Phases of translation corpus

Compilation and analysis

Andrius Utka
Vytautas Magnus University, Kaunas

The absolute majority of scholarly work in descriptive translation studies is product-oriented. In this article, the focus is moved from product-oriented to process-oriented translation studies by compiling an English – Lithuanian Phases of Translation Corpus (PT corpus). The PT corpus is analysed using quantitative and qualitative analyses. The quantitative analysis using frequency information highlights the difficult word types that either are missing or are inconsistently translated in successive Lithuanian translated versions. The qualitative analysis continues the quantitative research by help of parallel concordancing. The problematic cases of translation are extracted and cases of normalization, systematic replacement of terminology, and influence by the original language are reported.

Keywords: Lithuanian, phases of translation corpus, process-oriented translation studies, alignment, parallel concordancing

1. Introduction

European Community law documents have been translated into most European languages for several decades now. The translation of European Community Law is especially important to new Member States of European Union, as the translated European Union law is serving as a legal basis for creating new legislation systems in these states. The language of EU law documents inevitably leaves its traces in these translations, as a result of this the language of the whole legislation can be affected.

The article is aimed to present the findings of this research along descriptive lines of investigation, and avoid any prescriptive evaluation of the language. This is not so easy, as the quality of English language in EU documents is not

International Journal of Corpus Linguistics 9:2 (2004), **195–224.**
ISSN 1384–6655/E-ISSN 1569–9811

highly appreciated by an academic community. Translated texts in their own right are often seen as contradictory material: inconsistent, influenced by the original, but nevertheless important. Thus one is tempted to turn to evaluative and prescriptive analyses of these texts. In contrast, a descriptive analysis does not discriminate between good and bad translation, and grammatical and ungrammatical sentences, it just accepts the existing data as evidence in reality, which needs to be explained using appropriate methodology.

In this article, we would like to move the focus from product-oriented to process-oriented translation studies by putting forth an English – Lithuanian Phases of Translation Corpus (PT corpus). We will concentrate on processes that take place in successive Lithuanian translated versions.

A translated text is a result of many different factors: translator's personal choice, official regulations, and terminology. During its editing phase some elements in the language of translation are bound to be changed, such as terminology, vocabulary, grammar, and others. By using a combination of qualitative and quantitative methods on the Phases of Translation Corpus, we expect certain phenomena to be highlighted, which would be lost in an ordinary bilingual parallel corpus.

We are not making any prior assumptions in connection with the present research, as we are determined to apply a corpus-driven approach to the analysis of the process of translation, and all the steps of the research are based on the corpus evidence, and not on any prior theories.

2. Process-oriented translation studies

Until quite recently the absolute majority of scholarly work in descriptive translation studies has been product-oriented. Process-oriented research has been avoided not because it has been seen as unimportant, but because it has been very difficult (and still is) to get hold of objective "observational data", which is necessary for descriptive and empirical research.

In spite of these difficulties, a number of scholars have tried to account for the process of translation either by building a mental model of translation based on their own intuition, or relying on translator's retrospection, or relying on some experimentally acquired real-time data. All the attempts, which account for mental processes relying on a shared intuition or on other translator's retrospection, can be considered as speculative (Toury 1995: 233). It is relevant, therefore, to the present study to overview some research that made

use of some real-time data. By real-time data, here, we mean the data that is recorded during the process of translation.

The approaches to translation can be classified into *on-line approaches* and *off-line approaches*. Jakobsen (1998: 155) points out that on-line approaches attempt "to study translation while it takes place, whereas off-line approaches attempt to study translation after the event and rather more indirectly". In other words, on-line approaches deal directly with translator's behaviour, while off-line approaches are more focused on translation products.

What sort of real-time data can be recorded during the emergence of the final product of translation?

At present, we can identify three approaches that are dealing with the process of translation:

(a) think-aloud protocols (TAPs) *on-line*,
(b) *translog* system *on-line*,
(c) analysing successive draft versions *off-line*.

The first approach of *think-aloud protocols* (*TAPs*) is an on-line approach. TAPs can be defined as "concurrent verbalisation, or thinking aloud, which provides data on the mental states heeded by individual [translator] carrying out a task" (Bernardini 1999: 181). The performance of translator is recorded or videotaped. The internal process of translating is then judged by analysing the content of the verbal report itself, as well as translator's behaviour such as breaks in speaking, note-taking, speaking speed, or eye movements etc. However, it is argued that speaking aloud might interfere with the mental process, and that is why the approach cannot claim to account solely for the process of translating (Toury 1995: 235).

Jakobsen and Schou (1999) at the Copenhagen Business School (CBS) have recently developed the second on-line approach, known as *translog* system. The system allows recording the whole typing process during translating, so that a computer records the exact time of each keystroke. As a result one can study the entire process of emergence of a translated text by analysing all the real-time versions of it. This approach has been around only for a short time. The Copenhagen group TRAP (TRAnslation Process, at the Copenhagen Business School) has already done some experimentation with the system, and they tackle some of the the issues of translating in Hansen (1999).

We will present a more lengthy discussion of the third approach, which is the approach that has been selected for the present research. Here the data consist of successive written draft versions of translation, which are preserved

either on paper or on computer. The data may include the translator's first version and different revised versions, including or not including the final version.[1]

Analysing successive draft versions allows a researcher to spot difficult stretches of original text, since usually problematic cases are the ones that are being revised during the process of editing (Toury 1995). However, Toury (1995: 185) has primarily associated this approach with the printed text on paper. He has even predicted that "in the computer age, this kind of study would soon become impossible".

The present article proves that the situation is exactly the opposite, and that computers and corpus-based methods offer much greater possibilities than paper for storing and processing intermediate versions of translations.

3. Compilation of PT corpus

3.1 English-Lithuanian phases of translation corpus (PT corpus)

The article presents the *English-Lithuanian phases of translation corpus* of European community law documents. The corpus enables to conduct the process-oriented research of the language of translation, as Lithuanian translated texts were produced at three successive time points: the earliest version was the first translator's draft, the intermediate – second-edited draft, and the latest – the final version. These successive versions of translated text are aligned to corresponding English texts that were the original source of the Lithuanian translations. It must be noted, that the corpus thus compiled does not yet fit into existing classifications of translation corpora.

The possibility of creation of a *phases of translation corpus* has come about quite accidentally. While working on compiling the English-Lithuanian parallel corpus of EU documents at the University of Birmingham, we have noticed that a considerable number of Lithuanian documents have identical CELEX[2] numbers. The further investigation has shown that there are three types of special notes at the top of these documents such as:

Darbinis vertimas (vert.) – (translator) draft translation,
Darbinis vertimas – draft translation,
Autentiškas vertimas – authentic translation,

which turned out to be indications of different translated versions of EU English documents into Lithuanian:

(a) **the first translator's draft** (Lithuanian) – the first draft produced by a translator is the earliest translation, and therefore we may assume that it contains the most immediate language of translation,
(b) **the second edited draft** (Lithuanian) – the revised version of the first translator's draft,
(c) **the final version of translation** (Lithuanian) – the last version of translation, which is the goal of the earlier revisions, and which is considered to be the most suitable translation.

The Lithuanian files of translated EU documents were acquired from the "Centre of Translation, Documentation, and Information" in Vilnius. There were 1,689 files in total. Most of these files had only one existing version, but quite a few were in two or three versions. In order to avoid the confusion of dealing with different sizes of each version, we decided to work only with the files that existed in all three versions. Meanwhile, the corresponding English files had to be downloaded from the CELEX database of EU documents.

Subsequently all the three versions were aligned to each other and to the English documents at the sentence level (see Figure 1 below).

The whole corpus consists of 35 files in each version (i.e. an original English text and 3 Lithuanian versions). There are 112,745 words in EN files; 84,647 in D1 files; 84,655 in D2 files; and 85,034 in F3 files; a total of 367,081 words. No meta-data is added to the corpus except tags distinguishing between English and different Lithuanian versions, as we think that plain text corpora are among most easily shared resources.

<EN> 3. *Animals which might injure each other on account of their species, sex, age or origin must be kept and lairaged apart from each other.*

<D1> 3. *Gyvūnai, kurie gali vienas kitą sužeisti dėl to, kad yra tam tikros rūšies, lyties, amžiaus arba kilmės, turi būti atskirti ir patalpinti į atskirus gardus.*

<D2> 3. *Gyvūnai, kurie gali vienas kitą sužeisti dėl to, kad yra tam tikros rūšies, lyties, amžiaus arba kilmės, turi būti atskirti ir suvaryti į atskirus gardus.*

<F3> 3. *Gyvūnai, kurie gali vienas kitą sužeisti dėl savo veislės, lyties, amžiaus arba kilmės, turi būti atskirti ir suvaryti į atskirus gardus.*

Figure 1. An aligned sentence from the English – Lithuanian PT corpus

The PT corpus represents an off-line approach to the process of translation. The corpus does not have any temporal information such as how long it took to produce different versions, or how many and what sort of revisions could take place before the emergence of the final version of a document, nor any information about the number of persons involved in the revision phases after the first translator's draft was produced. We think that the above information is not necessary for our purposes, as we want to focus on the changing language of translation only and not on translator's mental activity or any other psychological aspects.

3.2 Alignment

The PT corpus has been aligned using the program "Vanilla Aligner", which has been developed by Pernilla Danielsson and Daniel Ridings (1997).[3] In the first phase, "Vanilla Aligner" aligns chosen "hard delimiters" (in our case paragraphs), and then it aligns "soft delimiters" (usually sentences) within each hard-delimited chunk of text. The program uses the Gale and Church algorithm, so that each pair of sentences within hard-delimited text is assigned a probabilistic score, which is "based on the ratio of lengths of the two sentences (in characters) and the variance of this ratio", and then "a dynamic programming framework is used for finding the maximum likelihood alignment of sentences" (Gale & Church 1993: 79).

Before proceeding to the alignment, "Vanilla Aligner" requires some preprocessing of input files. The input files should have "soft delimiters" (sentence boundaries) and hard delimiters (in our case paragraph boundaries) explicitly marked up (or segmented) as well as tokenised into the form of one word per line. The tasks of inserting paragraph boundaries and tokenising can be solved fully automatically; and they do not present any real difficulties. For performing segmentation of paragraphs a simple Perl script replaces paragraph marks of a text by a tag (in our case ".EOP" for End Of Paragraph) based on the existence of an extra line break between paragraphs, which is treated by the aligner as the end of paragraph. Similarly, for tokenization the script replaces spaces between words with end of line marks to get the required one word per line.

The task of sentence segmentation is somewhat more complicated than just a straightforward replacement of strings. Obviously, the initial assumption is that the punctuation marks ".!?..." signal the end of a sentence. However, we do not want the program to place sentence boundaries after initials and shortenings in the middle of a sentence, as for example, in English after *etc., cf.,*

pp., and in Lithuanian after *t.t., žr., p*. Simple heuristics enable us to update our Perl program so that it does not treat certain strings in English and Lithuanian as the end of a sentence.

It should be noted though that during segmentation the definition of a sentence is treated quite loosely. For example, we would also mark the boundary of a sentence before each paragraph mark even if it is proceeded by a comma (,). In many cases, it is reasonable to treat a semi-colon or a colon as the end of a sentence.

The decision whether to treat a particular punctuation mark or a string of characters as the end of a sentence depends on the type of text. If we dealt with coherent literary prose, the text would be segmented in a more conventional way, and in contrast, a poem would be segmented keeping to very different principles (more on problems of segmentation and tokenisation in Grefenstette & Tapanainen 1994). An aligned segment of a reasonable size makes the analysis of the parallel corpus easier. In our case, we have a collection of legal documents, which tend to have quite large segments of language that are not separated by conventional sentence boundaries, but which are independent enough to be considered as proper candidates for alignment.

After the tokenization and segmentation of files is completed, we can proceed with the next step of compilation of the PT corpus, which is the actual alignment. The major challenge of aligning the phases of translation corpus (the PT corpus) lies in the fact that "Vanilla Aligner" (as the majority of aligners) is designed for aligning two text files, whereas our project requires four parallel files to be aligned. Therefore the alignment of the PT corpus consisted of two stages.

In the first stage, we have aligned two pairs of files separately. As a result, we have produced two aligned files, where the English version has been aligned to the second draft (EN-D2), and the first draft to the final version (D1-F3), see table below.

Table 1. An extract from files after the first stage of alignment

EN-D2 file	*D1-F3 file*
<EN> 2. Animals must be unloaded as soon as possible after arrival.	*<D1> 2. Gyvūnai turi būti iškeliami iš karto po to, kai atvežami*
<D2> 2. Gyvūnai turi būti iškeliami iš karto juos atvežus.	*<F3> 2. Gyvūnai turi būti iškeliami iš karto juos atvežus.*

In the second stage, we have aligned the result files of the first stage, so that the line of EN is aligned to D1, and D2 to F3. As a result of this, we have produced a file where all the four sentences are aligned to each other:

<EN> 2. Animals must be unloaded as soon as possible after arrival.
<D1> 2. Gyvūnai turi būti iškeliami iš karto po to, kai atvežami.
<D2> 2. Gyvūnai turi būti iškeliami iš karto juos atvežus.
<F3> 2. Gyvūnai turi būti iškeliami iš karto juos atvežus.

The alignment of the PT corpus involved solving typical aligning problems, as well as some specific ones. Commonly, the process of alignment consists of a number of aligning iterations until the correct alignment is achieved. The process involves the following steps:

1st step: the first alignment,
2nd step: spotting errors and mis-alignments in the aligned file, finishing the process if it is correct,
3rd step: editing the intermediate source or target files,
4th step: going back to the 1st step.

The absolute majority of mis-alignments is caused by the different number of "hard delimiters" in two parallel files. As we chose paragraph marks as "hard delimiters", it is essential that both files would contain the same number of paragraphs. The different number of paragraphs is usually due to translator's errors, different layout of files, different splitting of titles, or additional information in one of the files. The correction involves either removing erroneous paragraph marks, or inserting empty paragraph boundaries in one of the files in the case of extra information in the other. Such insertions and deletions of paragraph delimiters are performed until both files have the same number of paragraph marks. As long as these differences are systematic, we can insert the appropriate tags automatically, and there is no other way except manual to correct any non-systematic differences.

Alignment of sentences causes yet another set of complications. The aligner is more likely to fail when some sentences that occur in the source text are omitted in translated versions, or when new sentences are added. Most other problems are caused by segmentation errors, which can be corrected simply by updating the program for segmentation or by correcting sentence delimiters manually.

4. Bilingual and multilingual parallel concordancers

At least two concordancing programs need to be mentioned, they are: MULTICONC developed by David Woolls at the Birmingham University (Woolls 1997), and PARACONC developed by Michael Barlow at the Rice university (Barlow 1995).[4] Both these programs have several features in common: they are both designed to produce and manipulate multilingual parallel concordances, they work on the Microsoft Windows platform, and they are well known. There is one important difference, however, which conditioned our preference for PARACONC rather than MULTICONC. Although it is claimed that MULTICONC is a multilingual parallel concordancer, "the actual concordancing is done with any two languages at any one time" (Ulrych 1997: 429). Thus the program can produce only bilingual concordances. For example, if one wants to compare English, German, and French, one needs to create concordances for English – German, English – French, and German – French pairs.

Meanwhile, PARACONC can create multilingual concordances for up to four languages. This ability of the program is especially relevant for the analysis

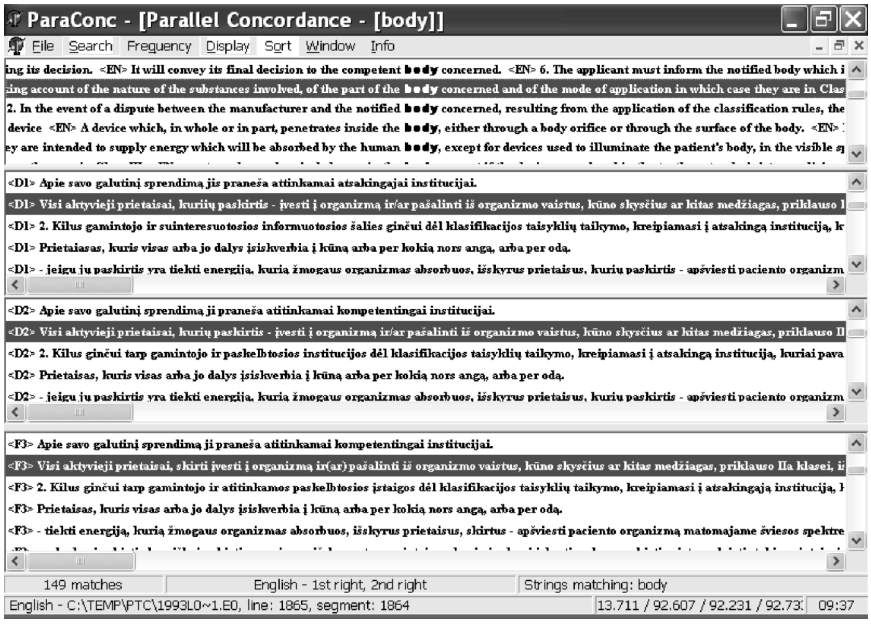

Figure 2. A concordance of "*body*" from PT corpus by PARACONC

of the phases of translation corpus. Although the corpus is bilingual, it consists of four parallel texts: one English and three Lithuanian (see Figure 2).

The program also provides possibilities of regular and advanced search, which enables the extraction of relevant concordance lines from the source language. Then the extracted concordance lines can be sorted and counting of collocates for the source language can be performed.

PARACONC has also a function that helps to explicate the most likely translational equivalents in the target language. Basically, the program creates a ranked list of words in the target text, which occur unusually frequently in a concordance if compared with their frequencies in the whole corpus. The program refers to these words as *hot words*. One can then pick up the correct translational equivalents from the list, and highlight them in the target text. The program now knows the node in concordance lines of the target text, and thus these lines can also be sorted and otherwise analysed.

5. Analysis of phases of translation corpus

5.1 Quantitative analysis

5.1.1 *Frequency lists and type-token ratio*

As it was mentioned above the PT corpus consists of just 35 legal documents in each version (112,745 words in the original English version; 84,647 in the first draft; 84,655 in the second draft; and 85,034 in the final version).

A corpus of this size imposes considerable practical constraints, as we can only analyse the most frequent lexical items. For example, most multi-word units or collocations are too infrequent to provide reliable evidence. Hence, analysis must be limited to the most frequent single words and their patterns.

It should also be noted that the PT corpus reflects phenomena in a very specific language domain, which is the language of translation of the EU legislation. Therefore any tendencies that may be observed in this corpus can be irrelevant or only partially relevant to other domains of the language of translation. For example, we cannot compare the process of translation of an EU document to translating of a literary novel, as the former is more aimed at rendering the meaning of the source language in a very exact and formal way, while the latter may rather be more concerned with artistic techniques and originality.

Table 2. An extract from frequency lists

No	*EN*	*freq.*	*D1*	*English equivalent*	*freq.*	*D2*	*freq.*	*F3*	*freq.*
1	the	9855	ir	(and)	2848	ir	2841	ir	2870
2	of	5193	ar	(or)	911	ar	913	ar	904
3	to	3469	būti	(to be)	660	būti	668	arba	650
4	in	3321	arba	(or)	640	į	644	į	637
5	and	3125	į	(to)	635	arba	636	būti	621
6	a	1662	kad	(that)	621	kad	603	kad	618
7	or	1557	turi	(have)	611	turi	577	**straipsnis**	**524**
8	be	1542	yra	(is)	537	**straipsnis**	**525**	turi	521
9	for	1483	**straipsnis**	**(article, Nom.)**	**535**	yra	523	su	517
10	**article**	**1226**	su	(with)	510	kadangi	504	yra	508
11	shall	1207	kadangi	(as/since/because)	489	su	489	kadangi	507
12	on	972	pagal	(according)	472	pagal	476	**straipsnio**	**505**
13	with	923	dėl	(due to/because of)	456	dėl	456	dėl	453
14	by	917	direktyvos	(directive/directives)	449	direktyvos	438	pagal	445
15	directive	836	gali	(can/may)	418	**straipsnio**	**438**	direktyvos	434
16	which	823	**straipsnio**	**(article, Gen.)**	**409**	gali	415	gali	413
17	this	809	eec	(eec)	401	eec	405	eec	408
18	member	745	kaip	(like/as)	390	valstybės	392	valstybės	394
19	as	723	narės	(members)	365	kaip	386	narės	373
20	is	657	atliekos	(waste/scrap)	362	bei	372	kaip	372
21	from	622	apie	(about/around)	361	narės	371	bei	369
22	not	598	bei	(and)	361	atliekos	359	atliekos	363
23	that	564	valstybės	(states/state's)	345	apie	351	apie	344
24	states	544	jei	(if)	307	a	336	jei	326
25	are	544	a	(–)	305	jei	323	iš	315
26	must	526	iš	(from)	299	d	310	d	311
27	whereas	512	**straipsnyje**	**(article, Loc.)**	**293**	iš	300	a	309
28	other	449	d	(–)	284	**straipsnyje**	**287**	būtų	278
29	may	437	šios	(these)	267	m	263	m	268
30	eec	428	jų	(their)	260	būtų	256	**straipsnyje**	**261**

First of all, we have produced four frequency lists for PT corpus: one for English and three for different Lithuanian versions (see Table 2). In English, not at all surprisingly, the most frequent word is the article *the* followed by prepositions *of, to,* and *in*, while in Lithuanian the two most frequent words are the conjunction *ir* followed by the other conjunction *ar*.[5] The fact that there are no corresponding translation equivalents for the most frequent English words *the, of,* and *in* is a significant one, since it confirms that we are dealing with very different language systems.

Lithuanian is a highly inflectional language. Where English employs various prepositions, Lithuanian makes use of endings, prefixes, and suffixes. This can also explain why in Table 2 the top English words are much more frequent than the top Lithuanian words. Although the first noun *article* (*straipsnis*) appears very high on all the lists, its frequency is 1226 on the English list to compare with just 535, 525 and 524 on Lithuanian lists. This is due to the fact that frequencies of the Lithuanian noun are spread across its different cases: nominative – *straipsnis,* genitive – *straipsnio,* locative – *straipsnyje* and others.

Clearly, lemmatisation would make some changes in frequency order by boosting the frequencies of the lemmatised entries, and thus moving them higher in the frequency list. This kind of generalisation, however, would hide the actual frequencies of different lemmata and it is not always considered as a positive practice. The preference of form to lemma is especially emphasised in John Sinclair's works, who claims that "each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity" (Sinclair 1991: 8). This assumption is also supported and exemplified in studies by Stubbs (1996: 172–173) for English, by Tognini-Bonelli (1996: 124–131) for Italian, and by Utka (forthcoming) for Lithuanian. This principle is particularly important for Lithuanian, where noun endings add different shades of meaning to words not just plurality or gender.

Type/token ratio is useful in evaluating the richness of vocabulary. Table 3 below shows statistics for the size of subcorpus (*tokens*), all different words (*types*), part of types (*types %*), and repetition of each type (*type repetition*). Not surprisingly each English type is repeated on average 20,9 times to compare with just 7,0–7,3 times for Lithuanian texts.

The values of type/token ratio show that each subsequent version is more repetitive than the previous one: a type in the first translator's draft is repeated an average of 7 times, in the second draft 7.2 times and in the final version 7.3

Table 3. Type and token statistics

Subcorpora	Tokens	Types	Types (%)	Type Repetition
English documents (E0)	112 745	5 386	4.78%	20.9
First translator's draft (D1)	84 647	12 119	14.32%	7.0
Second draft (D2)	84 655	11 822	13.96%	7.2
Final version (F3)	85 034	11 707	13.77%	7.3

times. The reduction of types in the editing phases can reflect the tendency of editors to normalize the vocabulary of the translated language.

Normalization is often seen as one of the universals of translation (Laviosa-Braithwaite 1998: 288), which is characteristic to all translated texts. The phenomenon of normalization has been commonly associated with the translated language as compared to the original language; in this case, however, we have probably come across the normalization of lexis during the phases of the translation. In order to claim this we need to have a closer look at the types that have been discarded in the later editing phases.

5.1.2 *Comparison of frequency lists*

The calculations of types and tokens in the previous section have shown that the number of types is decreasing with each version of translation (see Table 3). The first obvious and easily realizable step is to track down which word types have been removed or replaced in the later stages of translation. This analysis may point to specific problems in translated texts, which have been handled consistently by replacing or removing all the occurrences of a particular type with a different one. By comparing frequency lists to each other we have produced lists of words that are present in one subcorpus, but are not found in the other. See Table 4 for numbers of missing types for all possible comparison pairs.

If we compare D1 and D2 frequency lists, we find out that 933 types are missing in version D1 which means that they are newly introduced in version D2, and 1229 types that have been present in version D1 have disappeared from version D2; the difference 296 shows that from version D2 have disappeared 296 types more if compared to the number of newly introduced types. The numbers show that during the first editing phase many more types (296) have disappeared than during the second (115), while the numbers of newly introduced types are similar (933 and 911).

Yet another comparative study between our frequency lists has been performed by joining the frequency lists of versions D1 and D2 and comparing them to version F3. Thus D1-D2 list represents all word types that have oc-

Table 4. Statistics of missing types

Compared pairs	Types missing in D1	Types missing in D2	Types missing in F3	Diff.
D1 – D2	933	1229	×	296
D1 – F3	1480	×	1891	411
D2 – F3	×	911	1026	115

Table 5. Statistics of missing types

Compared lists	Types missing in D1/D2	Types missing in F3
D1/D2 – F3	807	2151
	Types missing in D1	*Types missing in D2/F3*
D1 – D2/F3	1740	1125

curred in the versions prior to the final. We have also joined D2 and F3 lists, and compared them to D1, where D2-F3 list represents all the types that have occurred after the first translator's draft has been produced (see Table 5).

The comparison between D1-D2 and F3 lists yields particularly interesting results. While there are just 807 newly introduced types in F3, 2151 word types, which have been present in D1-D2, have disappeared from the final version (F3) altogether. The numbers, clearly, show that there has been a considerable reduction of types in the final version if compared to the two previous versions. This implies that there is a strong tendency to normalize lexis of translated texts in the process of editing.

Let us consider the actual words that are introduced and disappear in the phases of editing. We have produced lists of word forms that are missing from one of the versions, but are present in the other (see Table 6 below). The numbers of frequencies show how many times a word form occurs in the version which a given frequency list is compared to. The lists are sorted in frequency order, so that we get the "most wanted" types for different versions at the top of the list.

The investigation of missing types allows tracing a complete removal of a word form or a replacement of one word form by another in the later editing stages. For example, the word *šlamai* (*sludges*), which occurred 80 times in the version D1, has been replaced by the more general Lithuanian word *dumblai* in D2. While *šlamai* is missing in D2, *dumblai* is missing in D1.

The parallel concordancing has helped us to extract a whole chain of replacements starting with English original (E0) and ending with the final translated version (F3). We have only considered the missing types that have occurred more than 10 times in the PT corpus. The procedure of extracting this chain has consisted of the following steps:

– a parallel concordance is made for a missing type from Table 6, so that if a term is missing in one subcorpus we look for it in the other;
– the corresponding terms from the other three subcorpora are picked up (see Table 7 below) and counted.

Table 6. The most frequent missing types

D1 compared to D2				D2 compared to F3			
Types missing in D1		*Types missing in D2*		*Types missing in D2*		*Types missing F3*	
dumblai	94	šlamai	80	autentiškas	35	darbinis	35
paskelbtoji	54	notifikuotoji	54	iia	33	viešosios	12
halogenintų	21	vert	25	iib	23	etiketavimą	9
paskelbtajai	17	instaliacijos	23	centras	20	įdedamasis	8
halogenintos	15	halogeninių	19	lr	19	etiketavimo	8
sertifikatą	13	halogeninės	15	elektrotechniniai	12	ge	8
paskelbtosios	13	šlakai	15	informacinis	11	išsami	8
gstn	13	angliarūgštės	14	elektrotechninis	10	viešoji	8
monitoringas	12	iia	13	antrinės	7	paragrafo	7
halogeninti	12	gftn	13	pakavimą	7	svarstant	7
destiliacijos	12	halogeniniai	12	sudarant	7	betarpės	6
vaisto	11	pažymėjime	11	informacinio	6	pagrinde	6
tyrės	11	vertėjas	10	respublikos	6	stebėjimą	6
spinduliuotės	11	tyrelės	10	tiesioginiam	6	dauginimosi	5
sertifikate	11	kompetetingos	9	elektrotechninio	5	nepridedant	5
vaistas	9	švitinimo	9	galutinis	5	neprieštaraujant	5
išsiskyrimo	9	auksčiau	9	išmetimo	5	pagrindinai	5
antagonistų	9	atšaukimo	9	nededant	5	periodo	5
užtikrinimas	8	raguočių	8	paminėtos	5	sutinkamai	5

Table 7. An extract from parallel concordance of "*notified body*"

<EN> The applicant must make a 'type' available to the **notified body**.
<D1> Pateikiantis prašymą asmuo **notifikuotoji įstaigai** turi pateikti "tipinį pavyzdį" ...
<D2> Pateikiantis prašymą asmuo **paskelbtajai įstaigai** turi pateikti "tipinį pavyzdį".
<F3> Prašymą pateikiantis asmuo **paskelbtajai įstaigai** turi pateikti tipinį pavyzdį.

<EN> The **notified body** may request other samples as necessary,
<D1> Esant reikalui, **notifikuotoji įstaiga** gali pareikalauti ir kitų pavyz ...
<D2> Prireikus, **paskelbtoji įstaiga** gali pareikalauti ir kitų pavyzdžių,
<F3> Prireikus **paskelbtoji įstaiga** gali pareikalauti ir kitų pavyzdžių,

<EN> 4. The **notified body** must:
<D1> 4. **Notifikuotoji įstaiga** privalo:
<D2> 4. **Paskelbtoji įstaiga** privalo:
<F3> 4. **Paskelbtoji įstaiga** privalo:

A closer investigation of these replacements gave the following observations:

- more voluminous replacements have taken place in the first editing phase (D1-D2) if compared to the second (D2-F3);
- the disappearance of a more frequent word in a later version usually implies that the systematic replacement of one term by the other has taken place. However, "clean" replacements, when the whole lemma of a term is replaced by a lemma of some other term, are quite rare;
- a case of simplification is found, when five English words (*wastes, sludges, slag, muds, dross*) are gradually reduced just to one Lithuanian word (*dumblai*);
- Latinization of Lithuanian terms can be observed. While in some cases Latin roots are replaced by Lithuanian (capitalised in the example), for example, *NOTIFIKuotoji* into *paskelbtoji* (*notified*), *INSTALIACijos* into *įrengimai* (*installations*), in other cases we see Lithuanian roots translated into Latin, for example, *tyrimas* into *MONITORINGas* (*monitoring*), *pažymėjimas* into *SERTIFIKATas* (*certificate*);
- the method is also useful for finding cases of "translationese" (the term coined by Gellerstam (1986)), as they tend to be replaced systematically in later editing phases. For example, in D1 the English word *medicinal product* is often translated on a word-by-word basis by the unnatural compound *vaistų gaminys*. Clearly, the translation here has been influenced by the source language. *Vaistas,* a more conventional term in Lithuanian, is chosen for D2 and F3 versions. The translations of *installation* as *instaliacijos, įpakavimo lapelis* as *package leaflet,* and *public limited-liability company* as *viešoji ribotos atsakomybės bendrovė* are other instances of translationese.

The findings are interesting as far as exemplification of the above-mentioned phenomena is concerned. However, we cannot claim that they represent characteristic tendencies in the process of translation, as most of the words are relatively infrequent and most of these replacements occur within a single file.

As the PT corpus is too small to provide reliable evidence that is based on missing types, we need to consider more frequent words that occur at the top of a frequency list.

5.1.3 *Comparing frequencies of types*

The differences between type frequencies on the three frequency lists, which are in the first translator's draft, the second draft, and the final version, are likely

Table 8. Comparison of type frequencies in D1 and D2 frequency lists

Word	English equivalent	Freq. in D1	Freq. in D2	Difference	Difference (%)
ir	(and)	2848	2841	7	0.25%
ar	(or)	911	913	–2	0.22%
būti	(to be)	660	668	–8	1.20%
į	(to)	635	644	–9	1.40%
arba	(or)	640	636	4	0.62%
kad	(that)	621	603	18	2.90%
turi	**(have)**	**611**	**577**	**34**	**5.56%**
straipsnis	(article)	535	525	10	1.87%
yra	(is)	537	523	14	2.61%
kadangi	(as/since/because)	489	504	–15	2.98%
su	(with)	510	489	21	4.12%
pagal	(according)	472	476	–4	0.84%
dėl	(due to/bacause of)	456	456	0	0.00%
direktyvos	(directive/directives)	449	438	11	2.45%
straipsnio	**(article)**	**409**	**438**	**–29**	**6.62%**
gali	(can)	418	415	3	0.72%
eec	(eec)	401	405	–4	0.99%
valstybės	**(states/state's)**	**345**	**392**	**–47**	**11.99%**
kaip	(as)	390	386	4	1.03%
bei	(and/as well as)	361	372	–11	2.96%

to spotlight the types that have been treated differently at the different phases of translation. Seeing that the manual comparison of two frequency lists is a very ineffective and tedious practice, we have used an automatic comparison of frequency lists. See Table 8 for comparison of frequencies of the 20 most frequent types in D1 and D2.

Differences between frequencies are represented by a percentage, as it gives an approximate value of relative difference.[6] For the present research the differences that are more than 5 per cent have been considered as noteworthy and looked more closely at. In Table 8 the frequencies of the words *turi* (*has/have*), *straipsnio* (*article*), and *valstybės* (*state*) have changed more than 5 per cent: frequencies of *straipsnio* and *valstybės* have increased by 6.62 and 11.99 per cent, while the frequency of *turi* has decreased by 5.56 per cent. And what is happening in the last stage? See Table 9 for comparison of frequency lists of D2 and F3.

The frequency of *turi* continues to diminish by 9.7 per cent, the frequency of *straipsnio* continues to grow by as much as 13.27 per cent, while the frequency of *valstybės* stabilises. In the last stage we have other words that have

Table 9. Comparison of type frequencies on D2 and F3 frequency lists

Word	English equivalent	Freq. in D2	Freq. in D3	Difference	Difference (%)
ir	(and)	2841	2870	–29	1.01%
ar	(or)	913	904	9	0.99%
būti	**(to be)**	**668**	**621**	**47**	**7.04%**
į	(to)	644	637	7	1.09%
arba	(or)	636	650	–14	2.15%
kad	(that)	603	618	–15	2.43%
turi	**(has)**	**577**	**521**	**56**	**9.71%**
straipsnis	(article)	525	524	1	0.19%
yra	(is)	523	508	15	2.87%
kadangi	(as/since/because)	504	507	–3	0.59%
su	**(with)**	**489**	**517**	**–28**	**5.42%**
pagal	**(according to)**	**476**	**445**	**31**	**6.51%**
dėl	(due to/bacause of)	456	453	3	0.66%
direktyvos	(directive/directives)	438	434	4	0.91%
straipsnio	**(article)**	**438**	**505**	**–67**	**13.27%**
gali	(can)	415	413	2	0.48%
eec	(EEC)	405	408	–3	0.74%
valstybės	**(states/state's)**	**392**	**394**	**–2**	**0.51%**
kaip	(as)	386	372	14	3.63%
bei	(and/as well as)	372	369	3	0.81%

considerable frequency falls or rises: while *būti* and *pagal* have falls of 7.04 and 6.51 per cent respectively, *su* has a rise of 5.42 per cent.

Thus we have 6 types – *turi (has/have), straipsnio (article), valstybės (state), būti (to be), su (with), pagal (according to)* – which have significant frequency fluctuations across the three versions of Lithuanian translated texts, and which belong to 20 most frequent types in these texts. While the former fact is important as it is likely to highlight problematic cases during different phases of translation, the latter fact ensures that the further findings are sufficiently frequent to lead to plausible generalisations.

We arbitrarily have chosen to look more closely at three of the six types that have significant frequency fluctuations: the noun – *valstybės* (*state*), the verb – *turi* (*have*), and the preposition – *pagal* (*according to*).

At this stage the potential of quantitative analysis has been exhausted, as we have come to the point where the question should be answered what could possibly cause these fluctuations of frequencies. In order to answer this question we need to have a closer look at the language patterns that surround these

types. In other words we have come to the stage of qualitative analysis, which is presented in the following sections.

5.2 Qualitative analysis

5.2.1 *Analysis of the noun* "valstybės" *(*state*)*

Before looking at the parallel concordances, it is relevant to give a brief description of grammatical features of the Lithuanian noun. Most Lithuanian nouns have 7 grammatical cases (nominative, genitive, dative, accusative, locative, instrumental, and vocative) both in plural and in singular, which can be of feminine or masculine gender. These grammatical features, namely case, number, and gender, are expressed with different endings.

Although the quantitative analysis has singled out as a significant case just one word form from the whole lemma, we cannot overlook the fact that one word form represents just one-fourteenth part of the lemma. So the question is whether other constituents of the lemma show similar extent of fluctuations of frequencies. The answer can be found in Table 10.

Table 10 demonstrates that singular genitive or plural nominative of *valstybės* are the most problematic cases as compared to the other cases of the lemma. The next interesting case is plural genitive that rises by 11 in the second draft (D2). It can also be observed that frequencies of all the cases of *valstybė* tend to rise in version D2, while they stabilise in the version F3. In the later analysis we will concentrate on evidence for the types *valstybės* apart from

Table 10. Lemmata of "valstybė"

Type	Number/Case	D1	D2	D2-D1	F3	F3-D2
valstybės	**sg.gen/pl.nom**	**345**	**392**	**+47**	**394**	+2
valstybių	pl.gen	121	132	+11	134	+2
valstybėms	pl.dat	75	82	+7	84	+2
valstybė	sg.nom	58	66	+8	66	0
valstybėse	pl.loc	40	40	0	45	+5
valstybėje	sg.loc	29	33	+4	33	0
valstybei	sg.dat	9	11	+2	10	–1
valstybes	pl.acc	4	7	+3	5	–2
valstybėmis	pl.instr	1	3	+2	3	0
valstybe	sg.instr	2	1	–1	1	0
valstybėms	pl.dat	0	1	+1	0	–1
valstybė		**684**	**768**	**+84**	**775**	+7

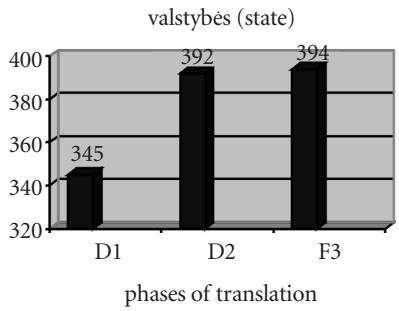

Figure 3. Changing of frequencies of the type "*valstybės*" (*state*)

their lemma, as the other constituents of the lemma are not so frequent, and fluctuations of their frequencies are not so distinct.

The type of *valstybės* (*state*) (as well as s*traipsnio* (*article*), *direktyvos* (*directive*) and *narės (member))* shows exceptionally high frequency for a content word. On the frequency list (see Table 2 in Section 5.1.1) its neighbours are function words such as *pagal (according to) 445–476, dėl (due to/because of) 453–456, kaip (as) 372–390,* and *bei (and/as well as) 361–372.* The equivalent English word *states* is also very high on the English list with frequency 544. Possibly, the fact implies that such types in EU documents are not simply conventional nouns, but they have acquired features of function words.

What are meaningful patterns associated with the word? The computer programs PARACONC and WordSmith Tools (Scott 1996) have helped us to look more closely at the relevant patterning.

First of all, the analysis of concordance for *valstybės* (*state*) has shown that the type *valstybės,* as well as other constituents of the lemma belong almost unanimously to the compound *valstybės narės (Member States)*. This is understandable as EU legislation is primarily directed towards EU Member States. The question is, however, why the term *valstybės narės* occurs almost 50 times less in the immediate version of translation (D1) than in later versions (see Figure 3).

In Table 11 below, the commonest patterns of *valstybės* are presented.

The majority of patterns in Table 11 and the other patterns that have been left outside the table reveal the authoritative manner of the analysed texts. *Member States*, the primary object of regulations and directives, are told what they *may do, need to do, shall do,* and *should do.*

Table 11. The common patterns of the type "*valstybės*" (*state*) in D1, D2, and F3

Pattern	D1	D2	F3
valstybės narės gali	*61*	*63*	*61*
(Member States may / need)	*18%*	*16%*	*15%*
valstybės narės, kurioje / kurios / kuri	*25*	*27*	*26*
(Member States in which / where)	*7%*	*7%*	*7%*
valstybės narės turėtų / turi	*21*	*15*	*15*
(Member States shall take)	*6%*	*4%*	*4%*
valstybės narės imasi	*18*	*23*	*22*
(Member States shall take / shall adopt / adopt)	*5%*	*6%*	*6%*
valstybės narės priima / priėmė / priims / priimdamos	*16*	*17*	*14*
(Member States adopt / shall bring into force)	*5%*	*4%*	*4%*
valstybės narės teritoriją / teritorijoje / teritorijos	*14*	*15*	*15*
(in / to the teritory of a Member State)	*4%*	*4%*	*4%*
valstybės narės privalo	*11*	*26*	*24*
(Member States shall ensure / should recognize)	*3%*	*7%*	*6%*
Others	*179*	*206*	*217*
	52%	*53%*	*55%*
TOTAL	*345*	*392*	*394*

The investigation shows that in D1 *Member states* are sometimes translated by synonymous expressions *šalys narės* or *bendrijos narės (members of community):*

<EN> 1. ***The Member States*** *shall take all appropriate measures to ensure that ...*
<D1> 1. ***Bendrijos narės*** *turi imtis visų reikiamų priemonių, siekiant ...*
<D2> 1. ***Valstybės narės*** *privalo imtis visų reikiamų priemonių, siekdamos ...*
<F3> 1. ***Valstybės narės*** *privalo imtis visų reikiamų priemonių siekdamos ...*

<EN> ***The Member States*** *shall take all appropriate measures to ensure that if ...*
<D1> ***šalys narės*** *turi imtis visų reikiamų priemonių, siekiant užtikrinti, kad ...*
<D2> ***Valstybės narės*** *privalo imtis visų reikiamų priemonių, siekiant ...*
<F3> ***Valstybės narės*** *privalo imtis visų reikiamų priemonių siekdamos ...*

As could be expected the later versions have corrected this irregularity, which caused the frequency drop in the type *valstybės.*

Frequency fluctuations of word types at the top of frequency lists have enabled us to find a structural pattern *valstybės narės.* In spite of the fact that this construction plays a central role in the analysed texts, the earlier versions of translation have not always consistently translated it.

5.2.2 *Analysis of the Preposition* "pagal" *(*according to*)*

The preposition *pagal (according to / in accordance with)* is among the most frequent prepositions in Lithuanian. Grammatically it requires to be followed by an accusative case. It has been shown in Section 5.1.3.2, that *pagal* has similar frequencies in D1 and D2 versions, but has a fall in the final version (see Figure 4 below).

Investigation of parallel concordances has shown that the type *pagal (according to)* is most often used in cross-referencing constructions (see Table 12 below).

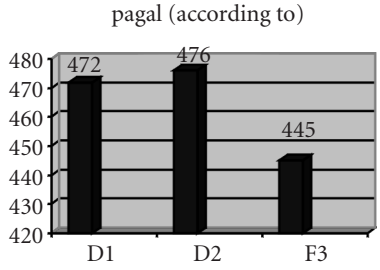

Figure 4. Changing of frequencies of the type "*pagal*" (*according to*)

Table 12. The common patterns of the type "*pagal*" (*according to*) in D1, D2, and F3

Pattern	**D1**	**D2**	**F3**
pagal # straipsnio / straipsnyje / straipsnį #	*82*	*91*	*76*
(*in accordance with Article #, according to Article #, persuant to Article #)*	*17%*	*19%*	*17%*
pagal direktyvos / direktyvą / direktyvas #EEC	*35*	*41*	*43*
(in accordance with Directive #EEC)	*7%*	*9%*	*10%*
pagal šią direktyvą, pagal šios direktyvos	*28*	*31*	*27*
(in accordance with this Directive)	*6%*	*7%*	*6%*
Other	*327*	*313*	*299*
	69%	*66%*	*67%*
TOTAL	*472*	*476*	*445*

The most frequent patterns in Table 12 account just for 31% of the usage of *pagal.* Generalisation could go only as far as this, because the remaining patterns are too infrequent to be generalised. We have found a number of identical English patterns, which have been translated inconsistently across all the versions, we could not find any explanation or justification for such choices:

IN ACCORDANCE WITH ARTICLE #
pagal # straipsnį (D1, D2, and F3)
kaip nustatyta # straipsnyje (D1, D2, and F3)
kaip numatyta # straipsnyje (D1, D2, and F3)
kaip numato # straipsnis (D1, D2, and F3)

IN ACCORDANCE WITH DIRECTIVE #
pagal Direktyvą # (D1, D2, and F3)
kaip nurodyta Direktyvoje # (D1, D2, and F3)
textitDirektyvose (D2)
Direktyvos (D2)

PURSUANT TO ARTICLE #
pagal # straipsnį (D1, D2, and F3)
kaip numatyta # straipsnyje (D1, D2, and F3)
kaip tai numato # straipsnis (D1, D2, and F3)
nurodytų # straipsnyje (D2)
vadovaujantis # straipsniu (F3)
laikantis # straipsniu (F3)
 # straipsnyje numatyta procedūra (D1, D2, and F3; procedure pursuant to Article #)

Thus *in accordance with Article #* is most often translated into *pagal # straipsnį*, but in quite a few cases the translator employs the construction *kaip nustatyta / numatyta / numato # straipsnyje (as defined in # Article)* when translating the phrase.

While in Table 12 we could represent the most frequent expressions for *pagal*, the list is not at all exhaustive. This is due to the specific characteristic of very frequent grammatical words, which is sometimes referred to as *collocational neutrality* (Halliday 1966), in other words, they are collocationally unrestricted. Grammatical words owing to their multifunctional nature and high frequency belong to a great variety of patterns. These patterns can only be generalised by grammatical frameworks. The following grammatical framework can define the usage of *pagal:*

... *PAGAL* + (intervening words) + NP (Acc.) ...

The above framework indicates, that *pagal* requires to be followed by a noun phrase, where a noun (its head) is in the accusative case. There can also occur any number of intervening words in between, which are very often in genitive case ("of-phrase" equivalent), as in the following concordance lines:

...medžiagų klasifikavimas PAGAL didžiausio laipsnio pavojingumą...
(classification of dangerous substances according to the greatest degree of hazard)
...PAGAL 3 straipsnio nuostatas...
(in accordance with the provisions of Article 3)
...PAGAL gamintojo valstybėje narėje galiojančių standartų saugos reikalavimus...
(in accordance with the safety provisions of the standards in force in the Member State of manufacture)

In Lithuanian the meaning of the prepositional framework *pagal+NP (Acc.)* can be achieved by other competing frameworks such as:

vadovaujantis + NP (Instr.)
remiantis + NP (Instr.)
NP (Instr.)
atsižvelgiant + į + NP (Acc.)
and others.

It seems that in the final version of translation many prepositional *pagal* phrases have been replaced by the competing constructions, even though there is nothing wrong with the grammar of *pagal* phrases. It could be the case that editors of the final version were more likely to replace the prepositional phrases, in order to get further from the organisation of source texts. The lack of finding more frequent regularities prevents making any strong claims of this sort.

5.2.3 *Analysis of the verb* "turi" *(*has/have*)*

In bilingual dictionaries, *turi* is traditionally translated as "to have" (see for example in Piesarskas & Svecevičius 1991). As the type *turi* denotes the verb in the third person singular or plural, the core meaning in English can be translated either as *has* or *have*. Besides possessive meaning the verb *turi* just like English *have* can express the authoritative requirement. Analysis of such a verb is particularly interesting in the legal documents as the main function of a legal document is to set rules and requirements. Discrepancies of frequencies in

different phases of translation, thus show that there is no general agreement among translators as to how to translate these important words in the legal discourse.

As it is shown in Figure 5, the frequency of *turi (has/have)* is gradually decreasing: falls by 34 in D2 as compared with D1 and by as many as 56 in F3 as compared to D2. As in previous analyses, first of all, we have examined parallel concordances, in order to find the commonest patterns associated with the verb *turi (has/have)*. These patterns are presented in Table 13.

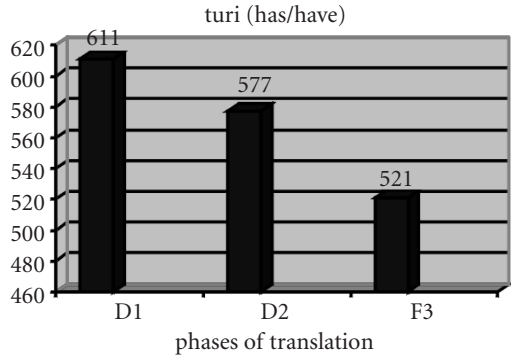

Figure 5. Changing of frequencies of the type "*turi*" (*has/have*)

Table 13. The common patterns of the type "*turi*" (*has/have*) in D1, D2, and F3

Pattern	**D1**	**D2**	**F3**
turi būti	*358*	*350*	*317*
(shall (be) + V, should (be), are (to be), must (be) etc.)	*59%*	*61%*	*61%*
įstaiga / įstaigos / įstaigai / įstaigoms turi	*22*	*22*	*22*
(body shall (be))	*4%*	*4%*	*4%*
turi pateikti	*21*	*17*	*17*
(shall report / forward / give / supply)	*3%*	*3%*	*3%*
turi atitikti	*13*	*11*	*12*
(shall comply / must meet /	*2%*	*2%*	*2%*
turi teisę	*12*	*12*	*10*
(shall be authorized, entitled, being eligible, have the right)	*2%*	*2%*	*2%*
turi imtis	*10*	*3*	*4*
(shall take)	*2%*	*1%*	*1%*
Other	*175*	*162*	*139*
	29%	*28%*	*27%*
TOTAL	*611*	*577*	*521*

The table suggests that the verb *turi* has a strong association with the verb *būti (to be)*, as their pattern *turi būti* accounts for 60% of the usage of *turi.* In contrast, the other leading patterns are very infrequent. Drawing on this evidence we might claim that except for the co-occurrence with *būti*, the verb *turi* is collocationally unrestricted. Its right collocates, however, are restricted grammatically, as most of the patterns as well as *turi būti* fall into the grammatical framework ***turi* + *infinitive*** equivalent to English ***have* + *to infinitive*, *must* + *infinitive*, *ought* + *infinitive*** (Piesarskas & Sveceviˇcius 1991), the frameworks, which commonly imply a command, requirement, obligation, or duty to do something.

We can also infer from the table above that the overall fall of frequency of *turi* in D2 (–34) is not so much related to the fall of the collocation *turi būti* (–8), but it is also scattered amongst the other patterns (other = –13), while the fall in F3 as compared to D2 is primarily associated with the fall of *turi būti* (–33), it also has a significant fall of 23 in all the other patterns.

It is also obvious that the fall of frequency of *būti* in F3 is correlated with the disappearance of the collocation *turi būti* (see Figure 6 below).

Investigation of concordance lines for the three versions has shown that the English patterns ***shall be* + *verb (Past participle)*** and ***shall* + *verb (Infinitive)*** have been the most problematic cases. They are most commonly translated as *turi būti* in all three versions, however, the versions very often disagree about their translations, as is shown in the example below.

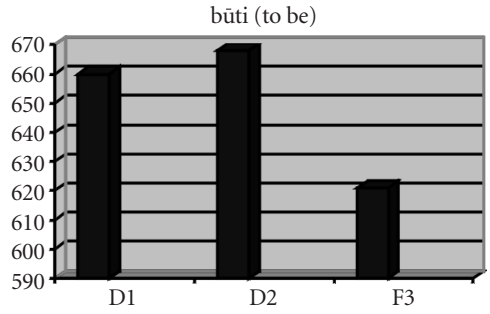

Figure 6. Changing of frequencies of the type "*būti*" (*to be*)

<EN> ... company **shall be regarded** as indirectly holding voting rights...
<D1> ...bendrovė **turi būti traktuojama** kaip netiesiogiai turinti balso teisių ...
<D2> ...bendrovė **turi būti traktuojama** kaip netiesiogiai turinti balso teisių ...
<F3> ...bendrovė **traktuojama** kaip netiesiogiai turinti balso teisių ...

<EN> ... company **shall be regarded** as having been effected by ...
<D1> ... bendrovė **privalo būti traktuojama** kaip ...
<D2> ... bendrovė **privalo būti traktuojama** kaip ...
<F3> ... bendrovė **traktuojama** kaip ...

The confusion of the translators might have been caused by the fact that the modal verb *shall* has essentially three different functions. One of its functions is that of an auxiliary verb, the second function is to express various polite intentions, and the third function that is the most common in EC law documents is to communicate a strong requirement by a rule or law, as in "*The Security Council shall decide what measures shall be taken to restore peace and security*".

Yet another pattern that has been inconsistently translated during different phases of translation is the modal expression ***should be,*** especially between the versions D2 and F3.

<EN> ...these harmonized standards **should be established** by common agreement by ...
<D1> ...šie harmonizuoti standartai **turi būti patvirtinti** bendru susitarimu visų šalių narių...
<D2> ...šie suderinti standartai **turi būti patvirtinti** bendru susitarimu visų valstybių narių...
<F3> ...šiuos suderintuosius standartus **reikėtų tvirtinti** bendru susitarimu visų valstybių ...

While there is some variety in translation of identical English phrases, *turi būti* has been often interchangeable with *privalo būti, privalu, yra + V, reikėtų, turi* or has been simply removed or added in the editing phases. These expressions are close in meaning, however, they do convey different levels of strictness. For example while *turi būti* and *privalo būti* imply strong requirement, *reikėtų (should)* implies desire that something is done rather than required.

The general tendency has been observed, that earlier versions are more likely to keep to an original "more wordy" structure of translation, while later versions, especially the final version, tend to synthesize the English verbal phrase into the more compact Lithuanian expression.

Thus frequency fluctuations during different phases of translation of the verbs *turi (has/have)* and *būti (to be)* have pointed to difficulties in translating the English modal expressions to Lithuanian. The variety of different choices of translating English modals might mean that translators do not follow any explicitly set rules for translation of particular expressions, and just follow their intuition of language.

6. Conclusions

The underlying thought of this article is that a corpus-based approach should facilitate translation studies. We attempted to demonstrate how innovative methodology and expertise of corpus linguistics can contribute to the development of process-oriented translation studies.

It has been shown that a relatively small, but carefully designed corpus of successive translated versions might be a rich source of issues for corpus-driven and process-oriented research into translating, provided that a combination of qualitative and quantitative methods is applied.

The study has focused on three different areas, which are the following: justification for the corpus-based and process-oriented approach, methodology of compilation of the PT corpus, and the actual analysis of the corpus.

The quantitative analysis of the corpus has presented the methodology, which has allowed capturing the problematic word types of legal discourse by assessing frequency fluctuations of word types for the different phases of translation. First of all, we have dealt with missing types from one version as compared to another. The analysis has led to the discovery of systematic replacements of entire groups of words, which leads to a number of general observations: for instance, we believe that the analysis of missing types allows us to spot difficult terminological problems such as simplification, latinization, and translationese in translated texts. Then we have concentrated our attention to the top of frequency lists of the different subcorpora. The comparison of the lists has allowed us to recognize word types that have considerable frequency fluctuations across different versions of translation.

The qualitative analysis has answered some questions in connection with frequency fluctuation of the problematic word types. The analysis of parallel concordances of these types has shown that the most frequent content words are commonly employed for very functional purposes in legal discourse, such

as, for example, cross reference. We have also reported a number of cases where the influence of original texts over the target language has taken place.

We hope that linguists will more readily use phases of translation corpora of a similar design in the future, as they provide exciting opportunities for analysis of the language of translation.

In a broader perspective, comparing successive stages in a translation has similarities to other successive emergence of text, such as updating of documentation (for example software manuals) or text-critical editions of texts. Perhaps the presented methodology of the phases of translation corpus could be adapted for other texts not just translation.

Notes

1. This kind of data could only relatively be considered as "real time", since although the versions represent some gradual emergence of a translated text, they are fixed products of some earlier mental activity. The approach does not represent the continuing real-time process, as do on-line approaches. Thus it belongs to the off-line rather than on-line approaches and the data can hardly account for any continuing mental processes. The approach can only be claimed to represent the translated language in certain phases of translation.

2. CELEX (Communitatis Europaeae Lex) is a computerized system produced by the Office for Official Publications of the European Communities (EUR-OP) accessible on-line at http://europa.eu.int/celex.

3. "Vanilla Aligner" is written in C, it runs on UNIX or MS-DOS platforms. "Vanilla Aligner" for MS-DOS can be freely downloaded at:
http://spraakbanken.gu.se/lb/English/downloads.phtml

4. See http://www.ruf.rice.edu/~barlow/parac.html

5. Translations of Lithuanian words taken from *Lithuanian-English Dictionary* (Piesarskas & Svecevičius 1991) are not exhaustive, as they are just meant to serve as aids for non-Lithuanian speakers.

6. It should be noted though that a percentage is very sensitive to the size of word frequency, and thus it cannot be used as a decisive factor for evaluating differences. While it is a useful factor for contrasting frequent types that are neighbours on a frequency list, it is not a reliable measure in relation to the whole frequency list.

References

Barlow, M. (1995). *A Guide to ParaConc*. Houston: Athelstan.

Danielsson, P., & Ridings, D. (1997). Practical Presentation of a *Vanilla* aligner. In *TELRI Newsletter No 5*. Manheim: Institute für Deutsche Sprache.

Bernardini, S. (1999). Using Think-Aloud Protocols to Investigate the Translation Process: Methodological Aspects. In N. J. Williams (Ed.), *RCEAL: Working papers in English and applied linguistics 6* (pp. 179–199). Cambridge: University of Cambridge.

Gale, W. A., & Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics, 19* (1), 75–102.

Gellerstam, M. (1986). Translationese in Swedish Novels Translated from English. In L. Wollin & H. Lindquist (Eds.), *Proceedings of the Scandinavian Symposium on Translation Theory (SSOTT) II* (pp. 88–95). Malmö: Liber förlag.

Grefenstette, G., & Tapanainen, P. (1994). What is a Word, What is a Sentence? Problems of Tokenization. In *Proceedings of the 3rd International Conference on Computational Lexicography (COMPLEX'94)* (pp. 79–87). Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences.

Halliday, M. A. K. (1966). Lexis as a linguistic level. In C.E. Bazell et al. (Eds.), *In Memory of J. R. Firth* (pp. 148–162). London: Longman.

Hansen, G. (1999). *Probing the process in translation: Methods and results*. Copenhagen: Samfundslitteratur.

Jakobsen, A. L. (1998). Understanding the Process of Translation: The Contribution of Time-Delay Studies. In B. E. Dimitrova (Ed.), *Översättning och tolkning – Rapport från ASLA.s höstsymposium Stockholm, 5–6 november 1998* (pp. 155–172). Stockholm: ASLA.s Skriftserie 12.

Jakobsen, A. L., & Schou, L. (1999). Translog Documentation. In G. Hansen (Ed.), *Probing the process in translation: methods and results* (pp. 151–186). Copenhagen: Samfundslitteratur.

Laviosa-Braithwaite, S. (1998). Universals of Translation. In M. Baker (Ed.), *Routledge Encyclopedia of Translation Studies* (pp. 288–291). London: Routledge.

Piesarskas, B., & Svecevičius B. (1991). *Lithuanian-English Dictionary.* Vilnius: Mokslas.

Scott, M. (1996). *Wordsmith Tools Manual.* Oxford: Oxford University Press.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.

Tognini-Bonelli, E. (1996). *The Role of Corpus Evidence in Linguistic Theory and Description*. Unpublished PhD thesis. Birmingham: University of Birmingham.

Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins Publishing Company.

Ulrych, M. (1997). The Impact of Multilingual Parallel Concordancing on Translation. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC'97: International Conference on Practical Applications in Language Corpora* (pp. 421–436). Lodz: Lodz University Press.

Utka, A. (forthcoming). Lemmatisation and Collocational Analysis of Lithuanian Nouns. In G. Barnbrook, P. Danielsson & M. Mahlberg (Eds.), *Meaningful Texts: the Extraction of Semantic Information from Monolingual and Multilingual Corpora* (pp. 105–112).

Woolls, D. (1997). *MULTICONC: software for multilingual parallel concordancing*. Birmingham: CFL Software.