执迷不悟 的人的结局。(AI)

Y: 我们从这段惨史能看到玛赛拉多么无情，格利索斯托莫多么痴心，你的友谊多么诚挚，而一个人 一纳头 走上爱情的迷途，会落到什么下场。(Literal translation)

44. y si él, con todo este desengañado, quiso porfiar contra la esperanza y navegar contra el viento, ¿qué mucho que se anegase en la mitad del golfo de su desatino?

L: 既然我讲得这样明白了，他还执迷不悟，逆风行舟，怎能不迷途翻船呢？(I-FIG)

Y: 我讲得这样明白，他还不死心，偏要 逆水行船，他掉进地狱去有什么说的呢？(I-FIG)

45. porque los buenos amigos han de probar a sus amigos y valerse dellos, como dijo un poeta, usque ad aras; que quiso decir que no se habían de valer de su amistad en cosas que fuesen contra Dios. (M.R. p. 350, “hasta los alteres”, no parece que se trate de ningún poeta sino de Pericles, a quien atribuye estas palabras Plutarco en uno de sus tratados morales)

L: 好朋友之间应该彼此信任，就像一位诗人说的，光明磊落，不应该利用友谊做违反上帝意志的事情。(FIG)

Y: 良朋好友之间的依赖和利用，应该象诗人所说的：‘能供在祭坛上’^①。这就是说，不该利用友谊干违反上帝的事。（“能供在祭坛上”（usque ad aras），这是纪元前五世纪古希腊政治家贝利克雷斯（Pericles）的话，见普鲁塔克（Plutarco）《杂文集》（Moralia）《论羞愧》（Del amaleverg(enza)）一文中。塞万提斯记错了，以为是一个诗人的话。(Literal translation)

46. Cuatro veces sosegó, y otras tantas volvió a su risa, con el mismo ímpetu que primero; de lo cual ya se daba al diablo don Quijote, y más cuando le oyó decir, como por modo de físga:

L: 这回唐吉珂德 怒不可遏 了。这时，只听桑乔以嘲笑的口吻说：(AI)

Y: 堂吉珂德瞧他这样，已经 冒上火来，禁不起他又连讥带讽，学着自己的腔吻说：(Literal translation)

47. especialmente de su español, en quien tenía siempre puestos los ojos y traía colgada el alma.

L: 不过她特别注意观察她那位西班牙人，眼睛始终不离开他，为他 牵肠挂肚。(FIG)

Y: 她尤其关心她的西班牙人，一双眼直盯着他，一片心也直围绕着他。(Literal translation)

48. de manera que el arte, imitando a la naturaleza, parece que allí la vence.

L: 真可谓巧夺天工。(FIG)

Y: 这种艺术摹仿天然而巧夺天工。(FIG)

49. el cual, viéndose tratar mal de aquel villano malandrín, puesta la cólera en su punto y crujiéndole los huesos de su cuerpo

L: 唐吉珂德见这个坏蛋竟敢如此对待自己，立刻气得七窍生烟！(FIG)

Y: 堂吉珂德瞧这混蛋对自己撒野，火气冲天，浑身的骨头都要爆裂了(I-AI)

50. y aquella gomia o esponja y polilla de la infinidad de dineros que allí sin provecho se gastaban

L: 这个贪得无厌、巧取豪夺、消耗了无尽钱财的要害。(AI)

Y: 这座城堡是个祸根，它好比饕餮的妖魔，好比海绵，好比蠹虫，吞吸和消耗了无穷无尽的金钱，(literal translation)

51. después que hubo reprehendido mi atrevimiento y afeado mi determinación, viéndome resuelta en mi parecer, se ofreció a tenerme compañía, como él dijo, hasta el cabo del mundo

L: 对我的大胆设想大加指责，可是看到我主意已定，就同意陪我去，还说哪怕是陪我到天涯海角。(FIG)

Y: 那长工先是怪我鲁莽，不赞成我的主意，可是瞧我很坚决，就自告奋勇要陪我，据他说，陪我到天涯地角也愿意。(I-FIG)

52. Pero haga el cielo lo que fuere servido, que tanto seré más estimado,

L: 不过还是听天由命吧，(FIG)

Y: 不过一切听天吧。(Literal translation)

53. —Pues como eso sea —respondió Sancho—, no hay sino encomendarnos a Dios y dejar correr la suerte por donde mejor lo encaminare.

L: 桑乔说，“咱们就求上帝保佑，听天由命吧。”(FIG)

Y: 桑丘说：“那么咱们只要靠上帝保佑，随命运去安排得了。(Literal translation)

54. y fue tanto su valor que, sin subir por los torpes medios y caminos que los más privados del Gran Turco suben, vino a ser rey de Argel, y después, a ser general de la mar, que es el tercero cargo que hay en aquel señorío.

L: 他没有像土耳其大公的心腹那样靠歪门邪道往上爬，而是靠自己的勇气终于成了阿尔及尔的国王，而后又成了海军统帅，成了那个统治阶层的第三号人物。(SB)

Y: 土耳其大皇帝的宠幸多半靠卑鄙的途径爬上高位，他却不然；他勇猛无比，因此做了阿尔及尔国王，后来又做了海上的统帅，这在土耳其帝国就是第三把交椅了。(Literal translation)

55. que os veis levantado del polvo de la tierra a ser señor de título, y correspondéis a tan buena obra con decir mal de quien os la hizo!

L: 一旦平步青云，受封晋爵，就以诽谤来回报一直扶植你的人呢！(FIG)

Y: 把你从泥土里提拔出来，封了你爵位，你却用混话来报答人家的恩情 (Literal translation)

56. Quítenseme delante los que dijeren que las letras hacen ventaja a las armas, que les diré, y sean quien se fueren, que no saben lo que dicen.

L: 如果有人说舞文弄墨比舞刀弄枪好，就让他从我面前滚开吧，那是信口胡言。(FIG)

Y: 谁说拿笔杆子的行业比拿枪杆子的高，那就请他们滚开去！凭他们是谁，我都要骂他们胡说八道。(Plain explanation)

57. —Es tan bueno —respondió Ginés—, que mal año para Lazarillo de Tormes y para todos cuantos de aquel género se han escrito o escribieren.

L: 希内斯说，”与之相比，《托尔梅斯河的领路人》以及其他所有那类书都相形见绌。(AI)

Y: 希内斯说：“好得很呢！压倒了《托美思河上的小癞子》①那类的书，不管是从前的或将来的，比了我的自传都一钱不值了。(FIG)

58. Viendo, pues, mi padre que, según él decía, no podía irse a la mano contra su condición, quiso privarse del instrumento y causa que le hacía gastador y dadivoso, que fue privarse de la hacienda, sin la cual el mismo Alejandro pareciera estrecho.

L: 据他自己说，他见自己积习难改，就想剥夺自己挥霍无度的手段和病因，也就是剥夺自己的财产。(AI)

Y: 据我父亲说，他瞧自己结习难改，就想铲除病根，就是说，分散自己的财产；(I-AI)

59. y, desta manera, andaban tan a una sus voluntades, que no había concertado reloj que así lo anduviese.

L: 两人总是心心相印，形同一人。(FIG)

Y: 这样呢，两人同心同意，便是准确的钟表也不能象他们那样协调 (literal translatio)

60. —Señor —respondió Sancho—, que el retirar no es huir, ni el esperar es cordura, cuando el peligro sobrepuja a la esperanza, y de sabios es guardarse hoy para mañana y no aventurarse todo en un día.

L: “退却不等于逃跑，等着也不算聪明。如果危险超过了希望，明智的办法就是养精蓄锐，而不是孤注一掷。(AI)

Y: 桑丘说：“先生啊，回避不是逃跑。凶险很大、出路很少的场合；死挺着算不得聪明。聪明人留着自己的身子等待来日，不在一天里拚掉性命 (Literal translation)

61. permitas que estos dos amantes le tengan, sin impedimento tuyo, todo el tiempo que el cielo quisiere concedérsele;

L: 让这一对有情人在天赐的良辰顺利地结成眷属，这样才能显示出你高贵的宽广胸怀，让大家看到你的理智战胜了欲望。(Literal translation)

Y: 让这一对有情人终成眷属，白头偕老，别再去破人好事，这才见得你的高尚慷慨，大家也就知道你能够以智胜情 (FIG)

62. Y lo que yo saco en limpio de todo esto es que estas aventuras que andamos buscando al cabo al cabo nos han

de traer a tantas desventuras, que no sepamos cuál es nuestro pie derecho.

L: 我 把 这 些 都 挑 明 了， 是 想 说， 咱 们 到 处 征 险， 结 果 给 自 己 带 来 许 多 不 幸， 弄 得 自 己 简 直 无所适从。(AI)

Y: 我 现 在 明 白 了 一 个 道 理： 咱 们 四 处 冒 险， 无 非 落 得 吃 尽 苦 头， 连 自 己 的 左 右 脚 都 分 辨 不 出。(Literal translation)

63. a la más necesaria ocasión y cuando es menester dar una traza que importe, se les yelan las migas entre la boca y la mano, y no saben cuál es su mano derecha.

L: 关 键 时 刻 需 要 他 们 拿 主 意 的 时 候， 他 们 却 举棋不定， 手 足 无 措。(FIG)

Y: 在 紧 要 关 头， 必 须 有 急 智 的 时 候， 这 些 人 往 往 拿着面包不会往嘴边送， 自 己 的 左 右 手 都 分 辨 不 出。(Literal translation)

64. a la más necesaria ocasión y cuando es menester dar una traza que importe, se les yelan las migas entre la boca y la mano, y no saben cuál es su mano derecha.

L: 关 键 时 刻 需 要 他 们 拿 主 意 的 时 候， 他 们 却 举棋不定， 手 足 无 措。(FIG)

Y: 在 紧 要 关 头， 必 须 有 急 智 的 时 候， 这 些 人 往 往 拿着面包不会往嘴边送， 自己的左右手都分辨不出。(literal translation)

65. el cual, viéndose tratar mal de aquel villano malandrín, puesta la cólera en su punto y crujiéndole los huesos de su cuerpo, como mejor pudo él, asió al cuadrillero con entrambas manos de la garganta, que, a no ser socorrido de sus compaÑeros, allí dejara la vida antes que don Quijote la presa.

L: 唐 吉 诃 德 见 这 个 坏 蛋 竟 敢 如 此 对 待 自 己， 立 刻 气 得 七窍生烟！ 他 用 双 手 紧 紧 掐 住 了 团 丁 的 脖子。(FIG)

Y: 堂 吉 诃 德 瞧 这 混 蛋 对 自 己 撒 野， 火气冲天，浑身的骨头都要爆裂了。 他 拚 命 用 两 手 卡 住 那 巡 逻 队 员 的 脖子， 那 家 伙 要 没 有 同 伙 帮 忙， 不 等 堂 吉 诃 德 松 手 早 送 命 了 (literal translation)

66. deste modo vendrá a ser mi libro, al cabo de haberme quemado las cejas por guardar los preceptos referidos, y vendré a ser el sastre del cantillo.

L: ... 我 的 书 也 会 是 这 样。 如 果 我 想 保 持 它 的 艺 术 性， 即 使 我 呕心沥血 地 写 出 来， 也 只 能 落 个 费 力 不 讨 好 的 结 局。(FIG)

Y: 我 按 照 上 面 说 的 艺 术 规 律， 精心费力 写 出 来 的 书， 也 逃 不 了 这 样 的 遭 遇， 我 就 成 为 ‘ 四 岔 路 口 的 裁 缝 ’ ② 了。(SSY)

67. que, rompiendo imposibles y malezas,

L: 不 畏 艰 难， [[披荆斩棘]]， (FIG)

Y: 你 不 顾 困 难、 突破障碍， (plain explanation)

68. —No le mana, canalla infame—respondió don Quijote encendido en cólera—, no le mana, digo, eso que decís, sino ámbar y algalia entre algodones; y no es tuerta ni corcovada, sino más derecha que un huso de Guadarrama.

L: “无耻的恶棍，”唐吉珂德怒气冲天地说，“她眼里流出的不是你说的那些东西，而是珍贵的琥... (I-FIG)

Y: 堂吉珂德勃然大怒，喝道：“无耻的混蛋！她眼睛里不流那些东西！不流你说的那些东西！流的是龙涎香和裹在棉花里的麝香②！（AI）

69. pero estaba ya el mozo picado y no quiso dejar el juego hasta envidar todo el resto de su cólera; y, acudiendo por los demás trozos de la lanza, los acabó de deshacer sobre el miserable caído,

L: 可骡夫已经怒不可遏，直打到怒气全消才住手。然后，骡夫捡起其余几截断矛 (AI)

Y: 那小子却打上火来，定要打个畅快才罢。他拣起其余的断柄，一股脑儿全撒在那摔倒的可怜虫身上。(plain explanation)

70. de lo cual ya se daba al diablo don Quijote, y más cuando le oyó decir, como por modo de fisga:

L: 这回唐吉珂德怒不可遏了。这时，只听桑乔以嘲笑的口吻说：（AI）

Y: 堂吉珂德瞧他这样已经冒上火来，禁不起他又连讥带讽，学着自己的腔吻说：(plain explanation)

71. Duró este engaño algunos días, hasta que, al cabo de pocos meses, volvió Fortuna su rueda y salió a plaza la maldad con tanto artificio hasta allí cubierta, y a Anselmo le costó la vida su impertinente curiosidad

L: 这个骗局持续了一段时间。几个月后，命运女神扭转乾坤，他们精心设置的骗局昭然若揭，安塞尔莫则因为自己无谓的 (FIG)

Y: 事情一时上没闹破，直到几个月后，命运的轮子转了过来，掩盖得非常巧妙的丑事就传扬开了。(literal translation)

72. Duró este engaño algunos días, hasta que, al cabo de pocos meses, volvió Fortuna su rueda y salió a plaza la maldad con tanto artificio hasta allí cubierta, y a Anselmo le costó la vida su impertinente curiosidad.

L: 这个骗局持续了一段时间。几个月后，命运女神扭转乾坤，他们精心设置的骗局昭然若揭，安塞尔莫则因为自己无谓的 (FIG)

Y: 事情一时上没闹破，直到几个月后，命运的轮子转了过来，掩盖得非常巧妙的丑事就传扬开了。(plain explanation)

73. mas, a lo que se dice, en fin se hará lo que Ambrosio y todos los pastores sus amigos quieren, y mañana le vienen a enterrar con gran pompa adonde tengo dicho.

L: 明天，他们要到我刚才说的那个地方大张旗鼓地安葬。(FIG)

Y: 据说后来还是按照安布罗修和他那些牧羊伙伴的主张办事。明天他们要在我说的那地方举行别致的葬礼，(plain explanation)

74. Porque mi Dulcinea del Toboso osaré yo jurar que no ha visto en todos los días de su vida moro alguno, así como él es, en su mismo traje, y que se está hoy como la madre que la parió:

L: 我敢发誓，托博索的杜尔西内亚这辈子从未见过一个穿着摩尔人衣服的摩尔人。她至今仍守身如玉。(FIG)

Y: ... 子 也 没 看 见 过 一 个 穿 摩 尔 服 装 的 地 道 摩 尔 人 , 她 现 在 就象 生 她 的 妈 妈 一 样 (Literal translation)

75. el cual alcanzado, a muchos hemos visto que, habiendo pasado por estas Sirtes y por estas Scilas y Caribdis, como llevados en vuelo de la favorable fortuna, digo que los hemos visto mandar y gobernar el mundo desde una silla,

L: 他们 含辛茹苦, 一旦达到了目标, 就好像插上了时来运转的翅膀, 开始坐在椅子上统治世界, (FIG)

Y: 咱们看见许多人 历尽艰难困苦, 到了这一步, 就飞黄腾达了; 咱们看见他们坐在安乐椅里辖治世界 (plain explanation)

76. Y lo que sería mejor y más acertado, según mi poco entendimiento, fuera el volvernos a nuestro lugar, ahora que es tiempo de la siega y de entender en la hacienda, dejándonos de andar de ceca en meca y de zoca en colodra, como dicen.

L: 我觉得最好咱们掉头回老家去。现在正是收获季节, 咱们去忙自己的活计, 别像俗话说的‘东奔西跑, 越跑越糟’啦。” (SSY)

Y: 依我浅见, 现在正是收获的季节, 最好还是回村料理咱们的田地去, 别象老话说的‘[[东奔西走]], 乱撞乱投’” (SSY)

77. pero pensar que hemos de volver ahora a las ollas de Egipto, digo, a tomar nuestra cadena y a ponernos en camino del Toboso, (M.R. p. 227, *volver a la mala vida pasada*)

L: 但是, 如果以为我们现已回到了 太平盛世, 可以拿着锁链去托博索了, (FIG)

Y: 您如要我们这会子 回到埃及的肉锅旁边去 ①, 换句话说, 要我们扛着这副链子到托波索的大道上去 (Literal translation)

78. es pensar que es ahora de noche, que aún no son las diez del día, y es pedir a nosotros eso como pedir peras al olmo.

L: 那简直是白日说梦, 让我们 [[缘木求鱼]]. (I-FIG)

Y: 那就等于说, 目前不是上午十点, 却是夜晚; 您要我们干这件事就仿佛‘要榆树结梨’ ②。(② 西班牙谚语, 指不可能的事。)(Literal translation)

79. Si no, por vida de... Basta, que podría ser que saliesen algún día en la colada las manchas que se hicieron en la venta, y todo el mundo calle y viva bien y hable mejor, y caminemos, que ya es mucho regodeo este.

L: ... 不客气点儿, 我发誓 ... 行了, ‘说不定哪天客店的事情就会 [[水落石出]] 呢’. 谁也别说了, 你好好待着, 说话客气点儿. 已经费半天 ..(FIG)

Y: 你要是不客气, 哼哼.. 我不用多说。‘客店里沾上的肮脏, 说不定有一天会漂洗干净’ ①。大家别闹, 好好过日子, 说话放和气些。咱们耽搁得够了, 上路吧。” (plain explanation)

80. Digo que oyeron que daban unos golpes a compás, con un cierto crujir de hierros y cadenas, que, acompaÑados del furioso estruendo del agua, que pusieran pavor a cualquier otro corazón que no fuera el de don Quijote.

L: ...伴随着水的咆哮声，除了唐吉珂德，任何人听到这种声音都会[[毛骨悚然]]。(FIG)

Y: 那是有节奏的敲打声，夹杂着铁片和铁链的碰擦声，再加上汹涌的水声。除了堂吉珂德，谁听了都会害怕的。(plain explanation)

81. el cual alcanzado, a muchos hemos visto que, habiendo pasado por estas Sirtes y por estas Scilas y Caribdis, como llevados en vuelo de la favorable fortuna, digo que los hemos visto mandar y gobernar el mundo desde una silla, trocada su hambre en hartura, su frío en refrigerio, su desnudez en galas, y su dormir en una estera en reposar en holandas y damascos:

L: 他们含辛茹苦，一旦达到了目标，就好像插上了时来运转的翅膀，开始坐在椅子上统治世界，饥肠辘辘变成了脑满肠肥，忍寒受冻变成了怡然自得，缺衣少穿变成了穿着阔绰，铺席而眠变成了铺绫盖缎。(SSY)

Y: 咱们看见许多人历尽艰难困苦，到了这一步，就飞黄腾达了；咱们看见他们坐在安乐椅里辖治世界，他们吃得好，住得暖，穿上鲜衣华服，睡在铺着细布和花缎的床上，再也不挨饿受冻、衣衫破烂、垫着草席子睡觉了。(Literal translation)

82. Y quien lo contrario dijere -dijo don Quijote-, le haré yo conocer que miente, si fuere caballero, y si escudero, que remiente mil veces.

L: 不管他是骑士还是侍从，都是在说[[弥天大谎]]。(FIG)

Y: 哪个侍从说这话，我就要他承认自己是一千个撒谎，一万个撒谎！(literal translation)

83. Así que, Sancho amigo, no te congoje lo que a mí me da gusto: ni quieras tú hacer mundo nuevo, ni sacar la caballería andante de sus quicios.

L: ...所以，桑乔朋友，你别担心，我愿意要这种东西。你也不要别出心裁，惹游侠骑士生气。”(AI)

Y: 所以，桑丘朋友，你别为我甘心的人事担忧，别另出花样，也别去改革游侠骑士道的常规。”(plain explanation)

84. —Gracias doy al cielo por la merced que me hace, pues tan presto me pone ocasiones delante donde yo pueda cumplir con lo que debo a mi profesión y donde pueda coger el fruto de mis buenos deseos.

L: 苍天如此迅速赐给我机会，让我尽自己的职责，实现夙愿，旗开得胜。(FIG)

Y: 感谢上天照应，叫我马上有机会尽尽本分，实现自己的雄心壮志。(plain explanation)

85. ella va aquí de su voluntad, tan contenta, a lo que yo imagino, de verse en este estado, como el que sale de las tinieblas a la luz, de la muerte a la vida y de la pena a la gloria".

L: 我想，她心甘情愿走到这一步，可以说是弃暗投明，起死回生，由辱变荣。’(SSY)

Y: 我看她对当前的情况非常乐意，好象是从黑暗投入光明，从死亡投入永生，从烦恼投入欢乐 (literal translation)

86. Ese es el cuerpo de Grisóstomo, que fue único en el ingenio, solo en la cortesía, extremo en la gentileza, fénix en la amistad, magnífico sin tasa, grave sin presunción, alegre sin bajeza,

L: 这是克里索斯托莫的身体。他聪颖过人，温文尔雅，慷慨大度，友遍四方，尊贵无上；(I-

FIG)

Y: 死者 格利索斯托 莫 是最杰出的天才，最有礼貌，最温文，最笃于友谊，也最豪爽慷慨；
(plain explanation)

87. mezclados entre ellas pedazos de cristal luciente y de contrahechas esmeraldas, hacen una variada labor, de manera que el arte, imitando a la naturaleza, parece que allí la vence.

L: ...和祖母绿交织在一起，形成了一幅五彩缤纷的作品，真可谓巧夺天工。(FIG)

Y: 还镶嵌着闪亮的水晶和仿造的翡翠。这种艺术摹仿天然而巧夺天工。(FIG)

88. Pero, como suele decirse que un mal llama a otro, y que el fin de una desgracia suele ser principio de otra mayor, así me sucedió a mí, porque mi buen criado, hasta entonces fiel y seguro,

L: 人们常说祸不单行，逃出狼窝又进了虎口，我就遇到了这种情况。(FIG)

Y: 可是正应了老话说的‘祸不单行’，又说是‘灾祸往往由小到大，衔接而来’，我就碰到了这种情况。(FIG)

89. Quiso, digo, quitar a las galeras sus pies, poner en alboroto a la Santa Hermandad, que había muchos años que reposaba;

L: 因此我说呀，他放了那些苦役犯就是放虎归山，给圣友团带来了麻烦，本来圣友团已经好多年没有事干了。(FIG)

Y: 我说呀，他是存心剥夺海船上划船脚力，并且使安静了好多年的神圣友爱团又忙乱起来。(Literal translation)

90. que por el hilo se sacará el ovillo y quedaremos con esto satisfechos y seguros, y vuestra merced quedará contento y pagado;

L: 这样一了百了，我们满意了，放心了，您也高兴了，满足了。我们渴望瞻仰她的芳容。
(FIG)

Y: 因为‘拿到了线头儿，就抽开了线球儿’^①。这样我们才心安，您也可以满意。^①西班牙谚语。意思是有了线索，便知底蕴。(Literal translation)

91. y si vos queréis porfiar y espolear y dalle, será enojar a la fortuna y dar coces, como dicen, contra el aguijón.

L: 如果您还这么踢它，就会惹怒老天，就像人们说的，物极必反。(FIG)

Y: 您如果还要固执，只顾踢它，硬要它走，就会触犯造化的神道，就是老话说的‘向钉子上硬碰’。(Literal translation)

92. Con todo eso -dijo don Quijote-, mira, Sancho, lo que hablas, porque tantas veces va el cantarillo a la fuente..., y no te digo más.

L: “可也是，”唐吉诃德说，“你看你总是说，桑乔，喋喋不休，难免...行了，我不说了。”(FIG)

Y: 堂吉诃德道：“可是，桑丘，你说什么话得仔细想想。因为‘水罐儿一次次到井边去..’^①，底下我不说了。”(^①西班牙谚语：“水罐儿一次次到井边去，结果就砸碎了。”) (Literal translation)

93. Y ¡cómo si llevaba azogue! -dijo don Quijote-, y aun una legión de demonios, que es gente que camina y hace caminar, sin cansarse, todo aquello que se les antoja. (M.R. p.332, alude a una ardida de los gitanos consistente en echar azogue en los oídos de las caballerías para que corran más)
- L: “它 矫健如飞，” 唐吉珂德说，“因为有很多鬼怪簇拥着它呢。它们可以随心所欲地不间断地跑路或者带着人跑路。(I-FIG)
- Y: 堂吉珂德说：“仿佛灌了水银吗？大批的魔鬼簇拥着它呢！魔鬼自己能跑，如果高兴，还能带着人畜跑，叫他们跑了路不累。(Literal translation)
94. Vete de mi presencia, monstruo de naturaleza, depositario de mentiras, almario de embustes, silo de bellaquerías, inventor de maldades, publicador de sandeces, enemigo del decoro que se debe a las reales personas!
- L: 你这个万恶的魔鬼，竟敢造谣生事，[[蛊惑人心]]，真是卑鄙至极，愚蠢透顶，污辱贵人的尊严。(FIG)
- Y: 你给我滚！你这魔头！你这撒谎精、促狭鬼！你这诡计多端、造谣生事的家伙！你这不敬帝王的逆贼！你快滚！(literal translation)
95. aunque tengo para mí que aquellos que se holgaron conmigo no eran fantasmas ni hombres encantados, como vuestra merced dice, sino hombres de carne y de hueso como nosotros;
- L: 并非像您所说的那样是什么鬼怪或魔法师,而是和我们一样 有血有肉 的人。(FIG)
- Y: 不过我觉得捉弄我的那伙人不是您说的鬼怪，也不是魔法支使的，他们和咱们一样是 有皮肉筋骨 的人 (literal translation)
96. que con toda solicitud y advertimiento miraba por la honra porque no pareciese mal al vulgo ocioso y a los ojos vagabundos y maliciosos la entrada de un mozo rico, gentilhombre y bien nacido, y de las buenas partes que él pensaba que tenía, en la casa de una mujer tan hermosa como Camila;
- L: 他知道自己有些优越条件，因而在一些 [[游手好闲]]、别有用心的小人看来，一位如此富有、英俊而又出身高贵的...(FIG)
- Y: 因为象他这样一个富贵公子，自己知道颇有几分人材，如果经常到卡蜜拉那样漂亮夫人的家里去，那些 吃了闲饭没事干的人 就不免歪言散语，恶意中伤。(plain explanation)
97. aunque don Quijote fuera verdaderamente de los caballeros andantes de la Tabla Redonda, no le estimaran en dos ardites.
- L: 他们觉得唐吉珂德即使真是圆桌骑士，也 一文不值。(FIG)
- Y: 即使堂吉珂德真是圆桌骑士里的一员，在他们眼里也 不值半文钱。(plain explanation)
98. Ya que quieres, crüel, que se publique, de lengua en lengua y de una en otra gente del áspero rigor tuyo la fuerza,
- L: 狠毒的你，既然愿意，把你的冷酷，[[公诸于众]]，任人 街谈巷议 (SSY)
- Y: 狠心的姑娘，你既要 众口宣扬 你坚如铁石又冷若冰霜 (plain explanation)

99. y en verdad que estoy por condenarlos no más que a destierro perpetuo, siquiera porque tienen parte de la invención del famoso Mateo Boyardo, de donde también tejió su tela el cristiano poeta Ludovico Ariosto;

L: 接着又由基督教诗人卢多维科阿里奥斯托来添枝加叶。(FIG)

Y: 这些人物对名诗人玛德欧博雅铎②的作品有贡献；基督教诗人卢铎维戈阿利奥斯陀③又从博雅铎取材。(Plain explanation)

Appendix IV A complete situational framework for register studies: situational parameters of variation
 Biber, D. (1994) "An analytical framework for register studies", in *Sociolinguistic Perspectives on Register*, D. Biber and E. Finegan (eds.), 31-56

1. Communicative characteristics of participants

A. Addressor(s): single, plural or institutional

B. Addressee(s):

1. self/other

2. Single/plural/unenumerated

C. audience: yes/no

2. Relations between addressor and addressee

A. social role relations -- relative status and power of addressor and addressee

Addressor has more power /equal status/ addressee has more power

B. Extent of shared knowledge

1. Specialist knowledge of topic: high/low

2. Specific personal knowledge: high/low

C. Interactiveness: extensive/slight/none

D. Personal relationship: like, respect, fear, kin, friends, enemies, colleagues, etc.

3. Settings

A. Characteristics of the place of communication

1. private/public

2. Domain:

Business and workplace

Education and academic

Government and legal

Religious

Art and entertainment

Domestic and personal

Other

3. Audio /visual mass media (television, radio, cinema)

B. Extent to which place is shared by participants

Immediate/familiar/removed

C. Extent to which time is shared by participants

Immediate / familiar/removed

D. Specific place and time of communication

4. Channel

A. mode (primary channel)

Written/spoken/signed/mixed/(other)

B. permanence

Recorded/transient

C. Medium of transmission

5. Relation of participants to the text

A. addressor -- production circumstances: revised or edited/scripted/planned/on-line

B. Addressee -- comprehension circumstances: on-line/self-imposed constraints

C. Addressor's and addressees personal evaluation of text: important, valuable, required, beautiful, popular, etc.

D. Addressor's attitudinal stance towards the text:

1. Emotionally involved/removed

2. Reverence/everyday

3. Excitement

Etc

E. Addressor's epistemological stance towards the text:

Belief, conviction, doubt, etc.

6. Purposes, intents and goals

A. factuality:

(Purported to be) based on fact, speculative, imaginative, mixed

B. Purposes

1. Persuade or sell; high/medium/low

2. Transfer information: high/medium/low

3. Entertain /edify: high/medium/low

4. Reveal self (including expression of personal feelings, attitudes, or efforts at enhancing interpersonal relations): high/medium/low

7. Topic/subject

A. level of discussion: specialized, general and popular

B. specific subject: finance, science, religion, politics, law, people, daily activities, etc.

Appendix V Analysis of Don Quijote's increased use of archaic speeches in Subdivision VIII & IX

	AN	AU	ST	SSKT	DR	FR	PL	DO	ADE	AEE	ADA	ADEP	FA	PE	IN	SR	LE
4/29	1	2	3	3	3	3	2	2	2	3	3	4	1	2	1	2	1
11/30	1	2	1	1	1	1	2	2	2	1	3	4	2	2	2	2	1
12/30	1	2	3	3	3	3	2	2	2	3	3	4	1	2	1	2	1
18/30	1	2	3	3	3	3	2	2	2	3	3	4	1	2	2	2	1
19/30	1	2	3	1	3	3	2	2	2	3	3	4	2	2	2	2	1
20/30	1	2	1	1	1	1	2	2	2	1	3	4	2	2	2	2	1
23/30	1	2	1	1	1	1	2	1	1	1	2	3	3	1	1	1	2
24/30	1	1	1	3	1	1	1	1	2	1	3	2	2	2	1	1	2
27/30	1	1	1	1	1	1	1	1	2	1	3	3	3	2	1	1	2
36/31	1	1	1	1	1	1	1	2	2	1	3	4	1	2	1	1	1
40/31	1	1	1	1	1	1	1	2	2	1	3	4	2	2	2	2	1
41/31	1	1	1	1	1	1	1	2	2	1	3	4	2	2	1	2	1
44/31	1	1	1	1	1	1	1	2	2	1	3	4	1	2	2	2	1
46/31	2	2	2	2	2	2	2	2	2	2	3	4	3	2	2	2	1
48/31	1	2	1	1	3	3	2	2	2	1	3	4	3	2	2	2	1
50/31	1	2	3	3	3	3	2	2	2	1	3	4	2	2	1	2	1
6/37	1	1	1	1	1	1	1	2	1	1	2	4	1	2	1	1	1
8/37	1	1	3	2	3	3	2	2	2	2	3	4	1	2	1	2	1
12/37	1	1	2	1	2	2	2	2	2	2	3	4	3	2	1	2	2
13/37	2	2	2	1	2	2	2	2	2	3	3	4	1	2	1	2	1
14/37	2	2	2	1	2	2	2	2	2	3	3	4	3	2	1	2	1
15/38	2	2	2	1	2	2	2	2	2	3	3	4	3	2	1	2	1
17/43	1	1	3	1	3	3	1	2	2	3	3	4	1	2	1	2	1
18/43	2	2	1	1	3	3	1	2	2	1	3	4	2	2	1	2	1
25/43	2	2	1	1	3	3	2	2	1	1	1	1	3	1	1	2	1
26/44	2	2	1	1	3	3	2	2	2	1	3	4	1	2	1	2	1
33/45	2	2	2	1	2	2	2	1	2	2	2	3	2	2	1	2	2
36/45	2	2	2	1	2	2	2	2	2	2	3	2	2	1	1	1	1
37/45	2	2	2	1	2	2	2	1	1	2	2	2	3	1	1	1	2
40/45	2	2	2	1	2	2	2	2	2	2	3	4	1	2	3	2	1
41/45	2	2	1	1	3	3	2	2	2	1	3	1	2	1	3	2	1
42/46	1	1	3	1	3	3	2	2	2	3	3	4	1	2	3	2	1
43/46	1	1	3	1	3	3	2	2	2	3	3	4	1	2	3	2	1
45/46	1	1	1	1	1	1	2	1	1	1	2	3	3	1	1	2	2

46/46	1	1	1	1	1	1	2	2	2	1	3	1	2	2	3	2	1
50/46	1	1	3	3	3	3	2	2	2	3	3	4	1	2	1	2	1

Appendix VI Data encoding of Subdivision X (Liu, I)

CHAPTER	47			48		49						50	52			
Speech no.	1	4	6	7	9	12	14	15	16	17	18	20	24	27	28	29
AD	1	2	2	1	1	1	1	1	1	1	1	1	1	2	2	2
AU	2	2	2	1	1	1	2	2	2	2	2	2	2	2	2	2
ST	1	3	3	1	1	1	2	1	1	1	1	1	1	2	1	1
SSKOP	1	1	1	1	1	3	1	1	1	1	1	1	1	3	1	1
PL	2	2	2	1	1	1	2	2	2	2	2	2	2	2	2	2
DO	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2
ADE	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2
ADA	3	3	3	3	2	2	2	3	3	3	3	3	3	3	3	3
ADEP	2	4	4	3	1	3	3	4	1	1	4	4	4	4	4	4
FAC	2	1	1	2	3	2	1	1	3	3	2	2	2	2	3	2
INS	1	2	2	2	1	2	1	2	1	1	2	2	2	2	2	2
SR	1	2	2	2	1	2	2	2	1	1	2	2	2	2	2	2
LE	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1

Appendix VII Computation of factor-based scores

Chapter	Environment	Content	Chapter	Environment	Content
8	0.54	0.41	9	0.17	-1.53
8	0.54	0.41	9	0.17	-1.53
8	0.54	0.41	9	-0.96	-2.89
8	0.54	0.41	9	1.17	-0.35
8	0.54	0.41	9	0.17	-1.53
8	0.54	0.41	9	0.25	0.76
8	-1.15	-1.61	9	0.25	0.76
8	-2.35	-0.19	9	-1.05	-0.79
8	-1.95	0.28	9	-0.75	-0.43
8	-1.16	1.23	9	0.25	0.76
8	-0.66	1.83	10	-0.46	-0.78
8	-0.66	1.83	10	1.17	-0.35
8	-0.66	1.83	10	1.17	-0.35
8	1.17	-0.35	10	-1.16	1.23
8	0.54	0.41	10	-3.23	-1.25
8	0.54	0.41	10	-1.16	1.23
9	-1.44	0.9	10	0.03	-0.19
9	0.25	0.76	10	0.54	0.41
9	-0.26	0.16	10	-0.76	-1.14
9	1.17	-0.35	10	-0.76	-1.14
9	1.17	-0.35	10	0.54	0.41
9	1.17	-0.35	10	0.54	0.41
9	-0.66	1.83	10	0.54	0.41
9	0.54	0.41	10	1.17	-0.35
9	-0.13	-1.9	10	1.17	-0.35
9	1.17	-0.35	10	1.17	-0.35