Quantifying English and Polish *Lolitas*: A Corpus-Driven Stylistic Comparison

Łukasz Grabowski

Abstract The study presented in this article,which is a fragment of a larger study of translational and non-translational texts (Grabowski 2012), falls within the scope of descriptive translation studies (DTS) and corpus linguistics, with particular emphasis on the study of translation universals, on the example of English-original (written in 1955) and two independent Polish translations of the novel *Lolita* by V. Nabokov (by Stiller in 1991 and Kłobukowski in 1997). According to Baker (1995: 243), *universal features of translation* or *translation universals*, constitute specific textual characteristics (e.g. lexical, grammatical or stylistic) typical of translated texts, irrespective of languages involved in the translation process. In this study, which was completed with the use of corpus linguistics methodology, the texts were compared in terms of basic stylometric indicators presented through descriptive statistics, top-frequency wordlists, frequency profiles and frequency spectra. More specifically, the analysis aimed to compare the English-original and two Polish translations of *Lolita* in terms of text length, sentence length, number of repetitions (conciseness of style) as well as frequencies and distribution of both word-types (distinct words) and word-tokens (running words). Also, the aim was to find traces, if any, of translation universals (S-universals, after Chesterman 2004) attested in the Polish translations. The article concludes with suggestions as to research on translation universals in literary texts with the use of corpus linguistics methodology.

1 Introduction: English-Original and Polish Translations of *Lolita*

Lolita (Nabokov 1955) is one of the best known novels by Vladimir Nabokov, which firmly established him as an outstanding American novelist. Due to its

Ł. Grabowski (✉)
Opole University, Opole, Poland
e-mail: lukasz@uni.opole.pl

K. Piątkowska and E. Kościałkowska-Okońska (eds.), *Correspondences and Contrasts in Foreign Language Pedagogy and Translation Studies*,
Second Language Learning and Teaching, DOI: 10.1007/978-3-319-00161-6_14,

highly controversial–at that time–subject matter, Nabokov was unable to find the publisher of the novel in the United States of America, and instead the book was published in France in 1955 by Olympia Press. The first American edition was issued by G. P. Putnam's Sons Publishing House in New York only in 1958 (Boyd 1995: xlv). The novel is written in a highly artistic, masterful and precise style, which made Nabokov one of the most brilliant and idiosyncratic stylists of English (Stiller 1991: 421).

The history of translation of *Lolita* into Polish has been quite turbulent. The first attempt to translate the novel was undertaken in 1958 in Tel-Aviv, where a translation produced by an anonymous journalist made the papers of the Polish-language weekly *Przegląd*. According to Stiller (1991: 434), the author of this abridged version—which covered only three fourths of the length of the English original—was a journalist Moshe Balsam. Stiller (1991: 434) further writes that in the years to follow, there were more fragments of the novel translated into Polish, which made it to the papers, e.g. to the weekly *Przekrój*, where fragments of the novel translated by Juliusz Kydryński came out in 1959; to the Polish émigré weekly *Wiadomości* printed in London (in 1961) translated by Jerzy Tepa; to the *Odra* weekly with the fragments of the novel translated from Russian by Eugenia Siemaszkiewicz (in 1974); to the weekly *Tygodnik Kulturalny,* where the first seventeen chapters of the book translated by Robert Stiller were printed in 1987. The first full and unabridged Polish translation of *Lolita* appeared in 1991, and it was completed by Robert Reuven Stiller.

The translation by Stiller is accompanied by an extensive commentary concerning the project (Stiller 1991). The translator claims that four reference materials provided the basis for his translation, namely: (1) *The Annotated 'Lolita'* by A. Appel, Jr. and Nabokov, which is an annotated text of the novel accompanied by commentaries and notes, which provide further explanations of Nabokov's referrals, puns, archaic, foreign and invented words etc. as they appear in the English *Lolita*; (2) *Keys to 'Lolita'* by Carl A. Proffer, which is another extensive commentary on the novel; (3) the original English language version of *Lolita*, and (4) the Russian self-translation of the novel by Nabokov. As a result, Stiller's translation is based on both English and Russian language versions of the novel (Stiller 1991: 435–436).

The second full-version translation of the novel into Polish was completed six years later by Michał Kłobukowski (i.e. in 1997). Nevertheless, Stiller (1997) added piquancy to Kłobukowski's translation. Immediately after the second full-version translation of *Lolita* was released onto the market, Stiller (1997: 6–7) published an incriminating article in the literary journal *Wiadomości Kulturalne* accusing Kłobukowski of glaring incompetence and plagiarism of his own translation of *Lolita*.

Thus, it is believed that the tempestuous past of the English-original of *Lolita* as well as a stormy verbal duel between the two Polish translators of the novel make this particular text even more interesting object of a comparative study.

2 Universal Features of Translation

In her seminal 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 languages involved in the translation process. Further, Baker posits a number of hypotheses on the differences between translational and non-translational language, e.g. that translations tend to be, among others, 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 the one used in non-translated texts (Baker 1995).

In the same vein, Kenny (2001: 53–54) claims that translations exhibit 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 one. In addition, characteristics of translational language are a product of constraints inherent in the translation process and do not vary across cultures (Olohan 2004: 92). Thus, it is essential to study linguistic patterns which are specific to translated texts, irrespective of source and target languages (Laviosa-Braithwaite 1995: 153). Finally, Kenny (2001: 54) hypothesizes 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).

Therefore, in this study, the English-original version of the novel *Lolita* (henceforth 'ENL') will be compared with its two independent Polish translations (henceforth 'PLS' and 'PLK', respectively) to identify the differences as regards text length, sentence length, number of repetitions (conciseness of style), and to find 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 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 article, which deals with comparison of the English source-text and its two Polish translations, the search for S-universals will be pursued.

3 Methodology, Research Material, Tools and Stages of the Analysis

In order to provide answers to the aforementioned study questions, the corpus-driven methodology was applied. In contrast to the corpus-based approach, which always works within commonly accepted frameworks of theories of language, or—in other words—is theoretically-committed (which implies prior classification of linguistic data), the English-original and Polish translations of *Lolita* were not adjusted to fit any predefined categories or theoretical schemata. Thus, the study questions were addressed through empirical analysis of frequency distributions of words and recurrent patterns of language use as found in the aforementioned texts. As a result, the novels were compared through bottom-up observation of empirical linguistic data, which were presented in quantitative terms and, where necessary, supplemented with qualitative observations.

According to Hoover (2004: 517–533), the aim of such 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 Hoover (2004: 517-533). It is therefore believed that quantitative approaches, such as the corpus-driven one presented in this study, enable one to study translational style and its variations from a different perspective, and to put forward more fine-grained hypotheses or research questions to be addressed in qualitative studies in the future.

The texts used in the analysis, i.e. the English-original as well as its two Polish translations, were purchased in bookstores in paper format and they were further converted into machine-readable format supported by the software used throughout the study. To that aim, the texts were manually scanned and subjected to the OCR procedure. The scanned texts were then subjected to repeated proof-reading 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 edited and corrected using a spellchecker, or a search-and-replace facility of a word processor. Finally, the texts were saved in two files in a plain text format.

The corpus-driven analysis conducted in this study was facilitated by the use of the computer software WordSmith Tools 4.0 developed by Scott (2004), which is a suite of programs custom-designed for text analysis.

4 Corpus-Driven Comparison of Stylometric Indicators

The corpus-driven analyses used in this study encompass comparisons of descriptive statistics, which presents 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). The study ends with the comparison of frequency profiles and frequency spectra, which enable one to gain an insight into distribution of top-frequency and bottom-frequency words, respectively.

4.1 Descriptive Statistics

Descriptive statistics describes linguistic data in quantitative terms, and present basic indicators of style and lexical richness (Olohan 2004: 78–81). Hence, it provides a holistic view of the English-original of *Lolita* and its two Polish translations by Stiller and Kłobukowski (ENL, PLS, PLK, respectively). Their characteristics are presented in Table 1.

Hence no lemmatization was conducted on either ENL, PLS or PLK, the indicators such as the number of types, TTR and STTR are inflated for the Polish translations,[1] and thus impossible to serve as the basis for comparison. It is due to the fact that the texts represent typologically different languages, i.e. English, which is highly-analytical, and Polish, which is more synthetic as regards morphology. Nevertheless, as Sinclair (1991: 8) claims that each distinct inflectional form is potentially a unique lexical unit, the issue of non-lemmatization was ignored and the study focused on the remaining indicators, which are relevant and valid irrespective of typological differences between the two languages.

As far as the length of the original and the translations, one may arrive at the following conclusions. Firstly, the data show that both Polish translations are shorter than the source-text in terms of the number of running words, or tokens

Table 1 Descriptive statistics for ENL, PS and PK

Statistics	ENL	PLS	PLK
Number of tokens (text length/size in running words)	112,230	101,130	95,936
Number of characters (text length/size in characters or bytes)	1,261,546	1,370,082	1,331,058
Number of types (distinct words)	13,991	28,757	28,879
Mean frequency of a type	8,02	3.51	3.32
Type/token ratio (TTR)	12.46	28.43	30.10
Standardised TTR (STTR)	51.66	66.07	70.03
Standardised TTR std. dev. (STTRstd)	47.10	32.91	28.82
Standardised TTR basis	1,000	1,000	1,000
Mean word length (in characters)	4.40	5.50	5.66
Word length std.dev.	2.39	3.18	3.12
Number of sentences	5,549	5,628	5,529
Mean sentence length (in words)	20.22	17.96	17.35
Mean sentence length std.dev.	20.63	18.97	17.75

[1] Any lexeme is a set of inflectional forms, and each of these word-forms is treated as a separate word type, which overall inflates TTR and STTR for highly-inflectional languages. Obviously enough, even if lemmatization was conducted, it would not solve the problem of disambiguation.

(i.e. 112,230 versis 101,130 and 95,936 in ENL, PLS and PLK, respectively, which yields the ratio of original-to-translation at 1.11 and 1.17). In other words, Stiller required—on average—901 words to translate 1,000 English words in the original; Kłobukowski, on the other hand, required only 854 words to do the same. On the surface, this finding contradicts the hypothesis on explicitation in translation. According to Nida and Taber (1974: 163), one should expect translation to be longer than original text because translators tend to explicitate phenomena which are non-existent in the language of translation. This assumption has not been validated in this study.

On the other hand, if one takes into consideration the size of ENL, PLS and PLK measured in characters, the results are the opposite—the English-original is shorter than both Polish translations (1,261,546 versus 1,370,082 and 1,331,058 characters in ENL, PLS and PLK, respectively). It yields the original-to-translation ratio at 0.92 in the case of PLS, and at 0.94 in the case of PLK.

Overall, this case shows that comparison of length of texts written in different languages on the basis of the number of running words is misleading; the number of characters, including letters, digits, punctuation and spaces, constitutes a more reliable indicator in such comparisons (Mikhailov 2003: 167), particularly when one compares texts written in typologically or genetically unrelated languages.

The answer to this discrepancy is to be searched in typological differences concerning morphology. The frequent use of articles in the English language means that the number of running words in any English text is higher than in the translation into a language without articles, which is the case of Polish. On the other hand, it is dubious that every utterance in English is longer than an analogical utterance in Polish (particularly in a translation situation involving real texts). An important observation, however, refers to the fact that a synthetic language (such as Polish) has more synthetic (i.e. longer) word forms used, while a more analytical language, which is in this case English, has less synthetic word forms (i.e. shorter), which is due to poorer inflection. This difference is reflected in the mean word length, which accounts for 4.40 characters in ENL and 5.50 and 5.56 characters in PLLS and PLLK, respectively.

Therefore, one is made to conclude that Polish translations of *Lolita* are longer than its English original. However, it remains a debatable issue whether this pattern is typical of any English-to-Polish translation in general. The findings of a number of stylometric studies of originals and translations (Englund Dimitrova 1993, 1994; Mikhailov and Villikka 2001; Mikhailov 2003; Scarpa 2006; Rybicki 2007) show that the length of translation as compared with its source-text varies depending on language pairs and a direction of translation. Also, Baker (2000) suggests that this variation in original-to-translation ratios is due to translators' individual styles or idiolects. As a result, further studies[2] conducted on larger

[2] Apart from using larger corpora, which translates into more statistically significant results, it is possible to approach text length from mathematical perspective, e.g. using **entropy**, which is the quantity that measures information in texts (Oakes 1998: 58–60; Mikhailov 2003: 169).

parallel English-Polish corpora, containing texts representing different genres and types, are necessary to validate the universalist claim that Polish translations from English tend to be longer than their source-texts.

Also, the data presented in Table 1 show that the translation by Stiller is considerably longer than the one by Kłobukowski (by 5,194 running words or 39,024 characters). Further, taking into consideration the fact that the shorter translation by Kłobukowski has a higher number of word types than the longer translation by Stiller, one can conclude that PLS has more repetitions than PLK. This observation is further corroborated by the mean frequency of a word type, which is higher in PLS (3.51 versus 3.32 in PLK). Eventually, it shows that PLK has higher lexical density than PLS.

As regards lexical density measured by the STTR, the data show that Kłobukowski's translation is lexically richer than Stiller's translation. On average, there appear 700 word types per 1,000 word tokens in PLK, whereas in the case of PLS there are 660 tokens. It means that PLK is more complex and specific lexically and has fewer repetitions as compared with PLS.

The data on the number of sentences (5,628 versus 5,529 in PLS and PLK, respectively) and the mean sentence length (17.96 versus 17.35 in PLS and PLK, respectively) show that Stiller used 99 more sentences, which at the same time are slightly longer than the ones used by Kłobukowski. Further, the fact that Stiller uses 5,194 more words and longer sentences in the translation can mean that Kłobukowski's translated sentences are more concise and terse as compared with more explicit Stiller's sentences. On the other hand, the number of sentences in the English-original (5,549) shows that Kłobukowski was more consistent in translating in sentence-for-sentence fashion, whereas Stiller exhibited more flexibility in this respect. Overall, there are 79 more sentences in PLS than in ENL. Such a manipulation on the number of sentences on the part of Stiller is further confirmed by a higher value of the mean sentence length standard deviation in PLS (18.97 as compared with 17.75 in PLK). Thus, it is possible to put forward the hypothesis that Stiller's translated sentences are more explicit and precise as compared with Kłobukowski's more concise and terse sentences.

Taking into consideration the mean sentence length in the English-original version of the novel, which is 20.22 tokens, the corresponding figures for PLS and PLK show that both translators employed faithful sentence-for-sentence translation and used long-form constructions to translate the novel (Stiller, in particular). As the mean sentence length for the Polish prose is 11.90 tokens (Ruszkowski 2004: 34),[3] the data show that both Stiller's and Kłobukowski's sentences are untypical and differ from the ones in the non-translational texts, i.e. typical Polish novels.

[3] It requires clarification that Ruszkowski provided data regarding average utterance length. However, since the author adopted orthographic criterion regarding segmentation of utterances (Ruszkowski 2004: 30–32) in his study the very term utterance is therefore equivalent with the sentence.

4.2 Comparison of Wordlists

In order to compare ENL, PLS and PLK in terms of type, range and distribution of the most frequently used vocabulary, the wordlists were generated for these three texts. As a rule, wordlists highlight top-frequency grammatical words, which means that it is difficult to identify any lexical differences between the original and the two translations, which can be markers of translators' style. To remedy this inconvenience, grammatical words were deleted from the top-frequency items, and the most frequently used lexical (content) words in ENL, PLS and PLK are presented instead. Such a filtered-out wordlist with 25 top-frequency lexical words is presented in Table 2.

As, at least hypothetically, the three texts convey the same information, it is no surprising the most content words overlap in the source-text and its translations. These words include, among others, names of protagonists (*Lolita, Lo, Charlotte, Humbert*). However, the data also show that some differences between the source-

Table 2 Wordlists with top-frequency content words in ENL, PS and PK

ENL			PLS			PLK		
R*	Word	Freq	R*	Word	Freq	R*	Word type	Freq
9	WAS	1486	21	JUŻ	365	21	JUŻ	275
30	HAVE	388	29	LO	236	22	LO	274
37	SAID	344	33	JEST	219	28	JEST	228
39	WERE	305	37	JESZCZE	201	35	JESZCZE	202
42	LITTLE	287	43	BYŁO	181	42	BYŁA	166
51	LO	236	47	BYŁA	168	46	BYŁO	156
61	OLD	197	49	BYŁ	158	49	BYŁ	152
62	LOLITA	193	59	HAZE	135	57	HAZE	124
63	TWO	192	63	LOLITA	125	58	BARDZO	120
81	KNOW	139	71	PAN	112	64	POTEM	108
84	HAZE	137	80	BARDZO	99	69	DOLLY	100
87	WAY	135	81	HUMBERT	99	72	RAZ	99
89	CHILD	132	84	ZNÓW	98	76	LOLITA	96
93	ROOM	121	85	DOLLY	96	79	NIGDY	94
95	GIRL	120	98	RAZ	83	80	MA	92
100	AM	117	102	JESTEM	80	87	HUMBERT	87
101	CAR	117	103	DOMU	79	93	DOMU	82
102	GOOD	116	104	LAT	78	97	PAN	82
103	HUMBERT	115	107	WCIĄŻ	77	101	JESTEM	80
108	EYES	109	109	CZASU	73	102	MIAŁA	80
113	HAND	104	111	CHARLOTTE	72	110	MAM	76
114	MADE	104	112	MIAŁA	72	112	LAT	73
115	DAY	103	114	NIGDY	72	114	TERAZ	73
116	FIRST	103	118	DWA	69	117	LOLITY	71
120	LET	98	120	WRESZCIE	69	119	BYĆ	68

*R: rank of a word on a frequency list

text and its translation result from typological differences between language systems of the two languages, e.g. more analytical English morphology inflates frequencies of the most frequently used word types as compared with their lower values for Polish texts. For example, the high frequency (1,791 in aggregate) of the verbs *was, were* and *have* in ENL results from their functioning not only as inflectional forms of the verbs *to be* and *to have,* but also from being auxiliary verbs used in multiple grammatical tenses. It explains their higher frequency as compared with aggregated frequency (532 and 498 occurrences in PLS and PLK, respectively) of the corresponding verb forms in Polish, e.g. *było, była, był, byli, byłaś, byłeś, byliśmy, byliście, byłyśmy, byłyście*. Also, one may notice the high frequency of broad-meaning English verb forms, such as *said* and *made*, which do not have their potential equivalents in PLS and PLK among top-frequency content words.

The above examples also refer to one of specific problems of translation between English and other Slavic languages (e.g. Polish or Russian). Extending the assumption made by Comrie (1981: 31–79) with reference to Russian, it seems that the Polish language is more explicit semantically (i.e. words have more specific meaning distinctions) than English, which in turn is more ambiguous and vague in its surface forms. Hence, English largely depends on pragmatic and contextual information in specifying exact interpretation of its linguistic forms (e.g. a past tense reporting verb *said*), which are broad in meaning. According to Piotrowski (1994: 95–96), although the English language has both broad-meaning and specific lexemes, users of English tend to choose the ones with broad meaning rather than specific. Users of Russian and other Slavic languages, on the other hand, tend to choose specific lexemes, and that is the reason why they regard texts with multiple repetitions as ones with plain, simple, or even bad style (Piotrowski 1994: 96). As regards translation, the outcome can be that translation of English reporting verbs (or broad-meaning English lexemes in general) requires that more lexical words be used in Polish to produce a natural and acceptable translation.

Table 2 also reveals some characteristic features of the Polish translations. It shows that most top-frequency lexical words overlap in PLS and PLK. The exceptions to these are words such as *znów* (‘again’), *wciąż* (‘still’), *czasu* (‘time’, singular genitive case), *Charlotte*, *dwa* (‘two’), which are over-represented in PLS, whereas the words, such as *potem* (‘after’), *ma* (‘has’), *teraz* (‘now’), *Lolity* (singular genitive case), *być* (‘to be’) are over-represented in PLK. As regards the proper name *Charlotte* transferred by Stiller into the Polish text, Kłobukowski used *Charlotta* as an equivalent partly adapted to the Polish noun declension system. It is the only name of character that differs in the Kłobukowski’s translation. The remaining ones are the same in both texts. Thus, overall, one is made to conclude that Stiller’s translation has more repetitions among top-frequency grammatical words, which can pertain to sentences being more explicit and precise as compared with Kłobukowski’s translation. However, the two translations are similar in terms of high-frequency lexical words.

4.3 Frequency Profiles

In order to determine whether it is the English-original or the Polish translations of *Lolita* that has or have 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. It enables one to answer the question 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 cumulative percentage of the total word count (%cW) corresponding to frequency-based ranks. The results are presented in Table 3.

Although the data in Table 3 show that English *Lolita* (and any English text?) has more repetitions and lower lexical variety among top-frequency words, it is largely due to the lack of lemmatization. Furthermore, the typological difference regarding the character of morphology further confirms the above observation, e.g. articles and prepositions, which are frequently used in English, are treated as separate words, while in Polish various endings, prefixes and suffixes are bound with other stems or roots, which makes the frequencies of Polish words lower. Thus, it is no surprising to observe that the English text (actually, any English text), as compared with Polish, is dominated by top-frequency words (100 top-frequency words constitute almost 50 % of the total number of words used in the text, while in PLS and PLK the corresponding values are 36 % and 32 %, respectively). This observation may be therefore interpreted as the S-universal.

As regards the differences between the Polish translations, one may notice that in Stiller's translation 549 word types account for 50 % of the total word count, while in Kłobukowski's translation this threshold is reached at 758 word types. The data thus show that the translations are not uniform in that respect because PLK is unusually rich and considerably more varied lexically – there are 209 more word types in PLK which account for 50 % of the total word count as compared with PLS.

Table 3 Frequency profiles for top-frequency word types in ENL, PLS and PLK

ENL		PLS		PLK	
Rank	%cW	Rank	%cW	Rank	%cW
1	4.59	1	3.36	1	3.01
10	24.62	10	18.61	10	16.14
100	49.54	100	36.07	100	32.11
105	50.05	549	50.00	758	50.01

4.4 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 particular frequency (f) in the frequency spectra was substituted with information on the cumulative percentage of the vocabulary (%cV) and the cumulative percentage of the total word count (%cW) corresponding to word types with frequencies 1–25. The results are presented in Tables 4, 5 and 6 below.

Interpreting the above data, it is paramount to remember that some of the differences are attributed to different language systems—more analytical (with poor inflection) English versus more synthetic (with rich inflection) Polish, where each inflectional form of a particular word type (e.g. genitive, accusative or locative case of the noun, in either singular or plural, feminine or masculine) is treated as a single occurrence of a type. It is a problem typical of operating with non-lemmatized types and tokens in highly-inflectional languages, such as Polish. With the view of the above, one is in a better position to understand the discrepancy in the data.

As illustrated by the data in Tables 4, 5 and 6, it appears that the Polish translations of *Lolita* are considerably more creative lexically than the English-original. Although such a claim is not based on the analysis of distribution of lemmas, but word types, it is clear that Polish texts contain more low frequency words, where one can usually find creative and author-specific vocabulary (Kenny 2001: 127–134).

Firstly, as regards the number of hapax legomena (i.e. words which occur in a text only once) in ENL, PLS and PLK, the English text has 6,984 hapax legomena, which account for 49.91 % of the total vocabulary (%V) and 6.22 % of the total word count (%W). The PLS and PLK, on the other hand, have 19,560 and 19,586

Table 4 Frequency spectrum for ENL

ENL						
F	V (f)	W (f)	%V	%cV	%W	%cW
1	6,984	6,984	49.91	49.91	6.22	6.22
2	2,470	4,940	18.44	68.36	4.40	10.62
3	1,144	3,432	8.54	76.90	3.05	13.68
4	752	3008	5.61	82.52	2.68	16.36
5	471	2355	3.51	86.03	2.09	18.46
10	121	1210	0.90	93.76	1.07	25.38
25	10	250	0.07	98.83	0.22	34.84

Table 5 Frequency spectrum for PLS

PLS						
F	V (f)	W (f)	%V	%cV	%W	%cW
1	19,560	19,560	68.01	68.01	19.34	19.34
2	4,282	8,564	14.89	82.90	8.46	27.80
3	1,653	4,959	5.74	88.65	4.90	32.71
4	902	3,608	3.13	91.79	3.56	36.28
5	533	2,665	1.85	93.64	2.63	38.91
10	87	870	0.30	96.93	0.86	45.78
25	14	350	0.04	98.75	0.34	53.84

Table 6 Frequency spectrum for PLK

PLK						
F	V (f)	W (f)	%V	%cV	%W	%cW
1	19,586	19,586	67.82	67.82	20.41	20.41
2	4,389	8,778	15.19	83.02	9.14	29.56
3	1,770	5,310	6.12	89.15	5.53	35.10
4	782	3,128	2.70	91.86	3.26	38.36
5	490	2,450	1.69	93.55	2.55	40.91
10	89	890	0.30	96.90	0.92	48.33
25	17	425	0.05	98.78	0.44	57.31

hapaxes, respectively, which account for approximately 68 % of total vocabulary (%V) and 20 % of the total word count (%W). Statistically, it means that every 16th running word is a hapax legomenon in ENL, while in PLS and PLK it is every 5th word—with the false proviso that words are normally distributed in a text. If one takes into consideration overall vocabulary, then in ENL hapax legomena constitute almost 50 % of the text's lexis, while in the Polish translations they account for almost 70 % of all distinct words used.

As regards all word types with frequencies 1–25, the data show that the Polish translations have fewer repetitions and higher lexical variety among bottom-frequency words than ENL (i.e. all these word types account for nearly 35 % of the total word count in ENL and almost 55 % in PLS and PLK). Although this relationship can be treated as another S-universal in English-to-Polish literary translation, it is not known how far that result is influenced by the lack of lemmatization conducted on English and, in particular, on Polish language data. Finally, the data show that Stiller's translation is more varied lexically as regards the number of low-frequency words (i.e. with frequencies 1–25) than Kłobukowski's translation.

5 Conclusions

The aim of the study presented in this article was to compare—with the use of corpus-driven methodology—the English-original and the two Polish translations of *Lolita* by Stiller and Kłobukowski in terms of text length, sentence length, number of repetitions (conciseness of style) as well as frequencies and distribution of both word-types (distinct words) and word-tokens (running words). Also, the aim was to find traces, if any, of translation universals (S-universals, after Chesterman 2004) attested in the Polish translations.

Descriptive statistics revealed that Polish translations of *Lolita* are shorter than the English-original, and it is irrespective of the fact that the length measured by the number of running words indicates otherwise. It remains a contested issue, however, whether this pattern is typical of any English-to-Polish literary translation. Also, it was revealed that the sentences used in the Polish translations are shorter and thus more concise and terse than the ones found in the English-original. Hence, S-universal of explicitation in these particular translations was invalidated. On the other hand, the sentences used in the translations are longer than typical sentences found in Polish prose, which indicates that the translators used faithful sentence-for-sentence translation and long-form syntactic constructions. Comparison of lexical density showed that Kłobukowski's translation is overall lexically richer than Stiller's translation.

Comparison of wordlists showed that Stiller's translation has more repetitions among top-frequency grammatical words, which can point to sentences being more explicit and precise as compared with Kłobukowski's translation. Also, it was revealed that the source text and its two translations are largely similar in terms of high-frequency lexical words, except for the discrepancies due to typological differences between the morphology of the two languages, which were described in greater detail above, and which point to lexical explicitation in English-to-Polish translation.

Finally, comparison of frequency profiles and frequency spectra demonstrated that the English text, as compared with the Polish ones, is dominated by top-frequency words, an observation which may be interpreted as another S-universal, and that Kłobukowski's translation is more lexically varied in terms of the use of top-frequency words than Stiller's one, which has more bottom-frequency words. It was also found that Polish translations have fewer repetitions and higher lexical variety among bottom-frequency words than the English original.

To conclude, it seems that further qualitative research should be conducted to bring to life concrete illustrations of both typical and anomalous cases glossed over in a quantitative text analysis presented above. It is vital since it is still unknown what factors (and to what extent?) impact basic stylometric indicators presented throughout this study. The very impact of source language and target language, direction of translation, genre-specific characteristics, text type, register characteristics, translator's idiolect, author's idiolect, translator's and author's ideologies, source-language culture, target-language culture, onto basic stylometric indicators and, more generally, onto the scope and character of language universals still remain a debatable issue and account for a rather unexplored research area, particularly in the case of English-to-Polish literary translation.

References

Baker, M. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7(2): 223–243.

Baker, M. 2000. Towards a methodology for investigating the style of a literary translator. *Target* 12(2): 241–266.

Baroni, M. 2009. Distributions in text. In *Corpus linguistics: An international handbook Vol. 2*, eds. A. Lüdeling and M. Kytö, 803–821. Berlin and New York: Walter de Gruyter.

Boyd, B. 1995. Chronology of Nabokov's life and works. *The Garland companion to vladimir nabokov*, ed. V. Alexandrov. New York: Garland Publishing Inc.

Chesterman, A. 2004. Beyond the particular. In *Translation universals: Do they exist?*, eds. A. Mauranen and P. Kuyamaki, 33–49. Amsterdam/Philadelphia: John Benjamins.

Comrie, B. 1981. *Language universals and linguistic typology: Syntax and morphology*. Oxford: Basil Blackwell.

Englund Dimitrova, B. 1993. Semantic change in translation – a cognitive perspective. In *Translation and knowledge*, eds. Y. Gambier and J. Tommola, 285–296. Turku: Centre for Translation and Interpreting.

Englund Dimitrova, B. 1994. Statistical analysis of translations (on the basis of translations from and to Bulgarian, Russian and Swedish). *Scandinavian Working Papers on Bilingualism* 9: 87–103.

Grabowski, L. 2012. *A Corpus-driven study of translational and non-translational texts: The case of nabokov's "lolita"*. Opole: Wydawnictwo Uniwersytetu Opolskiego.

Hoover, D. 2004. Testing Burrows's Delta. *Literary and linguistic computing* 19(4): 453–475. Oxford: Oxford University Press.

Kenny, D. 2001. *Lexis and creativity in translation*. London: Routledge.

Laviosa-Braithwaite, S. 1995. Comparable corpora: towards a corpus linguistic methodology for the empirical study of translation. In *Translation and meaning* Part 3, eds. M. Thelen and B. Lewandowska-Tomaszczyk, 153–163. Maastricht: UPM.

Mikhailov, M. 2003. Parallel corpora of literary texts: principles of compilation and use in linguistics and translation studies. *Acta Electronica Universitatis Tamperensis*, 280. Tampere: Tampere University Press. Retrieved on 4 Feb 2011 from: http://acta.uta.fi/pdf/951-44-5754-4.pdf

Mikhailov, M. and M. Villikka. 2001. Is there such a thing as a translator's style?. *Proceedings of the Corpus Linguistics 2001 conference*, 378–386. Lancaster: Lancaster University Press.

Nabokov, V. 1955. *Lolita*. Paris: Olympia Press.

Nabokov, V. 1991. *Lolita: powieść* [Lolita: a novel]. Trans. R. Stiller. Warszawa: Państwowy Instytut Wydawniczy.

Nabokov, V. 1997. *Lolita: powieść* [Lolita: a novel]. Trans. M. Klobukowski. Kraków: DaCapo.

Nida, E. and C. Taber. 1974. *The theory and practice of translation*. Leiden: E.J. Brill.

Oakes, M. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Olohan, M. 2004. Introducing corpora in translation studies. London and New York: Routledge.

Piotrowski, T. 1994. *Problems in bilingual lexicography*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

Ruszkowski, M. 2004. *Statystyka w badaniach stylistyczno-składniowych* [Statistics in research into stylistics and syntax]. Kielce: Wydawnictwo Akademii Świętokrzyskiej.

Rybicki, J. 2007. Stylometria komputerowa w służbie tłumacza. Na przykładzie własnych przekładów. *Rocznik przekładoznawczy* ¾: 169–179. Toruń: Wydawnictwo UMK.

Scarpa, F. 2006. Corpus-based Quality-Assessment of specialist translation: A study using parallel and comparable corpora in English and Italian. In *Insights into specialized translation*, eds. M. Gotti and S. Sarcevic, 155–172. Bern: Peter Lang Verlag.

Scott, M. 2004. *Wordsmith Tools 4.0*. Oxford: Oxford University Press.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: University Press.

Stiller, R. 1991. "Lolita" jako gra i paradoks [Lolita as a Game and Paradox] In *Lolita: powieść* [Lolita: a novel], by V. Nabokov, 419–439. Warszawa: Państwowy Instytut Wydawniczy.

Stiller, R. 1997. Ty żagwio! Albo Nabokov Kłobukowskiego. *Wiadomości Kulturalne* 47 (183): 6–7.