A CORPUS METHODOLOGY FOR ANALYSING TRANSLATION

Rafal S Uzar
University of Lodz

1. Introduction

Translation is becoming an increasingly integral part of society throughout the world. Central Europe, as an example, is 10 years into a new economic system with expanding international business becoming one of the prime catalysts for the development of translation in this particular part of the globe. Translation companies are popping up like mushrooms after a cloudburst. Access to all manner of translation has meant that general standards have improved and as standards improve, more and better qualified translators are needed.

At the English department at the University of Lodz, the English MA degree has a translation option which lasts for the final three years of the five-year MA programme. Nowadays, more students are fulfilling the requirements needed to opt for taking this course. As numbers increase and standards get higher so more is needed from the teacher/trainer, thus, our standards must also rise. This has led to the need to develop different and more efficient tools with which to help us analyse translations, especially the translations of our students.

The English department prides itself on the quality of its translation students, however, there is always room for improvement especially with this rise in standards. With this aim in mind, work has begun on a corpus research project which will give the departmental translator trainers another perspective on the work that they are doing and the translations that the student/trainees are producing.

Within practical applications of language corpora and second language learning, corpora can be loosely divided into three groups:

a) monolingual

b) bilingual (parallel or comparable)

c) learner

For the purposes of translation training each of these corpora have their advantages and disadvantages. Translators are able to utilize all three kinds of corpus in the translation process in an attempt to improve the quality of their work.

The PELCRA project (Polish and English Language Corpora for Research and Applications) was set up at the University of Lodz in 1997. The project was set up to produce extensive corpus resources at both a local and national level. The project consists of a variety of corpora which fall into two main groups:

a) a Polish monolingual corpus

b) an English learner corpus

Translation students/trainees are free to make use of PELCRA using it both as a guide to avoid *learner* errors or erroneous learner tendencies and also as a reference point by using the Polish *national* corpus. The students also have access to the BNC (the British National Corpus) and in this way have two monolingual reference corpora for both of the languages they are working in.

Our students translate from the foreign language into the mother tongue, which is generally considered the norm but are also encouraged to translate from the mother tongue into the foreign language (i.e. from Polish into English). Most problems occur when the translator works into the foreign language and it is here that the learner corpus appears to be most useful.

Extensive work by the PELCRA team (for example, Lenko-Szymanska, 2000 and Lewandowska-Tomaszczyk, McEnery, Lenko-Szymañska, 2000) and other scholars at other institutions (e.g. Kaszubski, 2000) have given students and teachers valuable clues to *dangerous* areas in the production of FL texts so that not only do our students have a wide range of translation tools such as paper dictionaries/thesauri, e-dictionaries, e-glossaries but they also have important published academic work they can consult. Adding to this the reference and learner corpora, they have access to a large bank of knowledge to help them on their way. However, a much needed addition to the resources of PELCRA would be a translation corpus. Work on this corpus has already begun and soon the PELCRA team hope to add this particular element to the resources already currently on offer to the students.

2. The Learner Translation Corpus

In the process of creating these and other additional resources for PELCRA, it became apparent to us that due to the large amount of translation students the English Department has at its disposal a large amount of *student* translation data, texts translated from and into Polish and English, often with one original translated several times. A pilot study was prepared: a small corpus (15000 words) of student translations was compiled and then selected data from this corpus was given back to the same students in the form of collocation print-outs. All names were deleted from the corpus and the collocations were presented so that the students had no idea as to their origin. (An earlier and less advanced preliminary study was conducted and presented at the TELRI conference, Bratislava 1999)

The translator trainees were both surprised and interested by the errors/mistakes and the constructions used in these texts.

Together with the mother-tongue and foreign language monolingual corpora, they began to understand why certain constructions were inappropriate and some simply impossible. They were later informed as to the origin of the texts which led to an awareness for the need to use corpora in their work. It was our intention through this pilot study to gauge the usefulness of such a corpus resource. The students at the Department of English are only just beginning to understand the practical uses of corpora. By bringing home to them these benefits i.e. that corpora can help *them* in the production of FL texts and translations, the students have a greater urge to use them.

Using corpora for translation is now becoming not an uncommon thing for translator trainees but utilising the corpus paradigm for analysing and assessing translations *is* uncommon for the students and teachers here. What became obvious was a need for corpora that were tailored to the student's needs when assessing translation problems i.e. a corpus of their own language so that they could see what they were doing well or what needed more work. This meant the need for a more specialized learner corpus, one created with translation in mind i.e. a corpus containing student translations.

By producing a large corpus of student translations we, the translator trainers, will be given access to the kind of techniques employed by our students. As Coulthard tells us, "…a study of badly written text, or inadequate textualizations, may help us understand better the nature of successful textualization" (1996:2). We can point out to the students the quality of their translation by showing them similar (peer) textualizations and also by having a stockpile of common mistakes to hand compiled using the learner translation corpus. By batching many translations together and having the possibility of concealing the origin of the texts (i.e. making them anonymous), the corpus becomes much more user-friendly as we do not highlight any particular student's work or his/her errors.

Using a corpus approach allows us to annotate and store large amounts of translation data for later use. With corpus tools such as concordancers we can extract important statistical data from our student translations and therefore learn more about our student trainee translators. This brings a level of objectivity to the subjective task of assessing student work. Even a teacher's expertise will be stretched after hand-sifting through his/her 200th translation in a week.

This corpus, therefore, is an attempt to kill the proverbial two birds with one stone; the stone being our learner translation corpus. Firstly, we wish to fill a hole in the resources of PELCRA by providing a corpus of student translations for our students and, secondly, provide the translator trainers with a resource that will aid and perhaps objectify translation quality assessment (TQA).

However, we must be careful not to swing too far in the direction of a purely quantitative analysis and fall into the statistical trap. Translator trainers and trainees all too easily become slaves to statistics when beginning work with corpora. The corpus is a tool. Figures can be easily bent or ignored depending on our initial ideas and interpretations, therefore, our theoretical framework *must* be tight and our qualitative input must also be valuable. Stubbs tells us that, "Quantitative work with large corpora automatically excludes single and idiosyncratic instances, in favour of what is central and typical" (1983:233). Firstly, we can use corpora to help us make generalizations about language but at the same time corpora level out linguistic analysis so that detailed and specific information which we might obtain through hand analysis is lost. The latter is very dangerous in TQA when certain errors may occur only once or twice but give us valuable information about that particular piece of work, thus a level of hand analysis must be used.

3. The Corpus Data

The corpus project consists of a variety of different texts and is very much a specialized package. The first section of the corpus consists of three original Polish texts differing in style and content.

The first is a formal, rather complicated and long article from *Gazeta Wyborcza*, a Polish daily. The second text is an article from *Bravo*, a teenage magazine which uses much teenage slang. The third and final text in this section of the corpus is an *EU* document on accounting.

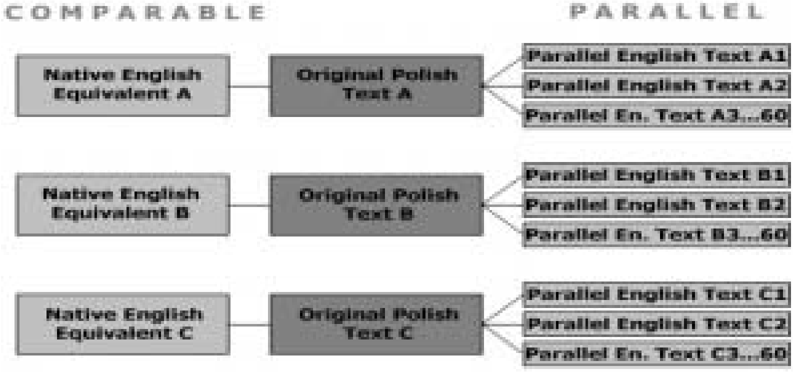
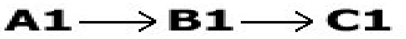


Figure 1: The Learner Translation Corpus Project

Initially, the idea was to focus solely on the creation of a learner translation corpus but our project grew somewhat and we introduced a comparable element to the corpus package. The second set of texts consists of three comparable original English equivalents. These texts were selected on the basis of their thematic and stylistic similarity to the original Polish texts. The first is an article taken from *The Times*, the second is a short piece, taken from *Smash Hits*, and the final text is, like the Polish, an *EU* document on accounting.

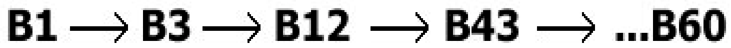
The final part, the parallel section, makes up the bulk of the corpus project and consists of a set of sixty English translations for each of the original Polish texts giving a total of 180 texts. Possessing so many translations allows us to have the best of several worlds. We can analyse the translations of a single student and take into consideration all his/her idiosyncrasies (see Stubbs' quotation above) across three different styles. For example:



A1 → B1 → C1

Figure 2: Analysing the Translations of a Single Author

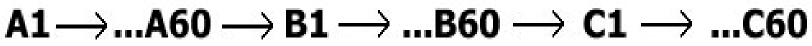
The letters in the diagrams correspond to each of the Polish source texts, the numbers correspond to each of the sixty students. As well as looking at one particular translator in action across three different styles, the corpus linguist can re-focus his/her attention and look at a particular style and how different students cope with the translation of one particular text:



B1 → B3 → B12 → B43 → ...B60

Figure 3: Assessing How Style/Register Affects Translation

The parallel section of the corpus also allows us to analyse all the translations of one particular style vis-à-vis another style illustrated below in figure 14.



A1 → ...A60 → B1 → ...B60 → C1 → ...C60

Figure 4: Comparing Styles

Essentially, the corpus is a sub-language project covering three different sub-languages. By combining both parallel and comparable elements as well as including the PELCRA corpus and the BNC, translation trainees and trainers are able to compare the roles specific and general language, style and register play in translations. In our project we can compare a particular translation with several reference points, namely:

a) the original source text

b) the comparable text

c) translations produced by other students

d) the foreign language reference corpus (the BNC)

This, of course, is all *post-production* work i.e. what the trainer can use to assess the translations or later use as a tool for teaching. However, in the pre-production phase whilst the student is translating, he/she can also use the native speaker corpus (in this case PELCRA) to compare with the source text to see how far the source (statistically and/or stylistically) deviates from the norm. This therefore gives us a fifth reference point:

e) the native-speaker reference corpus (PELCRA)

4. Looking at the Data

Several *corpus* methods can be employed to look at the actual translation *product*. Sixty students translated three different texts from Polish into English. All these texts were aligned. This was undertaken semi-automatically with subsequent hand-correction. Below is an example of a sentence from the Polish source text with five aligned translations. Each sentence is given a start-tag to signal the beginning of the sentence and also tell us the sentence number. This information is placed in angled brackets. The end of the sentence is also marked with an end-tag coming after the full-stop.

< s59 > O atrakcyjnosci tekstu decyduje jego klimat. Ksiazka moze go miec lub nie, bez wzgledu na date jej powstania. Wazne jest, by byl to klimat wlasciwy. < /s59 >

a) < s59 > What determines whether the text is attractive or not is its flavour. The book may have it or not, no matter when written. < /s59 >

b) < s59 > It is the atmosphere that makes a text attractive. Irrespective of the time when the book was written. < /s59 >

c) < s59 > It is the atmosphere of a text that decides about its attractiveness. A book may have that atmosphere or not, regardless of the date of its creation. < /s59 >

d) < s59 > A book may have it or not, regardless of when it was written. It is the right climate that matters. < /s59 >

e) < s59 > It may have a climate or not regardless of when it was written. It is important that the climate is proper. < / s59 >

A simple skim through a few texts followed by a collocation run can give us a list of possible lexical or even phrasal equivalents. In these examples, the Polish *klimat* offers up *flavour*, *atmosphere* and *climate*. What is more, we can see the various kinds of structures put to use in order to render a sentence into the FL. This kind of analysis can lead to:

a) An easier and more effective method of obtaining a more accurate qualitative assessment of the product.

b) A better understanding of what the student is doing when translating i.e. the process of translation.

We can see that through a direct comparison of various translations we find that certain translations read more easily than others and are more accurate renditions of the original Polish.

We are all aware of the value of translation but such a method of analysing work can substantially help us illustrate how:

Translating is one way of helping the learner to control mother-tongue interference by being more aware of its nature. Translating is also very useful in helping learners to become more aware of their own language and understand that it is not more natural or more logical than any other language, but simply a different system…Translation exercises can also show differences in structural patterns and pragmatic strategies between languages as well as the close relationship between language and culture. (Zabalbeascoa, 1997: 122)

Using a corpus methodology in which different styles are placed under scrutiny provides insight into this close link between language and culture. By providing a corpus with various realizations of three different registers we are able to observe how students cope with changes in register which is so often intertwined with culture (social context) as shown in the figure below:

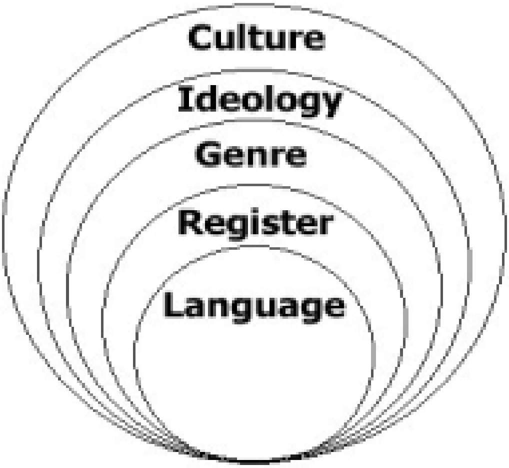
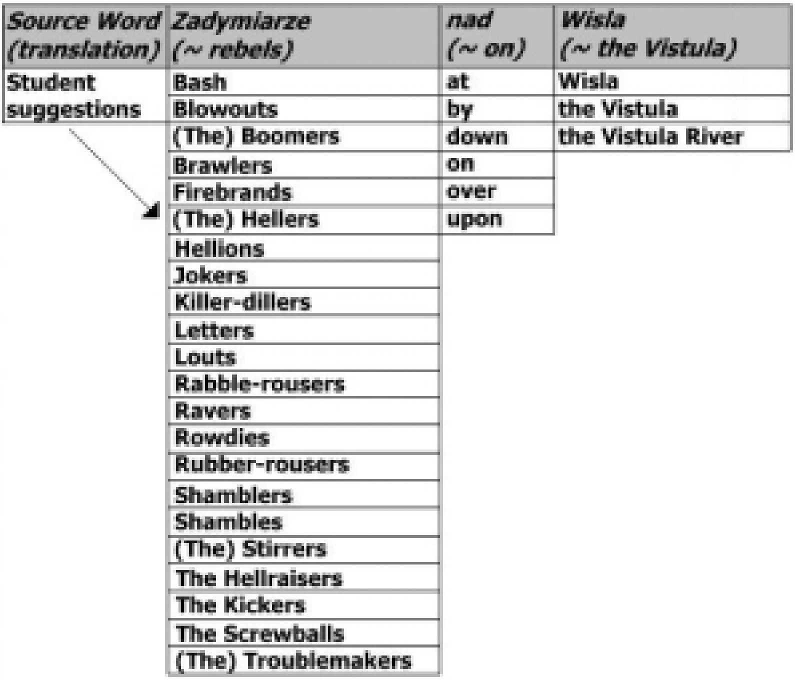


Figure 5: Language and Culture (adapted from Halliday & Martin 1996:38)

The styles of each text are different. For example, the first Polish text taken from *Gazeta Wyborcza* is a formal one and potentially difficult to translate as the writer expresses his opinion very creatively. From this creative yet rather formal text we move onto the second text, a teenage magazine article. Knowledge of a given culture or even sub-culture, namely teenage pop music culture, will undoubtedly be an obvious help in the translation of this text. With the corpus we are able to test these ideas about the language-culture link and see what actually goes on in practice. What is surprising about this teenage text is the fact that it bears some resemblances to the third text on EU accountancy as they are both rather restricted styles with set ways of expressing certain concepts. In these translations whole sections appear to be problematic and errors seem not to be restricted to the lexical level unlike the first source text.

Having a corpus which can highlight not only textualization errors, translation errors but also differences between translations and styles is no doubt helpful for TQA and students of translation. What we found, when putting our corpus data to use was that there were marked statistical differences between below average translations and better quality translations. This difference often lay in the way students dealt with culturally specific terms in one text and how they dealt with non-marked words in other texts.

When presenting our findings to students and other translator trainers, we experimented by giving them sentences, phrases and also individual words to compare. By using individual words, we were able to highlight particular problem areas stemming from a lack of contextualized knowledge. The figure below shows such an example taken from the second, rather informal text.



Source Word (translation)	*Zadymiarze (~ rebels)*	*nad (~ on)*	*Wisla (~ the Vistula)*
Student suggestions	Bash	at	Wisla
	Blowouts	by	the Vistula
	(The) Boomers	down	the Vistula River
	Brawlers	on	
	Firebrands	over	
	(The) Hellers	upon	
	Hellions		
	Jokers		
	Killer-dillers		
	Letters		
	Louts		
	Rabble-rousers		
	Ravers		
	Rowdies		
	Rubber-rousers		
	Shamblers		
	Shambles		
	(The) Stirrers		
	The Hellraisers		
	The Kickers		
	The Screwballs		
	(The) Troublemakers		

Figure 6: Comparisons at the Lexical Level

The greater the lack of contextualized knowledge across a larger group of students, the more varied the suggestions will be as we can see above. It seems strange that in a group of sixty advanced-level students none thought to use the word *rebels*, *yobs* or even *hooligans* as an equivalent for *Zadymiarze*. This is not to say these are more appropriate suggestions for the Polish *Zadymiarze*; it only highlights the lack of depth of knowledge in a particular cross-section of learners.

The simple sentence shown in figure 6 gives us many clues to the problems involved in translation as well as the possibilities and choices involved when producing a translation. If we look at the third word, we are reminded that many place names and geographical names have their equivalents in other languages e.g. *Milano* or *Roma* are Milan and Rome in English. In this way, the third word in the example above is restricted in the number of its permissible equivalents i.e. there is only one accepted translation. Frequency data has been omitted here, however, the majority of our potential translators chose the correct equivalent, some adding the word, *river* although a few of our translators simply took the Polish word across without translating it, hence only three equivalents in our cross-section of learners.

The second problem in the sentence above rests with the preposition. As we can see six possibilities have been given for the second word, *nad*. The word is largely unambiguous and this is reflected by the fact that most of our students used either *on* or *upon*. The very fact that this is a closed word category brings the number of equivalents down to a minimum, giving us only six different suggestions.

However, the most interesting word from the point of view of the translator trainer is the first word in the sentence shown in figure 6. It can be roughly translated as *rebels*. When we look through our list of student translations, we see by what factor the translation possibilities are raised. The first word, *Zadymiarze* gave as many as twenty-five *different* suggestions (some have not been included here).

This illustrates the choices faced in the process of translation and highlights to student translators what a linguistically challenging activity translation is. The sixty students produced approximately forty-five completely different translations of this one sentence alone. Had students been given access to a learner translation corpus would these errors have occurred? Probably, although, we believe that *fewer* errors would have been committed. Corpora often perform the duty of being a catalyst for introspection. Even access to a corpus of the native language (the language from which they were translating) would have helped. For example, the students would have greatly benefited from using the PELCRA corpus in their search for an equivalent and would have found that often, when using a corpus of one's own language, one is often made aware of unfamiliar collocations; *Zadymiarze*, for example, often collocated with members of Solidarity portrayed as *rebels* or disrupters of the peace.

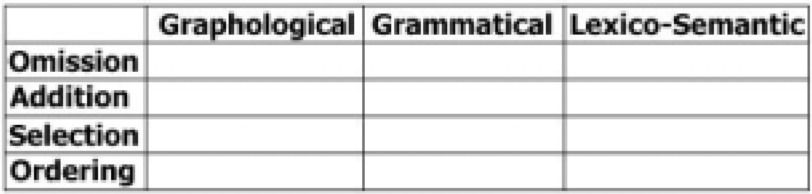
5. Annotating the Data

The student of translation has, therefore, a few extra spanners in the toolbox. We have found that putting together all of these various approaches to using corpora has been of enormous value to our budding corpus linguists and translator trainees. For teachers and translator trainers this has also helped, however, the corpus required the addition of information that might tell us even more about what was going on in the translation. These *additions* came in the form of various kinds of error tags based on a number of sources.

These *tags* are preliminary and will hopefully lead to an expanded error taxonomy for translation. Firstly, the entire corpus was annotated with every translated sentence in the corpus given a +, - , or 0 tag indicating whether a particular translated sentence was an appropriate, inappropriate or relatively appropriate sentence in relation to its original. This information was then placed in angle brackets at the end of each sentence. These are undoubtedly subjective ideas as to the quality of the translation but with this kind of system we make our assessment transparent and open to discussion from other translators and translator trainers.

The next step was to go into more depth with our ideas about the quality of the translations. A selection (fifty) of the 180 texts was analysed and assessed in terms of textual quality, readability and translation quality by a wide range of people from non-experts (English native speakers with no knowledge of translation, linguistics or Polish), to people with varying degrees of knowledge of translation and Polish, to professional translators highly competent in both languages. The corpus was then enriched with comments made by these individuals about the translations and about the errors in the translations. These took the form of extended tags.

Nevertheless, the problem of how to define an error remains. Errors in production/translation need to be categorized in some *systematic* way using an error taxonomy created for this purpose. In our search for an appropriate taxonomy, we consulted S. Pit Corder's work:



	Graphological	Grammatical	Lexico-Semantic
Omission			
Addition			
Selection			
Ordering			

Figure 7: The Corder Matrix

The matrix provided by Corder (1981:36) was our first step towards formulating an error taxonomy. Corder himself admits that the matrix is not perfect, for example, article omission or addition would be classified as different errors although they belong to a similar conceptual category. Nevertheless, the matrix is not overly complicated and it will not take a great deal of time to error tag the corpus by hand using these ideas. Eventually, an error taxonomy will be created that will take into consideration criteria such as:

a) a first impression of *each* sentence i.e. plus, minus or zero values
b) comments by our native-speakers/professionals
c) the Corder matrix

Point (c) will be implemented in the further stages of annotation. The corpus has, however, already been annotated using points (a) and (b). The examples below show three types of error tagging. The first example puts to use point (a) and is the result of a scan analysis of the text by an English (native speaker) who is also competent in Polish and is a translator.

Example 1 using point (a)

< AnalysisScan - English expert with Polish >
In short, beneath the Gombrowicz's deformations a student has to suspect that what is simple, obvious and clear. < - > In such case a young reader asks, why then sophisticated? < - > Why the author couldn't write directly what's his case. < - > If there's no tomfoolery, there's no fun. < + >

The second example uses point (b) and gives us extended comments made by an English native speaker linguist who has no knowledge of Polish. These comments have been put into a tag format. The sentence beginning *He notices something…* and ending *…the imagination game* is regarded as *uncertain* by the expert. Other elements within the sentence are seen as *problematic* therefore this information has been included within the sentence.

Example 2 using point (b)

< Analysis - English expert with no Polish >
< Uncertain > < problematic > He notices something with which < / problematic > young readers would probably agree; he says that literature < problematic > should not inform about reality but it should make from reality the imagination game < / problematic > . < /Uncertain >

The final analysis also uses point (b). This expert is again someone who has knowledge of both languages and is a translator. The comments differ slightly in their format from the previous tags due to the fact that the expert highlighted what he found to be an error (marked as *err*) and also added his suggestions/correction (marked as *cor*).

Example 3 using point (b)

< Analysis - English expert with Polish >
What determines whether the text is attractive or not is its

< err > flavour < /err > [cor]atmosphere[/cor]. The book may have it or not, < err > no matter < /err > [cor]irrespective of[/cor] when it was written.

Once our error taxonomy is completed, the error tags will be used as tools with which to compare the translation with the original and specifically highlight what is problematic in the translation and then attempt to discover the root of these problems. In this way the translation *product* tells us about the *process*.

6. Analysis and Assessment

As linguists and teachers we need to observe and understand both the product and the process. One of the goals is to improve the way in which our students translate. This can only be undertaken through the analysis of what is tangible i.e. the *product*. Therefore the assessment of the product and the feedback we obtain from the students will feed translation training which in turn will aid students in the translation process at a later stage. A similar process can be seen when we observe the *method* of analysing either the product or the process. Our goal is to improve the *quality* of translation. This can be tangibly analysed quantitatively using the corpus. Work with the corpus then feeds our introspection and gives rise to improved qualitative assessment and analysis. The translator trainer is therefore using corpus analysis as a direct feed for translator training. The diagram shows two concurrent phenomena:

a) What is assessed
b) How it is analysed

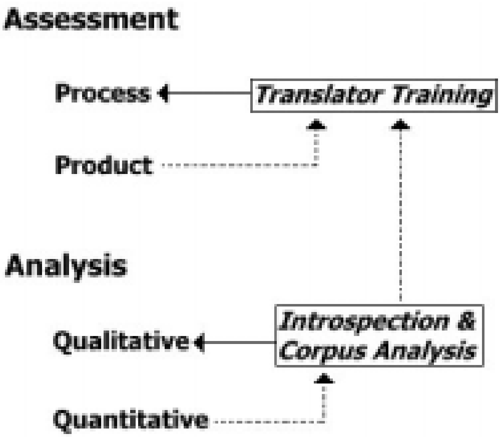
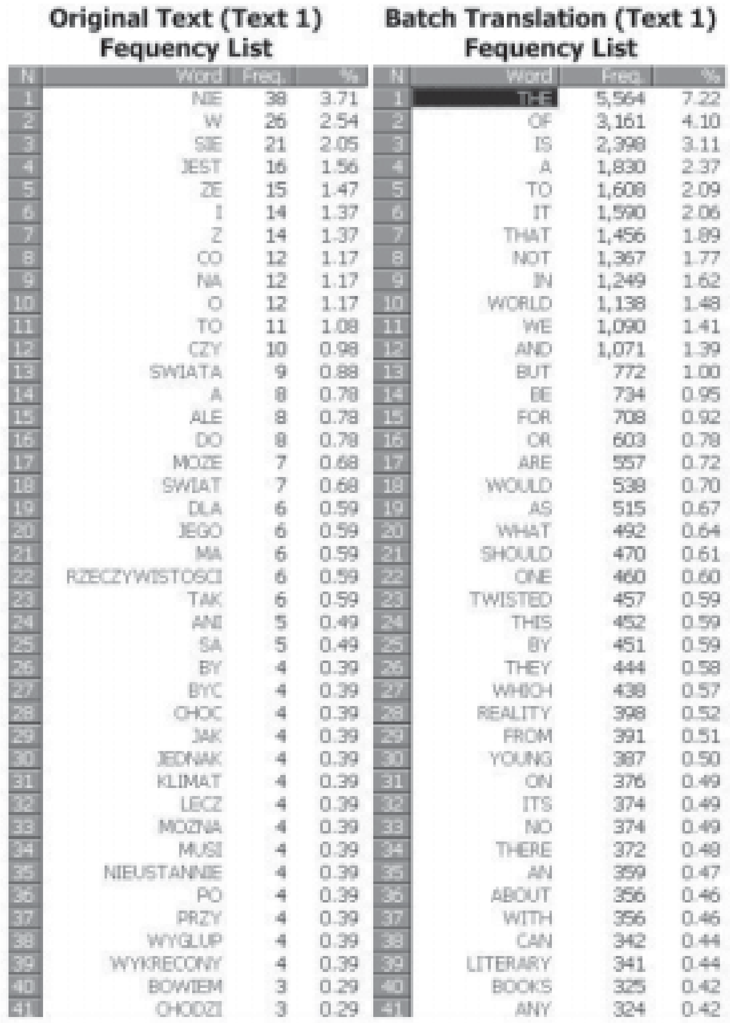


Figure 8: A Corpus Methodology for Analysing Translation

7. Comparing Parallel Texts Statistically

Thus, we are creating and then using corpus tools which will enhance our assessment of the translation product and process. We must bear in mind the close relationship between quantitative work and our introspections and the value that each have in TQA. By taking our quantitative analyses to a higher level, we are able to improve our qualitative assessment. We can, for instance, look at the frequency lists of parallel texts to give us some idea of the kind of lexis used by our translators and how often certain vocabulary items are used.

Looking at individual texts, for example, allows us to find the small and interesting idiosyncrasies of one translator. Time-consuming work but necessary for the trainer when assessing a piece of work. However, we can also *batch* translations together. In this way we can see the behaviour and statistical patterning of a large group of translators producing the same text. Figure 9 below shows wordlists for both the source text and the batch translation wordlist (of all sixty translations). If we look at figure 9 we see that the lexical equivalents *swiat – world* and *rzeczywistosc – reality* are statistically similar across sixty students but if we take the word *wykrecony* with a frequency of 0.39% we see that its most common *student* equivalent, *twisted*, has a frequency of 0.59%. Taking into account how statistically similar other equivalents are, for example *world* or *reality* this may highlight a point of interest in the translation process. May the students have over-used this English word? As it happens, all the students used this equivalent. None of the translations included synonyms like *bizarre*, *weird* etc. and the use of *twisted* seems a little forced and is generally overused throughout the text. These kinds of *corpus clues* allow us to make quicker assumptions when assessing a translation.



Original Text (Text 1)
Fequency List

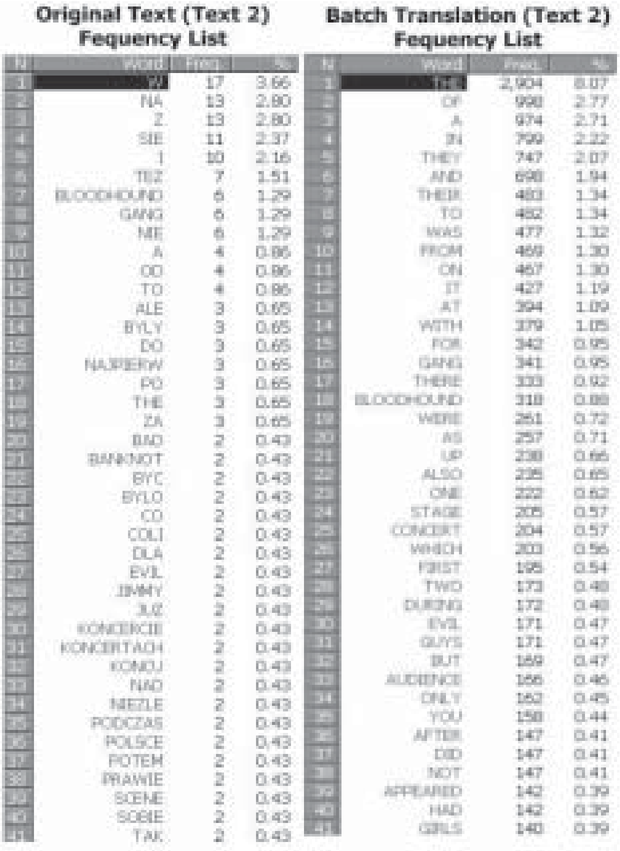
N	Word	Freq.	%
1	NIE	38	3.71
2	W	26	2.54
3	SIE	21	2.05
4	JEST	16	1.56
5	ZE	15	1.47
6	I	14	1.37
7	Z	14	1.37
8	CO	12	1.17
9	NA	12	1.17
10	O	12	1.17
11	TO	11	1.08
12	CZY	10	0.98
13	SWIATA	9	0.88
14	A	8	0.78
15	ALE	8	0.78
16	DO	8	0.78
17	MOZE	7	0.68
18	SWIAT	7	0.68
19	DLA	6	0.59
20	JEGO	6	0.59
21	MA	6	0.59
22	RZECZYWISTOSCI	6	0.59
23	TAK	6	0.59
24	ANI	5	0.49
25	SA	5	0.49
26	BY	4	0.39
27	BYC	4	0.39
28	CHOC	4	0.39
29	JAK	4	0.39
30	JEDNAK	4	0.39
31	KLIMAT	4	0.39
32	LECZ	4	0.39
33	MOZNA	4	0.39
34	MUSI	4	0.39
35	NIEUSTANNIE	4	0.39
36	PO	4	0.39
37	PRZY	4	0.39
38	WYGLUP	4	0.39
39	WYKRECONY	4	0.39
40	BOWIEM	3	0.29
41	CHODZI	3	0.29

Batch Translation (Text 1)
Fequency List

N	Word	Freq.	%
1	THE	5,564	7.22
2	OF	3,161	4.10
3	IS	2,398	3.11
4	A	1,830	2.37
5	TO	1,608	2.09
6	IT	1,590	2.06
7	THAT	1,456	1.89
8	NOT	1,367	1.77
9	IN	1,249	1.62
10	WORLD	1,138	1.48
11	WE	1,090	1.41
12	AND	1,071	1.39
13	BUT	772	1.00
14	BE	734	0.95
15	FOR	708	0.92
16	OR	603	0.78
17	ARE	557	0.72
18	WOULD	538	0.70
19	AS	515	0.67
20	WHAT	492	0.64
21	SHOULD	470	0.61
22	ONE	460	0.60
23	TWISTED	457	0.59
24	THIS	452	0.59
25	BY	451	0.59
26	THEY	444	0.58
27	WHICH	438	0.57
28	REALITY	398	0.52
29	FROM	391	0.51
30	YOUNG	387	0.50
31	ON	376	0.49
32	ITS	374	0.49
33	NO	374	0.49
34	THERE	372	0.48
35	AN	359	0.47
36	ABOUT	356	0.46
37	WITH	356	0.46
38	CAN	342	0.44
39	LITERARY	341	0.44
40	BOOKS	325	0.42
41	ANY	324	0.42

Figure 9: Comparing Frequency Lists (Text 1)

We can use this technique with all our texts, comparing the source text with the sixty translations. The data from the second text (figure 10) with its sixty translations does not, at first glance, throw up any problems, although the text (a teenage magazine article) was problematic for students. This may mean that the problem areas are not lexical. A case, perhaps, of not seeing the forest for the trees. We see no problems at the lexical level but problems may, in fact, be occurring at higher levels, at the phrasal or sentence level. As we can see, the corpus approach acts as a catalyst for introspection, throwing up questions about the how and why of the translation product and process.



Original Text (Text 2)
Fequency List

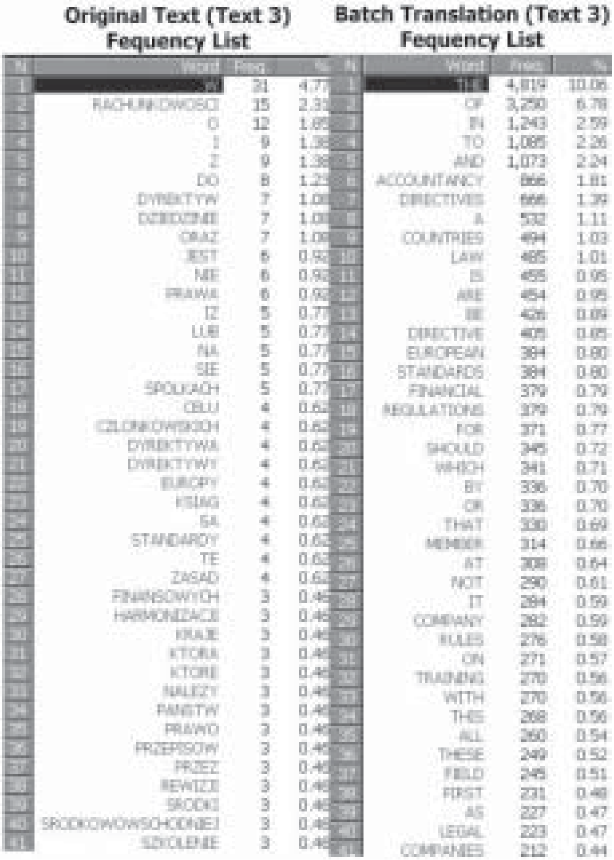
N	Word	Freq.	%
1	W	17	3.66
2	NA	13	2.80
3	Z	13	2.80
4	SIE	11	2.37
5	I	10	2.16
6	TEZ	7	1.51
7	BLOODHOUND	6	1.29
8	GANG	6	1.29
9	NIE	6	1.29
10	A	4	0.86
11	OD	4	0.86
12	TO	4	0.86
13	ALE	3	0.65
14	BYLY	3	0.65
15	DO	3	0.65
16	NAJPIERW	3	0.65
17	PO	3	0.65
18	THE	3	0.65
19	ZA	3	0.65
20	BAD	2	0.43
21	BANKNOT	2	0.43
22	BYC	2	0.43
23	BYLO	2	0.43
24	CO	2	0.43
25	COLI	2	0.43
26	DLA	2	0.43
27	EVIL	2	0.43
28	JIMMY	2	0.43
29	JUZ	2	0.43
30	KONCERCIE	2	0.43
31	KONCERTACH	2	0.43
32	KONCU	2	0.43
33	NAD	2	0.43
34	NIEZLE	2	0.43
35	PODCZAS	2	0.43
36	POLSCE	2	0.43
37	POTEM	2	0.43
38	PRAWIE	2	0.43
39	SCENE	2	0.43
40	SOBIE	2	0.43
41	TAK	2	0.43

Batch Translation (Text 2)
Fequency List

N	Word	Freq.	%
1	THE	2,904	8.07
2	OF	998	2.77
3	A	974	2.71
4	IN	799	2.22
5	THEY	747	2.07
6	AND	698	1.94
7	THEIR	483	1.34
8	TO	482	1.34
9	WAS	477	1.32
10	FROM	469	1.30
11	ON	467	1.30
12	IT	427	1.19
13	AT	394	1.09
14	WITH	379	1.05
15	FOR	342	0.95
16	GANG	341	0.95
17	THERE	333	0.92
18	BLOODHOUND	318	0.88
19	WERE	261	0.72
20	AS	257	0.71
21	UP	238	0.66
22	ALSO	235	0.65
23	ONE	222	0.62
24	STAGE	205	0.57
25	CONCERT	204	0.57
26	WHICH	203	0.56
27	FIRST	195	0.54
28	TWO	173	0.48
29	DURING	172	0.48
30	EVIL	171	0.47
31	GUYS	171	0.47
32	BUT	169	0.47
33	AUDIENCE	166	0.46
34	ONLY	162	0.45
35	YOU	158	0.44
36	AFTER	147	0.41
37	DID	147	0.41
38	NOT	147	0.41
39	APPEARED	142	0.39
40	HAD	142	0.39
41	GIRLS	140	0.39

Figure 10: Comparing Frequency Lists (Text 2)

Looking at the third source text's frequency list alongside the frequency list of all its translations, we are able to see some interesting points. We must remember that this kind of statistical information gives us a quick look at what is going on in the texts and can lead to deeper analysis later. If we look at figure 11, we see that the English word, *accountancy* is used less frequently in the translations than its equivalent, *rachunkowosc*, is used in the source text. Why? The answer is clear when we look further down the list. Some students decided to use the word *accounting* instead. Again we have another example of the corpus giving us clues about the translation patterns of a particular group of translators. By scanning the frequency lists, we might also notice the translation techniques employed by students. For example, the source language gives us two synonymous words, *panstwo* and *kraj*. The students have opted to translate both these words using one English equivalent, *country*. Only with corpus data can a translator trainer begin to make suppositions about student work with such speed.

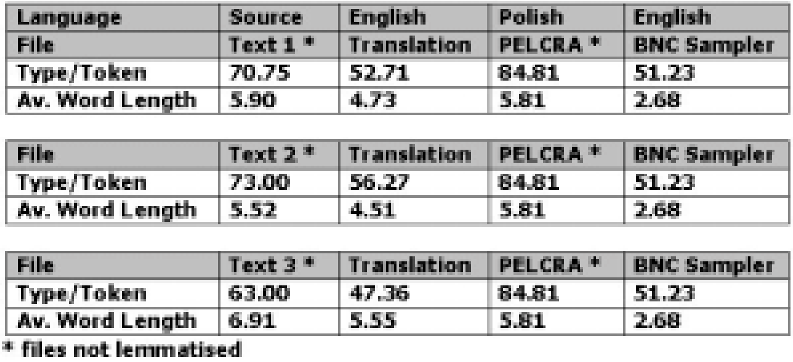


Original Text (Text 3) Fequency List				Batch Translation (Text 3) Fequency List			
N	Word	Freq.	%	N	Word	Freq.	%
1	W	31	4.77	1	THE	4,819	10.06
2	RACHUNKOWOSCI	15	2.31	2	OF	3,250	6.78
3	O	12	1.85	3	IN	1,243	2.59
4	I	9	1.38	4	TO	1,085	2.26
5	Z	9	1.38	5	AND	1,073	2.24
6	DO	8	1.23	6	ACCOUNTANCY	866	1.81
7	DYREKTYW	7	1.08	7	DIRECTIVES	666	1.39
8	DZIEDZINIE	7	1.08	8	A	532	1.11
9	ORAZ	7	1.08	9	COUNTRIES	494	1.03
10	JEST	6	0.92	10	LAW	485	1.01
11	NIE	6	0.92	11	IS	455	0.95
12	PRAWA	6	0.92	12	ARE	454	0.95
13	IZ	5	0.77	13	BE	426	0.89
14	LUB	5	0.77	14	DIRECTIVE	405	0.85
15	NA	5	0.77	15	EUROPEAN	384	0.80
16	SIE	5	0.77	16	STANDARDS	384	0.80
17	SPOLKACH	5	0.77	17	FINANCIAL	379	0.79
18	CELU	4	0.62	18	REGULATIONS	379	0.79
19	CZLONKOWSKICH	4	0.62	19	FOR	371	0.77
20	DYREKTYWA	4	0.62	20	SHOULD	345	0.72
21	DYREKTYWY	4	0.62	21	WHICH	341	0.71
22	EUROPY	4	0.62	22	BY	336	0.70
23	KSIAG	4	0.62	23	OR	336	0.70
24	SA	4	0.62	24	THAT	330	0.69
25	STANDARDY	4	0.62	25	MEMBER	314	0.66
26	TE	4	0.62	26	AT	308	0.64
27	ZASAD	4	0.62	27	NOT	290	0.61
28	FINANSOWYCH	3	0.46	28	IT	284	0.59
29	HARMONIZACJI	3	0.46	29	COMPANY	282	0.59
30	KRAJE	3	0.46	30	RULES	276	0.58
31	KTORA	3	0.46	31	ON	271	0.57
32	KTORE	3	0.46	32	TRAINING	270	0.56
33	NALEZY	3	0.46	33	WITH	270	0.56
34	PANSTW	3	0.46	34	THIS	268	0.56
35	PRAWO	3	0.46	35	ALL	260	0.54
36	PRZEPISOW	3	0.46	36	THESE	249	0.52
37	PRZEZ	3	0.46	37	FIELD	245	0.51
38	REWIZJI	3	0.46	38	FIRST	231	0.48
39	SRODKI	3	0.46	39	AS	227	0.47
40	SRODKOWOWSCHODNIEJ	3	0.46	40	LEGAL	223	0.47
41	SZKOLENIE	3	0.46	41	COMPANIES	212	0.44

Figure11: Compring Frequency Lists (Text 3)

8. Comparing Translations with Native Language Corpora

Other ways in which we can analyse translations to help us with assessing work is by comparing the statistics of translations vis-à-vis two reference corpora and by undertaking an inter-genre comparison. If we look at figures 12 and 13 we see that the average translation of the first source text has a very similar standardized type/token ratio to the BNC Sampler, the average word length however is somewhat higher. The second text has a higher Type/Token ratio than the first text and we find that the type/token ratio of the translations of the second text also increases. The average word length of the second text is shorter as is the average word length in the translation, which is interesting taking into consideration the fact that these are two different languages. The third source text sees a large drop in the Type/Token ratio and the average translation also sees a drop below the Type/Token ratio of the BNC Sampler. The average word length shows a sharp increase in the third text and we see the same correlation in the average translation of this text. Are these figures significant? We cannot always be completely sure with statistics but this does give us is an idea that native language interference could well be a problem and it is up to the teacher to then move forward and investigate these kind of results. We can compare statistics across genres to see if a particular genre causes the translation to gravitate towards the source language rather than the target culture by using reference corpora.



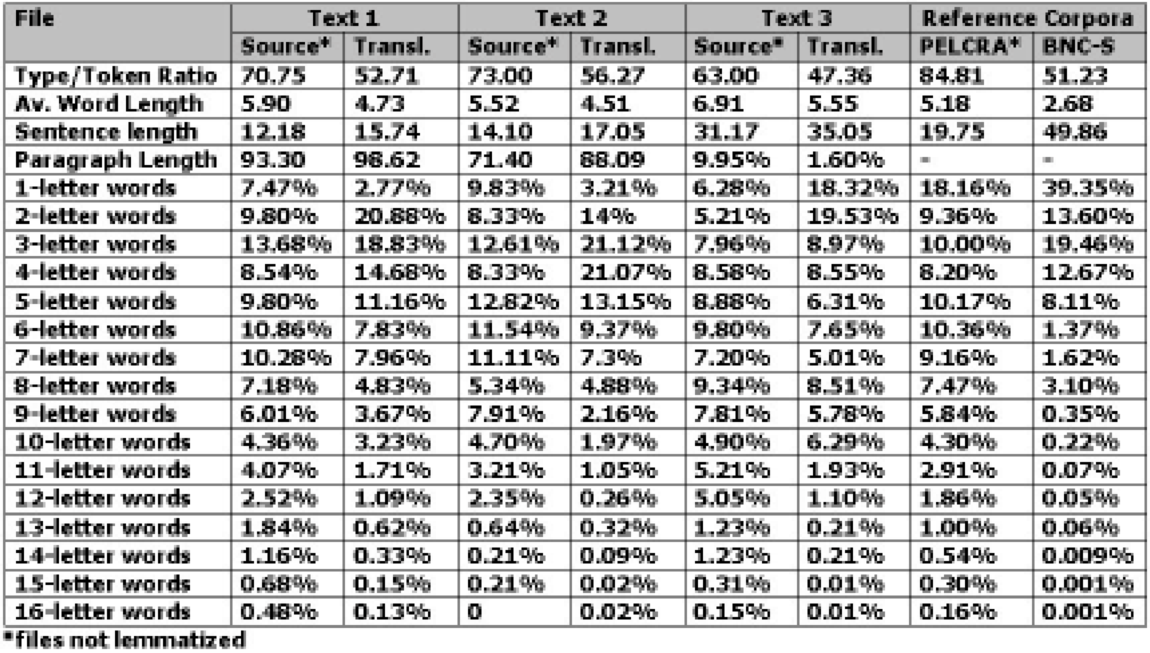
Language	Source	English	Polish	English
File	Text 1 *	Translation	PELCRA *	BNC Sampler
Type/Token	70.75	52.71	84.81	51.23
Av. Word Length	5.90	4.73	5.81	2.68

File	Text 2 *	Translation	PELCRA *	BNC Sampler
Type/Token	73.00	56.27	84.81	51.23
Av. Word Length	5.52	4.51	5.81	2.68

File	Text 3 *	Translation	PELCRA *	BNC Sampler
Type/Token	63.00	47.36	84.81	51.23
Av. Word Length	6.91	5.55	5.81	2.68

* files not lemmatised

Figure 12: Comparing Translations with Reference Corpora



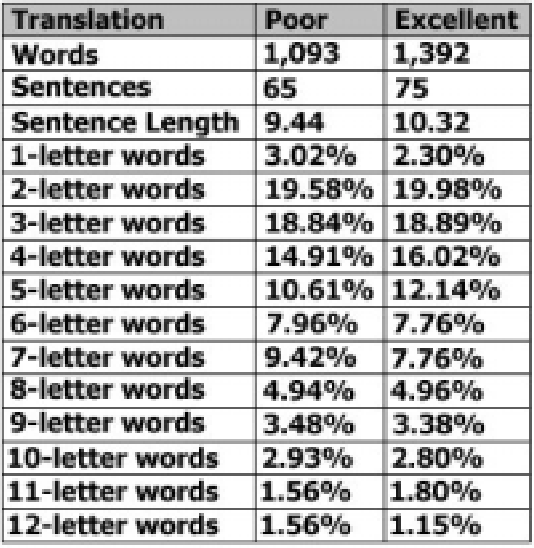
File	Text 1		Text 2		Text 3		Reference Corpora	
	Source*	Transl.	Source*	Transl.	Source*	Transl.	PELCRA*	BNC-S
Type/Token Ratio	70.75	52.71	73.00	56.27	63.00	47.36	84.81	51.23
Av. Word Length	5.90	4.73	5.52	4.51	6.91	5.55	5.18	2.68
Sentence length	12.18	15.74	14.10	17.05	31.17	35.05	19.75	49.86
Paragraph Length	93.30	98.62	71.40	88.09	9.95%	1.60%	-	-
1-letter words	7.47%	2.77%	9.83%	3.21%	6.28%	18.32%	18.16%	39.35%
2-letter words	9.80%	20.88%	8.33%	14%	5.21%	19.53%	9.36%	13.60%
3-letter words	13.68%	18.83%	12.61%	21.12%	7.96%	8.97%	10.00%	19.46%
4-letter words	8.54%	14.68%	8.33%	21.07%	8.58%	8.55%	8.20%	12.67%
5-letter words	9.80%	11.16%	12.82%	13.15%	8.88%	6.31%	10.17%	8.11%
6-letter words	10.86%	7.83%	11.54%	9.37%	9.80%	7.65%	10.36%	1.37%
7-letter words	10.28%	7.96%	11.11%	7.3%	7.20%	5.01%	9.16%	1.62%
8-letter words	7.18%	4.83%	5.34%	4.88%	9.34%	8.51%	7.47%	3.10%
9-letter words	6.01%	3.67%	7.91%	2.16%	7.81%	5.78%	5.84%	0.35%
10-letter words	4.36%	3.23%	4.70%	1.97%	4.90%	6.29%	4.30%	0.22%
11-letter words	4.07%	1.71%	3.21%	1.05%	5.21%	1.93%	2.91%	0.07%
12-letter words	2.52%	1.09%	2.35%	0.26%	5.05%	1.10%	1.86%	0.05%
13-letter words	1.84%	0.62%	0.64%	0.32%	1.23%	0.21%	1.00%	0.06%
14-letter words	1.16%	0.33%	0.21%	0.09%	1.23%	0.21%	0.54%	0.009%
15-letter words	0.68%	0.15%	0.21%	0.02%	0.31%	0.01%	0.30%	0.001%
16-letter words	0.48%	0.13%	0	0.02%	0.15%	0.01%	0.16%	0.001%

*files not lemmatized

Figure 13: Comparing Translations with Reference Corpora 2

9. Comparing Individual Texts

As well as comparing the batch translations with reference corpora, we can also compare the statistics of individual translations with each other. For example, we can verify our introspections about a particular translation by comparing it with another. We might, for instance, wish to assess our ideas as to what the difference between a *good* and *inadequate* translation is.



Translation	Poor	Excellent
Words	1,093	1,392
Sentences	65	75
Sentence Length	9.44	10.32
1-letter words	3.02%	2.30%
2-letter words	19.58%	19.98%
3-letter words	18.84%	18.89%
4-letter words	14.91%	16.02%
5-letter words	10.61%	12.14%
6-letter words	7.96%	7.76%
7-letter words	9.42%	7.76%
8-letter words	4.94%	4.96%
9-letter words	3.48%	3.38%
10-letter words	2.93%	2.80%
11-letter words	1.56%	1.80%
12-letter words	1.56%	1.15%

Figure 14: Comparing Individual Translations

We can select one text which we feel is above average and one we feel is below average as in figure 14. By producing a wordlist we can compare the number of words used in each text, the average sentence length or the number of 1-letter, 2-letter words etc. We can see that the better translation uses more words, contains more sentences and has a higher sentence length. The poorer text uses a marked percentage more simple 1-letter words (and strangely enough 7-letter words). The better translation uses a *marked* percentage more 4- and 5-letter words. Again, does this tell us anything significant? This method allows us to assess or re-assess *our* assessments of a translation. In a sense, it gives us the evidence to back up our ideas about the quality of a particular translation.

10. Conclusions and Future Work

A range of simple corpus methods have been put to use, methods which not only aid the translator trainee in producing a translation but also help the trainer assess the work of his/her students and the techniques used by these students. We can:

a) Compare the source text against a batch of translations using collocations,
    i. Compare how one word is translated across the whole group,
    ii. Compare how phrases are translated across the whole group,
a) Compare source and batch translation wordlists,
b) Compare source and batch translation wordlists with reference corpora wordlists,
c) Compare wordlists of individual translations,
d) Compare statistics across genres

This work is supported by the addition of error tags to the corpus allowing us to pinpoint problematic *hotspots* in the translations. The corpus will eventually be fully annotated with error tags and *positive* tags which will also highlight interesting translations of certain source phrases or sentences. The translations have already been aligned with the source texts to make life easier for the translator trainer. We also plan to work on the alignment of the source texts with the comparable texts using work already undertaken here in Lodz. With these steps completed we hope to move on and produce a language-independent concordancing tool for the translator which would allow us to view the concordance lists of source texts, comparable texts and translations together on one screen.

When working with texts and the translation of these texts, TQA is often based on the *feelings* of the professional translator trainer, whether a translation is a very good interpretation of the source text or not. It is often difficult to ascertain what *exactly* makes a particular translation *good*. This approach allows the translator trainer (or even the translation student) to quickly verify his/her intuitions through corpus evidence. With additional translation corpora and learner corpora being added to PELCRA as well as further work which has already begun on error tagging, we hope to create an invaluable resource and methodology which will help translators not only working in Polish but in other languages. We can then use these *learner* statistics to predict what might happen when students work with other texts and in other languages. It is our hope that work in learner translation corpora will be produced in other languages so that translators might have a better idea of the universal problems faced in translation much the same way that learner corpora across the world are giving us ideas about the universal problems of learners of foreign languages.

Bibliography

Corder, S.P. (1981) *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

Coulthard, M. (1996) *On Analysing and Evaluating Written Text*. in Coulthard, M. *Advances in Written Text Analysis*. London/New York: Routledge, pp.1-11.

Halliday, M.A.K. & J. R. Martin (1996) *Writing Science: Literary and Discursive Power*. London/Washington D.C: The Falmer Press.

Kaszubski, P. (2000) "Lexical profiling of English (learner) corpora: can we measure advancement levels?" in Lewandowska-Tomaszczyk, B. & P. J. Melia (eds) *PALC'99: Practical Applications in Language Corpora*. Frankfurt: Peter Lang.

Leñko-Szymañska, A. (2000) "Passive and active vocabulary knowledge in advanced learners of English" in Lewandowska-Tomaszczyk, B. & P. J. Melia (eds) *PALC'99: Practical Applications in Language Corpora*. Frankfurt: Peter Lang.

Lewandowska-Tomaszczyk, B., A. Leñko-Szymañska & A. M. McEnery (2000) "Lexical problem areas in the PELCRA learner corpus of English" in Lewandowska-Tomaszczyk, B. & P. J. Melia (eds) *PALC'99: Practical Applications in Language Corpora*. Frankfurt: Peter Lang.

Scott, M. (1998) *WordSmith: Software Language Tools for Windows*. Oxford: Oxford University Press.

Stubbs, M. (1996) *Text and Corpus Analysis*. London: Blackwell.

Zabalbeascoa, P. (1997) "Language Awareness and Translation" in van Lier, L. & D. Corson (eds.) *Encyclopedia of Language and Education: Knowledge about Language, Volume 6*. Dordrecht/Boston/London: Kluwer Academic Publishers, pp 119-130.