Variability and Stability in Foreign and Second Language Learning Contexts: Volume 1

Variability and Stability in Foreign and Second Language Learning Contexts: Volume 1

Edited by

Ewa Piechurska-Kuciel and Liliana Piasecka

**CAMBRIDGE
SCHOLARS**

P U B L I S H I N G

CONTENTS

PREFACE

Stability and variability are concepts that reflect important characteristics of reality, which have pushed humans forward in many spheres of life. On the one hand, we all would like to live in a safe and predicatable world where, free from fear and danger, we can pursue our daily activities. On the other hand, many associate stability with stagnation, which is the enemy of progress and creativity. But stability does not have to be interpreted in such a way. Recent decades have brought about many changes in countless manifestations of human activity, the most impressive being the development of information and communication technologies. Not only have they made interpersonal communication easy but they have also enabled communication among people from different, frequently very distant parts of the world. However, communication by means of the Internet, for example, has not changed the fact that it is based primarily on language – a stable though variable feature of *homo sapiens*. In addition, many people can use more than one language, which makes them more versatile communicators. Contributors to this book are all concerned with how stability and variability work in the context of language, language learning and teaching.

The brief overview (below) of individual chapters included in the two volumes implies a wide spectrum of topics organized within a relatively fixed framework of Applied Linguistics theory and practice, revolving around the concepts of stability and variability, which capture the dynamic nature of the phenomena characterizing language, learning and teaching. The indisputable strength of the papers lies in the fact that the vast majority report original empirical studies that were carried out in diverse second/foreign language learning contexts, investigating interesting issues across various nationalities, ages, and educational and professional groups of language learners and teachers. The issues under scrutiny embrace the 'classical', stable topics related to language learning and teaching, such as communicative competence, input, orality and literacy, learner characteristics and strategies, and teacher development, to mention just a few. In addition, 'recent arrivals', to borrow a marketing metaphor, are also present, as the authors consider learning and teaching implications resulting from the status of English as the language of international communication, and when they discuss related concepts of intercultural competence along with language learners' identity and creativity.

The contributors to the present volumes are researchers – foreign and second language learners and teachers themselves – who offer the reader a range of methodological designs that are successfully used in Applied Linguistics research. Last but not least, the multilingual and multicultural character of the contributors and their contributions needs to be underscored. In addition, the framework of stability and variability suggests that changes leading to progress and development derive from stable foundations that account for the sense of continuity and belonging in the applied linguists' communities of practice.

This book is divided into two volumes. **Volume One** focuses on language and language teaching. Part One of the volume comprises chapters related to such issues as English as an International Language, language in mass media, translation, interpreting, etymological aspects of dictionary-making, and acquisition of English articles.

Joanna Kapica-Curzytek, following Umberto Eco's (2002) search for a universal language, discusses English from this perspective, arguing that a change in the status of the language from national to international imposes a new role on the language teacher, namely that of a "broker of global (universal) values". Recognizing the role of English as International Language (EIL), or English as the *lingua franca* (ELF) of today, Karen Jacob discusses the current role of cultural information in English language learning and teaching (ELT). She is concerned whether Spanish learners of English as a foreign language perceive the cultural content of English language courses within the traditional EFL language paradigm or within the more recent proposal of a Global English paradigm. Elizabeth Woodward-Smith, in turn, shows how advertising discourse has evolved from informative to humorous techniques in response to historical, social, cultural and economic factors, basing her analysis on audiovisual advertising from television and the Internet, and on incongruity theory. Language use in mass media is also of concern to Sylwia Wrześniewska, who presents variable linguistic means employed to pass negative information in press headlines. Lech Zabor shares with readers the findings of his study on the acquisition of English articles by Polish learners through the lenses of definiteness and specificity. The shift of interest from morphology to meaning is evident in the chapter by Mariusz Kamiński, whose analysis of the history of general English lexicography shows how the etymological bias of dictionary makers was changed by empirical and systematic studies of words based on a large multi-lingual corpus, allowing lexicographers to separate meaning from etymology.

Two other chapters in this part of the volume refer to translation and interpreting. Searching for translation universals, Łukasz Grabowski has

compared translated and non-translated literary Polish texts and found that translated texts are more lexically complex and contain more precise and explicit sentences than the non-translated texts. The former also feature a higher lexical variety of bottom-frequency words than the latter. Grabowski suggests that these two differences may be regarded as T-universals. Aleksandra Wachla-Gierałtowicz, meanwhile, is concerned with simultaneous interpreting – a multitask requiring the use of a number of cognitive tasks – from the perspective of the quality of memorized and retrieved information in novice interpreters.

Part Two is devoted to current thinking and practice connected with foreign/second language teaching from the point of view of teaching contexts, input features that enhance language learning, the growth of implicit knowledge, as well as the role and development of foreign language teachers across the range of contexts in which they are active.

Maria Juan-Garau discusses variable effects of three different SLA learning contexts (formal instruction, content and language integrated learning, and study abroad) in terms of the opportunities for contact with the target language, opportunities for practice, and linguistic benefits they offer, but also in terms of learning outcomes. Context and input are relevant to Marek Kuczyński, who presents the case of two children, born and living in Poland, who have acquired English via natural communication at home (conversations with their father, TV programs and English language texts), and consequently developed a fluent oral communicative competence despite limitations in their L2 vocabulary knowledge. Similarly, Agnieszka Pietrzykowska is concerned with the influence of enhancing visual input by means of highlighting or boldfacing target language forms (for example, grammatical structures) on the process of noticing. The findings of the study she reports imply that increasing the saliency of a given form results in a higher level of noticing. The issue of variability is also explored by Mirosław Pawlak, researching the employment of implicit knowledge as a function of task type, proficiency level, and targeted linguistic form. On the basis of the research findings the author concludes that though the same type of knowledge was elicited, it also varied due to task difficulty and language proficiency, which required different types of processing.

Anna Niżegorodcew provides the reader with an insightful analysis into a workshop carried out within the European Master for European Teacher Training Project (EMETT), one of whose objectives was to train intercultural language teachers. Niżegorodcew describes how a workshop based on Czesław Miłosz's poetry was an attempt to develop the intercultural competence of international students. The issue of

intercultural competence is related to the status of English as the *lingua franca* of the multicultural world. Aleksandra Wach, then, explores the attitudes toward ELF expressed by Polish teachers of English as a foreign language, which range from enthusiastic to skeptical. A deep concern for teacher qualifications and development is evident in the chapter by Marek Derenowski, who examines students' and teachers' subjective perceptions of the roles that teachers perform, their professional development, the teachers' actual classroom practices, and the possible relations between teaching experience and professional development. The role of the teacher as a facilitator of the learning process and a group member is one of the leading themes in Dagmara Gałajda's chapter. The author explores how activities facilitating group dynamics, fundamental to the process of learning a foreign language, enhance classroom interaction, goal-orientedness, group identity and the skills of working as a team.

Volume Two focuses on the language learner. This perspective is quite meaningful as it highlights a large variety of factors that may have an impact on learning a foreign language. It also reflects the shift of interest from the language and the teacher to the learner – the most important actor engaged in the language learning process. The sixteen chapters address a wide gamut of factors, divided into three subsections. Part One covers individual cognitive and affective learner differences, Part Two centers around learning language skills and subsystems, while Part Three concerns learners with special educational needs (developmental dyslexia and autism).

In Part One (learner variables) Danuta Gabryś-Barker delves into multilinguals' learning stories with the intention to find out how their L2 and L3 learning experiences in a formal instructional context affect their motivations, attitudes, learning styles and strategies, perceived relations between the languages, and difficulties encountered when learning these languages. Gabryś-Barker's text introduces a very broad field of individual learner differences that have been investigated by the authors of other chapters in this section. Marina Mattheouidakis focuses on language aptitude and discusses the relationship between aptitude and young learners' English proficiency, stressing the dynamic nature of aptitude. Adriana Biedroń analyzes the personality of gifted foreign language learners, locus of control, and the style of coping with stressful and upsetting situations as variables modifying foreign language aptitude (FLA). Although she has found no significant relations between personality and cognitive factors, she concludes that her results evidence a dynamic and non-linear interaction between personality and cognitive factors, which is in line with dynamic systems theory.

Liliana Piasecka's chapter, in turn, is devoted to foreign language learners' identity, which is intertwined with and inseparable from the languages they use. Knowledge of an L2 can be viewed both as cultural and symbolic capital, therefore learning an L2 is an investment into one's present and future, and as such it may be a powerful motivating factor for language learners. While Piasecka discusses identity within a post-modern framework, Anna Mystkowska-Wiertelak elaborates on Dörnyei's (2005, 2009) L2 Motivational Self System theory and discusses how the tension between one's actual and ideal/desired self motivates the learner to learn an L2 and how this may affect his or her identity.

Gender – another important individual learner difference – is addressed by Ewa Piechurska-Kuciel and Ewa Jocher. The first author is concerned with communication skills in L2, and therefore she analyzes the relationship between gender and willingness to communicate in a foreign language. Her research findings reveal that female foreign language learners show higher levels of willingness to communicate in the L2 both in the classroom and outside than males. While Piechurska-Kuciel deals with oral communication, Ewa Jocher examines the possible influence of adolescents' gender and major (their main subject of study, such as humanities or science) on their L1 and L2 reading motivation, thus linking orality and literacy. On the basis of her study Jocher shows that females are more positively motivated to read in L1 and L2 than males, and that humanities majors have higher reading motivation than science majors, regardless of the language of the text.

In Part Two pronunciation is tackled by Magdalena Szyszka, who investigates pronunciation learning strategies and tactics that might improve the learners' pronunciation and, in consequence, their communicative competence. Arkadiusz Rojczyk and Andrzej Porzuczek report a study that shows how Polish learners of English cope with contexts that require vowel reduction, and compare their results with the results of native speakers of English. As regards language skills, Aleksandra Maryniak discusses reading effectiveness in a blended learning context that combines new technologies with traditional ways of teaching. Though the participants of Maryniak's study prefer to read from a piece of paper rather than from a computer screen, the mode of text presentation had variable effects on comprehension.

Three chapters on academic writing in a foreign language reflect three different perspectives on this manifestation of literacy. The authors agree that writing in an academic context is demanding and difficult because of the need to satisfy formal and linguistic conventions, and present one's own viewpoints and opinions while reflecting critical thinking, originality

and creativity. Magdalena Trepczyńska argues that the writers' critical thinking, originality and creativity are best developed through a dialogic interaction with a teacher who provides supportive feedback. Małgorzata Adams-Tukiendorf takes a closer look at creativity in academic writing and thoroughly discusses creative potential against the background of her research findings. Jan Zalewski gives the perspective of an academic who is concerned about the inadequacy of his students' academic literacy skills, which may seriously impinge on their participation in the academic community and on their independent writing of master's theses. Using his personal experience he suggests solutions to the problem, bearing in mind exosemiotic and endosemiotic aspects of literacy.

Part Three addresses learners with special educational needs. Aleksandra Schwierz introduces and defines the problem of these needs and presents contemporary practices in the Polish educational system on the basis of the views and opinions collected from special education teachers. Joanna Nijakowska discusses the actions that teachers can take in order to include students with dyslexia into regular activities in a foreign language classroom setting. She also stresses the need to raise the teachers' awareness of the problems that dyslexic learners face in order to prepare them to offer appropriate supportive activities. Beata Wiechuła-Napiórkowska describes autistic spectrum disorder, its various manifestations, methods of coping with it, and first and second language acquisition by autistic children, many of whom overcome problems with communication and may become good second language learners, though to a limited extent.

The book market is rich in numerous publications that concern the topics included in this proposal. Yet, despite competition, the proposed work will be of value to scholars and libraries because of the contributors, the cross-sectional nature of the chapters and the applied framework of stability and variability.

Since this book represents a slice of foreign language learning reality at the beginning of the second decade of the twenty-first century, it may be valuable and inspiring reading for foreign and second language teachers, college and university students of language and education, as well as for researchers and scholars who never cease asking questions about the human condition, which is so strongly connected with the ability to successfully communicate across languages and cultures.

Ewa Piechurska-Kuciel
Liliana Piasecka

PART I:

FOCUS ON THE LANGUAGE

CHAPTER ONE

BETWEEN STABILITY AND VARIABILITY. A CORPUS-DRIVEN STUDY OF TRANSLATION UNIVERSALS: THE CASE OF POLISH TRANSLATIONS OF *LOLITA*

ŁUKASZ GRABOWSKI

Abstract

This article presents selected results of a corpus-driven comparison of the two independent Polish translations of the novel Lolita by Vladimir Nabokov (1955), which were completed by Stiller (1991) and Klobukowski (1997), respectively, with a custom-designed reference corpus of typical literary Polish. According to Xiao (2009), if characteristics of translational language are to be generalized as translation universals (after Baker 1993), the language pairs involved must not be restricted to English or closely related–genetically or typologically–languages. Therefore, this study aims to examine T-universals (after Chesterman 2004) with an emphasis on core patterns of lexical use–as proposed by Laviosa (1998)–on the example of the differences between translational, represented by the Polish translations of Lolita, and non-translational literary Polish.

1. Introduction

According to Baker (1995, 233), descriptive Translation Studies (DTS) conducted today should not be limited to the comparison of source texts and their translations, but they should also be extended to comparisons of non-translated texts with translated texts, which are produced under different social, cultural, and sometimes even political circumstances. In the same paper, Baker (1995, 243) puts forward the idea of *universal features of translation* or *translation universals*, which are specific textual characteristics (e.g. lexical, grammatical or stylistic) typical of translated texts, irrespective of the languages involved in the translation process. Furthermore, Baker (ibid.) posits a number of hypotheses on the

differences between translational and non-translational language, e.g. that translations tend to be, among other things, more explicit as regards lexis and syntax than non-translated texts, their content and form is simplified if compared with non-translated texts, and that language used in translation is more conventional and less creative than that used in non-translated texts. According to Kenny (2001, 53-54), translations exhibit a distribution of lexical items that distinguishes them from original texts in the same language, which accounts for a symptom of specific translation strategies or tendencies, such as, among others, explicitation, simplification, normalization, sanitization and levelling-out. According to Olohan (2004, 92), these patterns are specific to translations and are seen to be more typical of translational language than of non-translational language. In addition, characteristics of translational language are a product of constraints inherent in the translation process and do not vary across cultures (ibid.). Thus, it is essential to study linguistic patterns that are specific to translated texts, irrespective of source and target languages (Laviosa-Braithwaite 1995, 153). Finally, Kenny (2001, 54) claims that translation universals have predictive power, which follows that if one accepts that some type of lexical or stylistic characteristics constitutes a translation universal, it means that one may predict the said characteristics in instances or samples of translation that one has not yet encountered (Kenny 2001, 54).

Laviosa (1998, 557-570), who studied distinctive features of translational English as compared with native English (represented by samples ellicited from the British National Corpus), found that translational English has four core patterns of lexical use: a relatively lower proportion of lexical words over function words, a relatively higher proportion of high-frequency words over low-frequency words, a relatively greater repetition of the most frequent words, and a smaller vocabulary (i.e. lower number of word types) frequently used.

Consequently, a number of distinctive features of translational English in relation to native English have been uncovered. Nevertheless, if the core patterns of lexical use in translational language that have been identified by Laviosa (1998) are to be generalized as *translation universals*, a language involved in the study must not be restricted to English. This observation provided motivation to undertake a project that examines features of translational Polish.

Therefore, in this chapter, two independent Polish translations of the novel *Lolita* (henceforth 'PS' and 'PK', respectively) will be compared with the custom-designed reference corpus of typical literary Polish

(henceforth 'PLRC') to identify core patterns of lexical use, as well as detect traces, if any, of translation universals.

For the purposes of this study, a typology of translation universals [TUs] proposed by Chesterman (2004, 6-7) was applied. Chesterman (ibid.) distinguished between two types of TUs: the S-universals, which are related to translation from the source to the target language, and the T-universals, which are related to comparisons of translational and non-translational texts (i.e. target-language texts, which are not translations). In this chapter, which deals with comparison of translational and non-translational texts, the search for T-universals will be pursued.

2. Methodology, research material, tools and stages of analysis

In this study a bottom-up corpus-driven methodology was applied. In contrast to the corpus-based approach, which always works within commonly accepted frameworks of theories of language (which implies prior classification of linguistic data), the texts were not adjusted to fit any predefined categories or theoretical schemata. The study questions were addressed through empirical analysis of frequency distributions of words and recurrent patterns of language use as found in PK, PS and PLRC, which altogether constitute the research material.

Lolita (1955) is one of the best known novels by Vladimir Nabokov, which firmly established him as an outstanding American novelist. The first full and unabridged Polish translation of the novel appeared in 1991 (PS). It was completed by Robert Stiller who used, among other reference materials, the English-original and Russian self-translation by Nabokov (1967) as source texts (Stiller 1991, 435-436). The second full translation of the novel into Polish was completed six years later (1997) by Michał Kłobukowski (PK), who used the English-original version as the only source text. Nevertheless, Stiller (1997) added piquancy to Kłobukowski's translation. Immediately after the second translation of *Lolita* was released onto the market, Stiller published an article in the literary journal *Wiadomości Kulturalne* accusing Kłobukowski of glaring incompetence, plagiarism and stealing Stiller's translation.

For the purposes of this study, the two Polish translations were manually scanned and OCRed. Then the scanned texts were repeatedly proofread in order to ensure spelling accuracy, and they were further verified against the paper format versions. At that stage, any cases of

misrecognition of characters were corrected using a spellchecker, or the search-and-replace facility of a word processor.

Table 1. Make-up of Polish reference corpus (PLRC)

No	Author	Title, (Publication date), Symbol[1]	Size (tokens)
1	Krystyna Siesicka	*Zapach rumianku (1969),* KS_ZR_N	30517
2	Krystyna Siesicka	*Ludzie jak wiatr (1970),* KS_LW_N	34566
3	Jerzy Andrzejewski	*Ciemności kryją ziemię (1957),* JA_CK_N	34597
4	Jerzy Andrzejewski	*Bramy raju (1960),* JA_BR_N	24867
5	Witold Gombrowicz	*Ferdydurke (1938),* WG_FD_N	71265
6	Tadeusz Borowski	*Wybór opowiadań (1951),* TB_WO_N	58116
7	Stanisław Lem	*Pamiętnik znaleziony w wannie (1961),* SL_PZ_N	57576
8	Olga Tokarczuk	*Prawiek i inne czasy (1996),* OT_PI_N	55155
9	Dorota Terakowska	*Poczwarka (1988),* DT_PO_N	68810
10	Jerzy Pilch	*Bezpowrotnie utracona leworęczność (1998),* JP_BU_N	46606
	TOTAL		478504

When compiling the reference corpus of Polish literary language, the most important criteria were representativeness and size. As regards representativeness, the reference corpora represent typical literary Polish contained in the novels published around the time when the English-original (1955), Russian auto-translation (1967) and the two Polish translations (1991, 1997, respectively) were launched onto the book market. As a result, the PLRC includes texts that were published in the years 1937-2001. As for size, the reference corpora are approximately five

---

[1] These symbols will be used throughout multivariate analyses in section 4 of this paper.

times larger than study texts. According to Berber-Sardinha (2000, 7-13), who conducted a study aimed at estimating the ideal size of a reference corpus to be utilized in WordSmith Tools, a reference corpus that is five times larger than the study corpus generates a similar number of keywords as reference corpora that are 100 or 200 times larger. The structure of the PLRC is presented in Table 1 above.

Subsequently, a quantitative corpus-driven analysis was completed with the use of WordSmith Tools 4.0 developed by Scott (2004), which is a suite of programs custom-designed for text analysis. According to Hoover (2004, 517-533), the aim of quantitative approaches to literature is to represent elements or characteristics of literary texts numerically, applying the powerful, accurate, and widely accepted methods of mathematics to measurement, classification, and analysis. Furthermore, the availability of texts in electronic format has increased the attractiveness of quantitative approaches as innovative ways of reading amounts of text that overwhelm traditional modes of reading.

The study was broken down into two successive stages. First, the PS and PK were compared with the PLRC in terms of descriptive statistics, frequency profiles and frequency spectra. As a result, in this part of the study corpus linguistics research procedures were used. Secondly, the PS, PK and PLRC were compared in terms of the distance between 2000 of the most frequent words (henceforth 'MFW') in the said texts and the reference corpus. In this part of the study, multivariate methods (Principal Components Analysis and Cluster Analysis) were used to visualize the differences between the translations and the reference corpus.

3. Corpus-driven analyses

The corpus-driven analyses used in this study include comparisons of descriptive statistics, which present basic stylometric indicators of style (number of running words, i.e. text length, number of distinct words, i.e. vocabulary used, TTR and STTR, which are measures of lexical variety, number of sentences and length of sentences used). Next, the number and distribution of words of different length will be compared. The analyses end with the comparison of frequency profiles and frequency spectra, which enable one to gain insight into the distribution of top-frequency and bottom-frequency words, respectively.

3.1. Descriptive statistics

Descriptive statistics describe linguistic data in quantitative terms, which are commonly accepted in corpus linguistics as basic indicators of style and lexical richness (Olohan 2004, 78-81). Hence, it provides a holistic view of the two Polish translations and the reference corpus PLRC, whose numerical characteristics are presented in Table 2.

Table 2. Descriptive statistics for EN (English-original version of *Lolita*) PS, PK (Polish translational texts) and PLRC (Polish non-translational literary texts)

Statistics	ENL	PS	PK	PLRC
Number of tokens	112230	101130	95936	478504
Number of types	13991	28757	28879	70722
Type/token ratio (TTR) (%/or *x* per 100 tokens)	12.46	28.44	30.10	14.77
Standardized TTR (STTR) (in %/or *x* types per 100 tokens)	**51.67**	**66.07**	**70.04**	**60.40**
Mean word length (in characters)	**4.40**	**5.50**	**5.66**	**5.32**
Number of sentences	5549	5628	5529	42936
Mean sentence length (in tokens)	**20.22**	**17.96**	**17.35**	**11.14**
Mean sentence length standard deviation	**20.63**	**18.97**	**17.75**	**23.37**

As the reference corpora are approximately 5 times larger in terms of the number of tokens than the novels, it is necessary to filter out those quotients that are not sensitive to differences in size (i.e. STTR, mean word length, mean sentence length, sentence length standard deviation). These are printed in bold in Table 2.

In general, STTR provides brief information as to the variety of the vocabulary used in a particular text or corpus, i.e. the complexity/simplicity or specificity/generality of the vocabulary (Baker 1995, 236, Kenny 2001, 34). More specifically, low STTR indicates narrow range of vocabulary, or narrow scope of subjects discussed in a

text or corpus, while high STTR indicates the opposite. The STTR value shows that on average in PS and PK there are 66 and 70 word types per 100 tokens, respectively, whereas in PLRC there are only 60 word types per 100 tokens. This perfunctory measure of lexical richness shows that translational texts represented by PS and PK are more complex and specific lexically and have fewer lexical repetitions as compared with PLRC.

As far as the mean word length is concerned, the data generated for PS, PK (5.50 and 5.66, respectively) and PLRC (5.32) provide no conclusive evidence of conspicuous differences between translational and non-translational texts. Moreover, the mean word length within the range 5.32-5.66 corresponds with the findings of the study conducted by Przepiórkowski (2006, 5), who revealed that in the Sample Corpus of Polish (size: 12,198,241) the most frequent word tokens were the ones with 5 and 6 letters, which further confirms the lack of significant differences between the texts and reference corpus.

As regards the mean sentence length, there are pronounced differences between PS and PK (17.96 and 17.35, respectively), on the one hand, and the PLRC (11.14), on the other. This discrepancy enables one to formulate a number of hypotheses. Firstly, the fact that on average sentences in the translations are 6 tokens longer than in the reference corpus shows that their style is more explicit and precise as compared with concise and terse sentences in PLRC. Secondly, longer sentences in PS and PK indicate that they are more lexically varied (which is also confirmed by higher STTR) and that there is a higher information load therein. However, the mean sentence length standard deviation shows that the PLRC has a less uniform distribution of sentences as, on the one hand, the mean sentence length is 11.14, but on the other hand, the high standard deviation of 23.37 signals that among these relatively short sentences one may identify extremely long ones. Overall, one may hypothesize that the style of the translations is more uniform in terms of length of sentences, while the style of typical literary texts in PLRC is more varied (i.e. one is bound to find there both conspicuously short and long sentences). However, the high mean sentence length in PS and PK may also be due to the influence of the style of the English-original (EN) version of the novel (with mean sentence length of 20.22 and mean sentence length standard deviation of 20.63). Thus, it may be the case that the translators refrained from the use of sophisticated syntactic translation strategies (e.g. the number of sentences in EN, PS and PK is very similar), such as transforming complex and compound sentences into shorter single sentences, and therefore did not interfere in the length of sentences translated from English into Polish. As

a result, long sentences attested in PK and PS do not conform to the typical literary Polish found in PLRC.

To sum up the analysis and interpretation of descriptive statistics, one may hypothesize that both PS and PK (translated texts) are, on average, more complex lexically and have more explicit and precise sentences than the reference corpus (non-translated texts)–the characteristics that may be construed as a T-universal. However, further and more detailed qualitative research is essential to bring to life concrete illustrations of the above differences and find the rationale underlying them, i.e. are they due to the author's style, or translator's style? Or are they due to interference from English and thus the direction of translation?

3.2. Word length

Although the data on the mean word length (i.e. 5.50 and 5.66 for PS and PK, respectively, and 5.32 letters for PLRC) provided no conclusive evidence of conspicuous differences between translational and non-translational texts, the issue of distribution of words of different lengths (henceforth 'n-letter words') has not yet been addressed. According to Kenny (2001, 127-134), creative and author-specific vocabulary is usually found among longer, multi-letter words rather than shorter ones, where usually grammatical (function) words prevail (i.e. they are repeated more often in a text). Obviously enough, the study conducted by Kenny (2001) focused on German to English translations of contemporary literary texts and its aim was to determine whether translators use more conventional word-forms in the target language to render lexically creative source-text forms. As a result, the aforementioned claim on the length of creative vocabulary is relevant to English and German language material. Nevertheless, since translation universals account for a rationale behind the present study, the said hypothesis should be extended to also cover Polish language material. As a result, comparison of word length may constitute provisional means to study lexical richness in texts (provided that the differences observed are statistically significant). Thus, the following Figure 1 illustrates distribution of n-letter words (1-22 letters long) and their proportion in the total word count in PS (dotted line), PK (broken line) and PLRC (straight line). It also enables one to see if translated texts (PS and PK) conform to the typical patterns found in PLRC.
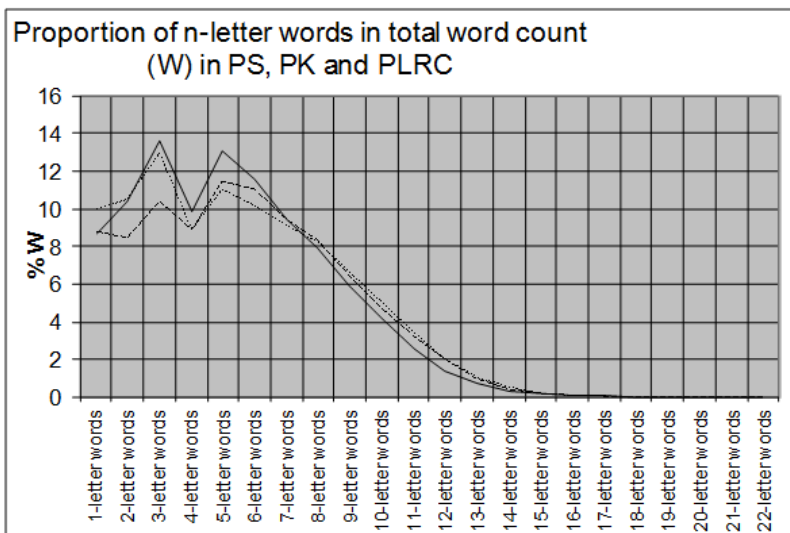
Figure 1.

Proportion of n-letter words in total word count (W) in PS, PK and PLRC

Figure 1 reveals the discrepancy as regards the frequency of n-letter words across PS, PK (translated texts) and PLRC (where non-translated texts were merged into one sample). The noticeable characteristic of PS is that the number of 5-letter words is considerably lower than the number of the most frequent 3-letter words, which is different than the data attested in PLRC (there, 5-letter words are almost as frequent as 3-letter words). PK, on the other hand, is unique in that 5-letter words are the most frequent there, followed by 6-letter words. As a result, it is PK that to a large extent does not resemble a typical Polish literary text, and it is considerably different than PS (which may be surprising in that both PS and PK convey–at least theoretically–the same information). Nevertheless, all PS, PK and PLRC share the characteristic feature observed in any Polish text, namely a specific bend at 4-letter words, which occur with considerably lower frequencies than neighbouring 3- and 5-letter words. Thus, it is almost a rule (however hilarious it may sound) that speakers of Polish, including both writers and translators, have the tendency to avoid 4-letter words, which are rather infrequent in Polish[2].

---

[2] Przepiórkowski (2006, 5) found the same tendency in the analysis of frequencies of n-letter words in the so-called Sample Corpus of Polish (size: 12,198,241 word tokens), which is a sub-corpus of larger IPIPAN Corpus of Polish. The said Sample Corpus represents

More detailed comparison of the number of n-letter words, i.e. a glimpse beyond their proportion in the total word count, revealed further irregularities in terms of the words' frequencies. In order to find whether different frequencies of n-letter words in PS, PK and PLRC are only random or whether they are statistically significant (i.e. they are caused by some other factor than chance), Dunning's (1993) log-likelihood test at the probability value p=0.0001[3] was used. To that end, the modified LL calculator (after Rayson 2000[4]) generated the log-likelihood (LL) scores presented in Table 3 (PS vs. PLRC) and Table 4 (PK vs. PLRC) below.

Table 3. LL scores calculated for frequencies of n-letter words in PLLS and PLRC (p=0.0001, d.f.=1, critical value=15.13)

n-letter words	Observed frequencies		Expected frequencies		LL score	Over-used in
	PS	PLRC	PS	PLRC		
1-letter words	10091	41462	8994	42558	**156.99**	**PS**
2-letter words	**10580**	**49581**	10496	49664	0.80	
3-letter words	**13167**	**65078**	13651	64593	**21.05**	**PLRC**

contemporary Polish and consists of scientific texts (10%), contemporary artistic prose (10:6%), older (late-nineteenth and early-twentieth century) artistic prose of the kind read at schools (9:7%), legal texts (4:9%), transcripts of parliamentary sessions (15:5%) and various newspaper texts (49:3%). Przepiórkowski (ibid.) revealed that the most frequent word tokens in the Sample Corpus are 6-letter-long, followed by 5- and 3-letter-long words. As regards 4-letter words, these are much less frequent as they occur even less often than 1-, 2-,7- and 8-letter-long words in the Sample Corpus of Polish.

[3] The value of p=0.0001 suggests 0.01 % danger of being wrong (1 chance in 10,000). As a 2x2 contingency table is used to compare frequencies of the n-letter words in PS vs. PLRC and PK vs. PLRC, a degree of freedom (d.f.) equals 1, which means that according to chi-square distribution table (which is the same for log-likelihood test) one can reject the null hypothesis whereby there is no statistically significant difference between observed frequencies in PS and PLRC or PK and PLRC only if the critical value of log likelihood exceeds 15.13 (at 1 degree of freedom).

[4] The log-likelihood calculator developed by Dr Paul Rayson (UCREL University Centre for Computer Corpus Research on Language, Lancaster University, UK) is available at: <http://lingo.lancs.ac.uk/llwizard.html>. The said calculator was adapted for the purposes of this research in order to calculate a large number of results.

4-letter words	8958	47347	9823	46481	**94.68**	**PLRC**
5-letter words	**11152**	**62882**	12917	61116	**303.49**	**PLRC**
6-letter words	**10319**	**55651**	11510	54459	**153.61**	**PLRC**
7-letter words	9222	45251	9504	44968	10.23	
8-letter words	8363	37924	8075	38211	12.25	
9-letter words	6670	28023	6053	28639	**74.23**	**PS**
10-letter words	5129	19712	4334	20506	**168.82**	**PS**
11-letter words	3498	12433	2779	13151	**211.47**	**PS**
12-letter words	2077	6822	1552	7346	**198.30**	**PS**
13-letter words	1061	3287	758	3589	**133.32**	**PS**
14-letter words	472	1562	354	1679	**43.35**	**PS**
15-letter words	228	832	184	875	11.48	
16-letter words	85	354	76	362	1.09	
17-letter words	39	174	37	175	0.11	
18-letter words	12	72	14	69	0.61	
19-letter words	4	40	7	36	2.50	
20-letter words	2	4	1	4	0.88	
21-letter words	1	4	0	4	0.02	
22-letter words	0	3	0	0	0	
TOTAL	101130	478504				

Table 4. LL scores calculated for frequencies of n-letter words in PLLK and PLRC (p=0.0001, d.f.=1, critical value=15.13)

n-letter words	Observed frequencies		Expected frequencies		LL score	Over-used in
	PK	PLRC	PK	PLRC		
1-letter words	8938	41462	8416	41983	**38.10**	**PK**
2-letter words	8524	**49581**	9703	48401	**178.15**	**PLRC**
3-letter words	**10504**	**65078**	12622	62959	**447.80**	**PLRC**
4-letter words	9024	47347	9414	46956	**19.63**	**PLRC**
5-letter words	**11553**	**62882**	12430	62004	**75.89**	**PLRC**
6-letter words	**11147**	**55651**	11155	55642	0.01	
7-letter words	**9509**	45251	9145	45614	**17.20**	**PK**
8-letter words	8402	37924	7736	38589	**67.19**	**PK**
9-letter words	6524	28023	5769	28777	**114.59**	**PK**
10-letter words	4722	19712	4080	20353	**116.35**	**PK**
11-letter words	3264	12433	2621	13075	**178.11**	**PK**
12-letter words	1986	6822	1470	7337	**199.37**	**PK**
13-letter words	990	3287	714	3562	**116.81**	**PK**
14-letter words	462	1562	338	1685	**50.11**	**PK**
15-letter words	210	832	174	867	8.49	
16-letter words	91	354	74	370	4.26	

17-letter words	44	174	36	181	1.81	
18-letter words	21	72	15	77	2.13	
19-letter words	9	40	8	40	0.10	
20-letter words	5	4	1	7	6.99	
21-letter words	3	4	1	5	2.64	
22-letter words	2	3	0	4	1.53	
TOTAL	95936	478504				

Summing up the results presented in Tables 3 and 4 above, one is made to conclude that 1-letter-long words as well as 7-, 8-, 9-, 10-, 11- 12-, 13-, 14-letter-long words, which are the long ones, are significantly more common in translated texts represented by PS and PK, whereas shorter 2-, 3-, 4-, 5- and 6-letter words prevail in the PLRC. Since according to Kenny (2001, 127-134), creative vocabulary is usually found among longer words, one is forced to conclude that translated texts (PS and PK) are lexically richer than non-translated texts attested in PLRC. Therefore, the T-universal related to lexical simplification typical of translated texts is invalidated in this particular case of the two Polish translations of *Lolita*.

3.3. Frequency profiles

In order to determine whether it is translational language (PS and PK) or typical literary Polish attested in PLRC (non-translational language) that has more repetitions and lower lexical variety in terms of top-frequency words, a frequency profile proposed by Baroni (2009, 805-806) was used. As a rule, the frequency profile is obtained by a replacement of words in a frequency list (which was completed with the use of WordSmith Tools 4.0) with their frequency-based ranks, by assigning rank 1 to the most frequent word, rank 2 to the second most frequent word, rank 3 to the third most frequent word etc. This enables one to answer the question of which frequency-based ranks (r) of words (tokens) have a particular frequency (f). However, a typical frequency profile was modified in that frequency information was substituted with information on the cumulative percentage of the total word count (%cW)

corresponding to frequency-based ranks. The results are presented in Table 5.

Table 5. Frequency profiles for PS, PK and PLRC

PS		PK		PLRC	
Rank	%cW	Rank	%cW	Rank	%cW
1-100	36.07	1-100	32.11	1-100	35.16

The data show that translational texts (PS and PK) and PLRC (typical literary Polish) feature no significantly different distribution of top-frequency words. As a rule, the higher the value of a frequency profile for the top ranks, the less lexically varied and more repetitive the text is. As a result, the only difference revealed in this analysis concerns the two translations–top-frequency words occur more frequently in PS, and hence PK is less repetitive and more lexically varied as regards top-frequency words. Nevertheless, there is no evidence to confirm that translational texts (PS and PK) have relatively greater repetition of the most frequent words or that a smaller vocabulary is frequently used therein than in typical Polish literary texts as attested in PLRC. Therefore, these core patterns of lexical use, or T-universals, were dismissed.

3.4. Frequency spectra

According to Baroni (2009, 806), frequency spectra enable one to determine how many word types (w) in a frequency list have a particular frequency [w (f)]. As creative or author-specific vocabulary usually occurs in a text with low frequencies, frequency spectra can be used to study lexical variety and degree of repetitions among bottom-frequency words. As a rule, a text is more varied lexically if proportion of bottom-frequency words in the total word count (%W) is higher. For the purposes of this study, a number of word types (w) corresponding to a particular frequency (f) in the frequency spectra was substituted with information on the cumulative percentage of the total word count (%cW) corresponding to word types with frequencies 1-25. The results are presented in Table 6.

Table 6. Frequency spectra for PS, PK and PLRC

PS		PK		PLRC	
w (f)	%cW	w (f)	%cW	w (f)	%cW
1-25	53.84 %	1-25	57.31 %	1-25	37.2%

The data show that both PS and PK (translated texts) have higher lexical variety among bottom-frequency words (i.e. words with frequencies 1-25) than PLRC (non-translated texts). Such a difference translates into more hapax legomena, dis legomena and tris legomena (which are words that occur in a text only once, twice or three times, respectively) used in the translated texts, where–according to Kenny (2001)–one may identify creative and author-specific vocabulary. It follows that a T-universal whereby translational Polish (represented by PS and PK) is more varied lexically in terms of bottom-frequency words than typical literary Polish (PLRC) was revealed.

4. Multivariate analyses

The analyses above focused on the data sets with only two vectors presented in a tabular form, i.e. specific observations (numbers) in the rows, with the vectors (column variables) specifying different properties of the observations (e.g. 11.14 being a mean sentence length in PLRC). This part of the study will attempt to focus on data sets with more than two vectors. More specifically, it will attempt to determine if the translational texts of *Lolita* (PS and PK) are similar to or different from typical non-translational literary texts (PLRC) in terms of the differences (i.e. distance) between distribution and frequency of 2000 the most frequent words (MFW) in each of the texts. In other words, the aim of the following research procedures is to measure and visualize the distance between translational and non-translational texts (i.e. the shorter the distance, the more similar the texts).

Studying frequencies and distributions of word types and word tokens in texts, Burrows (1987) and Baayen (2001), among others, questioned the idea that word-tokens appear randomly in texts. Using multivariate analyses, more specifically, multidimensional methods, they showed that there is a powerful 'force-field' attached to each occurrence of a word.

The distance between texts will be measured with the use of a Delta method developed by Burrows (2002), which is a simple measure of the