Keeping Czech in check: A corpus-based study of generalization in translation[1]

Jana Kubáčková

Generalization and specification of lexical meaning are studied, both quantitatively and qualitatively, as potential universals of translation on the basis of a modest English-Czech corpus comprising monolingual, multilingual and parallel subcorpora. Against the backdrop of recent research in this area, generalization and specification are outlined from the viewpoint of semantics, lexicology, stylistics and contrastive language typology, with a particular focus on the category of translation universals and the employment of corpus methodology. Methods and tools are tailored to the needs of the analysis and the two contrasting concepts are operationalized. The results obtained confirmed a weak overall tendency to generalization as well as occurrence of specification which may be explained as due to the influence of several factors.

Introduction: Arrival of corpora and universals

Translation universals have been around for quite some time now. They acquired substantial prominence in the mid-1990s when Mona Baker outlined the potential of corpus linguistics in the study of

> [...] universal features of translation, that is features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems (Baker 1993:243).

The introduction of the methods of corpus linguistics into descriptive translation studies (DTS) was hailed as "a turning point in the history of the discipline" (Baker 1993: 235). Indeed, in conjunction with corpus methodology, the study of translation universals has yielded some very interesting, indeed impressive results in terms of the nature of translations as well as research methodology (e.g. in Meta 43:4, Chesterman 2004). Yet I cannot avoid the impression that - for all that has been written about them - translation universals remain very elusive. After all, as Pym (2008) remarks, the degree of overlap among the various translation universals (those studied so far) is such that we might as well classify them as sub-categories of Toury's law of growing standardization, or interference. While this does not precisely bring us "back to square one", Pym's brilliant argumentation exposes the patchiness of our knowledge of translation universals and the need to understand translation universals in mutual relations.

And despite the methodological breakthrough made possible by large electronic corpora, such tools often remain difficult to harness. What Kenny wrote in 1998 still holds today:

[1] This article is based on an MA thesis defended at Charles University (Prague) in 2008.

> As has been shown time and time again in corpus linguistics, a new resource can give impetus to new research. The challenge is to know what questions to ask of a translation-oriented corpus, and how to ask them (Kenny 1998: 523).

In addition to the tricky nature of electronic tools and (largely) quantitative results, there is also the problem of linguistic systems that are more or less different from English. This has far-reaching implications for the use of tools originally tailored for English and for attempts to adapt them to, let's say, a Slavonic language, which turned out to be a problem in the present study when it came to dealing with the translators' tendency to generalize lexical meaning.

Universals, or tendencies?

The starting point, and in many respects the cornerstone of the study is the theory developed by Jiří Levý, the Czech translation scholar, who found that:

> Experiments with translators have shown that, when offered a group of near-synonyms, they exhibit a natural tendency to select from it the most generalised term, the least specific word (Levý 2008: 52).

In his theoretical work dating back to the 1950s (e.g. Levý 1955),[2] Levý addresses various general phenomena that have an impact on translating and translation, e.g. the tendencies towards generalization, stylistic levelling or 'intellectualization' - the latter concept represents a kind of rationalization and partly overlaps with what is today referred to as the explicitation universal).

In his approach to the process and product of translating, Levý proceeds from a prescriptive hypothesis to description, based on his experiments and observations, and towards explanation (e.g. in Levý 1971 a,b; Levý 2008: 47f.) and points to objective, but also subjective, i.e. psychological, cognitive, and pragmatic factors that may influence the outcome of the translation process. The three key factors pointed out are (a) the structure of the translators' linguistic memory and (b) their perception of their role as mediators between the text and the reader, but also (c) the principle of least effort, the "minimax theory":

> […] the translator selects from the range of alternatives the one which promises the maximum effect for the minimum effort (Levý 2008: 62-63).

The translator not only takes into account the reader's most likely expectations[3] but, more importantly, adopts a pragmatic approach to the process of translating, seeking (consciously or subconsciously) to strike a balance between his or her own efforts and potential results, looking for

[2] Other key studies and monographs include Umění překladu (The Art of Translation, first published in 1963) and studies in the anthology published in 1971 (for a recent translation into English see Králová - Jettmarová et al. 2008).

[3] Similarly to Gutt's theory of relevance; Gutt (2000).

> a sentence structure which broadly takes account of all the essential semantic and stylistic features, although a more perfect version might be found following a protracted period of experimentation and thought (Levý 2008: 63).[4]

However, opting for the solution that is readily available in the linguistic memory may easily result in a translation that is "colourless, general and vague" (Levý 2008: 52). According to Levý, good translators go

> deeper than the first, second or third level of the lexicon, selecting, as far as possible, words which contain all the semantic attributes of the source text (ibid: 52).

In his attempts to explain phenomena characterizing the process of translation and in his inherently interdisciplinary approach (Levý 1971a: 148), he in many respects anticipates the "most recent" trends in DTS and CTS. Sadly, having lived on the "wrong" side of the Iron Curtain, he is still too much of an outsider in the English-speaking world of translation studies.

In translation research, the position of generalization and specification is a rather marginal one. Lexical generalization is sometimes understood as a feature of simplification (Blum-Kulka – Levenston 1983, in Halverson 2003: 219; Klaudy 2003), while specification is often seen as an aspect of explicitation (Leuwen-Zwart 1990: 90; Klaudy 1993, in Baker et al. 1998; Øverås 1998). Leuwen-Zwart (1990: 93) and Munday (1998) suggest that specification is more prevalent than generalization, thus questioning the universal character of the latter.

In their classical *Stylistique comparée du français et de l'anglais: méthode de traduction* (1958; English translation in 1995), Vinay and Darbelnet use the term "generalization" to label a translation technique – not a universal tendency, but a conscious strategy "in which a specific (or concrete) term is translated by a more general (or abstract) term" (Vinay-Darbelnet 1995: 343).

In a similar vein, the concept of generalization (and, correspondingly, specification - referred to as "concretization") was developed by Kinga Klaudy (1996, 2003), who classifies these phenomena as (a) language-specific, (b) culture-specific and (c) translation-specific (Klaudy 1996). However, by taking examples from widely different languages (Finno-Ugric vs. Indo-European) and building on their well-known linguistic and stylistic preferences, she cannot account for generalization or specification as translation universals, even though she does include this category and describe it in terms similar to those used by Levý (2008: 62-63):

> [...] translators might be tempted to follow the line of least resistance, and if they cannot find a precise equivalent in the TL, they will select a word with a more general meaning [...]. (Klaudy 2003: 9)[5]

However, most studies addressing potential universals such as normalization, simplification, sanitization, Toury's law of growing standardization etc., subsume the tendency to use vague,

[4] In this, he is close to Pym (2008) and his notion of risk avoidance.

[5] The page numbers in references to works by Klaudy refer only to pages of texts printed from email attachments, not books. The texts have been kindly provided by Kinga Klaudy herself.

less specific vocabulary under the respective universal. For example, Laviosa, who focuses on simplification and has made a significant contribution to the use of electronic corpora in translation research, classifies the tendency to overuse high-frequency words[6] (the "core patterns of lexical use") as one of the criteria defining simplification (e.g. Laviosa 2002: 58n, 2003: 158-9).

This brings out the issue of genetic inter-relatedness between groups of potential translation universals, as highlighted by Pym (2008). Halverson (2003: 218n) suggests that behind these cognate tendencies there is a common cause, a sort of "gravitational pull" exercised by the most salient members of the semantic structure. Interestingly enough, in her article grounded in cognitive science, she endorses the arguments of Levý who also speaks of a "symptom of attraction exercised [...] by the best-known member of a group of synonyms" (1983: 143, my translation JK).

This highlights the need to study potential translation universals in their mutual relationships, horizontal as well as vertical. To use a rather crude example, the Czech language has a predilection for semantically rich verbs. A lack of specific verbs introducing direct speech in fiction translated from English into Czech may be seen as language-pair-specific generalization, which therefore cannot be considered as a universal. However, at a higher level of abstraction, it may be regarded as a feature of the *unique items hypothesis* proposed by Tirkkonen-Condit (2004) and simultaneously an instance of negative interference according to Toury (1995). Cases like these imply that objective conditions (linguistic and stylistic norms etc.) and subjective ones (the translator's linguistic memory, experience etc.) tend to combine and we can only make a more or less precise guess as to which of the two is more probable.[7]

Languages in contrast: the preliminaries

Any analysis of generalization involves the treatment of the issue of lexical meaning and synonymy both in general and language-specific contrastive aspects. After all, it is actually the existence of synonyms, differentiated by shades of notional, pragmatic or contextual meaning that provides the paradigm from which the translator can choose:

> A paradigm cannot of course be considered a set of equivalent elements but a set ordered according to a variety of criteria (e.g. 'shades of meaning', 'stylistic levels' etc.), as otherwise no choice would be possible. (Levý 2008: 51)

As regards stylistic levels, some formal aspects of synonyms can be distinguished as useful for subsequent corpus analysis, e.g. Czech expressive synonyms can often be identified thanks to certain suffixes and combinations of letters.

In Czech functional stylistics, the concept of synonymy is broader and more loosely defined than in lexicology (Filipec 1961: 145, Bečka 1948: 63) and is therefore more convenient for semantic analysis in translation. Levý (2008: 49-52), conceiving translation as

[6] High frequency of occurrence is assumed, already by Levý, to accompany a vaguer semantic content.

[7] There are of course many more factors to be taken into account, such as the translator's attitude to stylistic norms of the source and target cultures, the author's style etc.

a decision-making process, speaks of near-synonyms; J.V. Bečka, a leading Czech stylist of Levý's era, concedes:

> In stylistics, it is not only the word as such, but the choice between words that is at stake, [and sometimes] we have to decide between words whose meanings are close but by no means overlapping, i.e. between words that do not constitute true synonyms. (Bečka 1948: 63; my translation JK).

In other words, stylistic synonymy is very rich but rather unstable, context- and function-sensitive, sometimes verging on (co-)hyponymy. The reason is that in translation and any text analysis we deal with *parole,* which represents a projection of the paradigmatic axis onto the syntagmatic axis. Such a broad delimitation allows for the conception of synonymic chains where the dominant member, the "centre of gravity", tends to be the most frequent and general one (Filipec 1961: 205); this brings us back to Levý and his arguments explaining translators' tendency to generalize.

Contrastive language typology is another crucial aspect of the preliminary analysis, accounting for the principal differences between the source (English) and target (Czech) languages – it is all the more important as such typological differences have far-reaching implications for corpus research methodology. As Pym (2008) pointed out, comparable corpora represent an attempt to get rid of the influence of the source language, but in themselves are insufficient since they cannot account for interference and thus can lead to erroneous conclusions. This is one of the stumbling blocks of comparable corpora as conceived by Baker (1995). True, to a certain extent, the influence of linguistic systems can be harnessed using Jantunen's method of three comparable subcorpora (Jantunen 2004). In parallel corpora, where the source-target relations can be observed more directly, the first step is to isolate systemic differences in order to identify their influence. The next step in the present study was therefore the establishment of relevant typological features and stylistic preferences of Czech and English with special focus on the vocabulary and methodological implications.

The Czech scholar of English language and literature Vilém Mathesius draws a parallel between language typology and the meaning of lexical units.

> Roughly speaking, words in a language with a synthetic structure (such as Czech) usually have a more definite meaning than words in a language with an analytical structure (such as English or French) (Mathesius 1975: 18).

English is also classified among languages characterized by a high degree of polysemy (Čermák 2004: 205). While synthetic languages, including Czech, usually use affixes to create new words, English can often simply convert nouns to verbs etc., without changing the form. In addition, English vocabulary, known for its tendency towards monosyllabism, includes a significant proportion of homonyms (Vachek 1974:66).

This may have a significant impact on corpus research. For example, one cannot directly compare the frequencies and counts of word types in a parallel Czech-English corpus – one English type most probably stands for a number of context-dependent meanings, and may represent several different parts of speech. Nor can the type/token ratio be used to account for vocabulary richness in both languages – due to inflection, one Czech word can

occur in many cases with different endings, thus substantially increasing the number of types in relation to tokens. Thus the resulting ratio would be much lower than that for English.

Another crucial aspect of typological differences with direct consequences for corpus research methodology is the concept of "the word" itself.

> The definition of the word varies from one language to another. For example in Czech and other Slavonic languages of a predominantly synthetic type, the boundaries between words as opposed to collocations, sentences and morphemes are drawn more clearly than in English, a predominantly analytical language. (Filipec - Čermák 1985: 34, my translation JK).8

As Mathesius rightly points out, "there are borderline cases; besides independent words there are words approaching affixes" (Mathesius 1975: 24) – English, in particular, often uses apostrophes and hyphens, which can divide words as well as members of a compound. Czech and English differ also in their approach to and usage of various types of compounds.

In his guide to the ParaConc corpus manager, Michael Barlow pays special attention to the category of the word:

> [...] the first definition of a word that comes to mind is a string of letters (and perhaps numbers) surrounded by spaces. And with a little further thought, we would realise that we need to include punctuation symbols, in addition to spaces, as possible delimiters of words. Hence, we can define a word as a string of characters bounded by either spaces or punctuation (plus special computer characters such as the carriage return) (Barlow 2003: 75).

ParaConc, for example, treats the apostrophe as a part of the word. However, by changing search options, the apostrophe may be classified as a word delimiter. Similar precautions apply for using WordSmith Tools.

Compound words are even more problematic, mainly due to the varying degree of independence of hyphenated words. Moreover, a corpus manager cannot be expected to capture all instances of compounds since "it is largely a matter of personal choice whether we write match box, match-box or matchbox (Stubbs 2002: 31). Needless to say, phrasal verbs such as *give up, care for* etc., which usually have one-word Czech equivalents, are "invisible" for corpus managers. Finally, English uses many grammatical words (articles, auxiliary verbs etc.) and expressions where the grammatical and semantic functions are distributed between the members (*to have a swim, to give a laugh,* etc.). The whole – which is more than the sum of the parts – is unrecognizable in a frequency list.

Useful information concerning systemic differences between Czech and English can be gained from translated texts - as shown for example by Knittlová (2003), building on examples from 1960s-1980s translations. Like Baker (1992), Knittlová addresses various types of non-equivalence and speaks of generalization and specification as sub-categories of partial equivalence. She considers specification to be the prevalent tendency in translations from English into Czech and highlights the semantic richness of "multifaceted" Czech verbs:[9]

---

[8] Filipec and Čermák refer to the article by Josef Vachek (1961) – Some Less Familiar Aspects of the Analytical Trend of English. In Brno Studies in English 3, 9-78.

[9] Here, to some extent, Knittlová cannot avoid the blurring of purely semantic and grammatical (or semi-grammatical) categories such as the aktionsart. The present study excludes consideration of Czech verbal aspect and English tenses.

> Again, this is related to the typological difference between the two languages, to the nominal character of English and the rather verbal character of Czech (Knittlová 2003: 34, my translation JK).

Knittlová adds that "Czech equivalents of the most frequent groups of English verbs are semantically richer and more specific" (Knittlová 2003: 51, my translation JK).[10] She (2003: 51-52) also suggests that although English has verbs of similar specificity they are used much less frequently.

Linguistic typology also influences the way languages use markers of expressiveness:

> In English, expressiveness tends to be concentrated in lexical units which carry solely expressive connotational features and have a capacity to radiate, while in Czech texts expressiveness is spread more evenly over a greater number of units that carry both denotational and connotational features (Knittlová 2003: 106, my translation JK).[11]

As for generalization, most examples in Knittlová are illustrative of cultural differences rather than of a phenomenon occurring during the process of translation.

To be sure, a good translator ought to be able to come to terms with the incommensurable nature of language pairs – and the first step is to be aware of the problem and the remedy. As Levý (1983: 70) points out, routine Czech translations from English make insufficient use of diminutives and other means of expressing affection due to the typological differences. However, in the complex decision-making process of translation in general, and translation of fiction in particular, the issue of generalization vs. specification is only one of many.

Hypotheses and operationalization

Generalization in translation is defined as a "conscious or subconscious semantic loss of one or more specific semes (notional or pragmatic), in a lexical unit [...] in contrast to corresponding units of the original or, quantitatively, to comparable original texts written in the same language" (Kubáčková 2008: 65). Conversely, specification is defined as the opposite tendency, i.e. as "conscious or subconscious semantic enrichment [...]".

Bearing in mind that the boundaries between possible causes of these phenomena can be blurred and using the categories presented by Levý (1983) and Klaudy (1996), instances of generalization/specification are classified according to the following potentially independent variables: (a) differences between language systems, (b) stylistic norms, (c) pragmatic factors such as cultural knowledge and (d) as translation-inherent (universal tendencies, lack of time or experience, unwillingness to look for a better solution, etc.).

In line with Levý, who considered the vague and "grey" style to result from the tendency of translators to choose a more general word (1983: 137), the following set of hypotheses was established:

[10] These include the category of verbs introducing direct speech (*verba dicendi*) where the semantic richness of Czech verbs is reflected in the prevalent stylistic norm requiring lexical variation.
[11] E.g. certain endings and suffixes typical for spoken Czech.

I The vocabulary of a corpus of Czech texts translated from English will be more general and deprived of semantic colour in comparison with original Czech texts.
II The vocabulary of a corpus of Czech texts translated from several languages will be more general and deprived of semantic colour in comparison with original Czech texts.
III In translations, generalization will be more frequent than specification if we exclude instances of obligatory specification, non-equivalence due to differences in language typology and instances that can be accounted for by stylistic conventions.
IV The tendency to generalize will be observed in different translations of an identical original.

In terms of observable phenomena, the following were considered as indirect indicators of generalization:

1) A lower number of appellative autosemantic lemmas and a lower lemma-token ratio will be found in the translation corpus compared to the comparable corpus of original texts;
2) The first 200; 500; 1000 appellative autosemantic lemmas respectively in the frequency list of the translation corpus will cover a higher percentage of the corpus than the same numbers of lemmas in the comparable corpus of original texts;
3) The first 200 appellative autosemantic types in the frequency list of the translation corpus will cover a higher percentage of the corpus and include fewer lemmas than the same number of types in the comparable corpus of original texts;
4) The number of lemmas with a frequency of 1 up to 10 will account for a smaller part of the total number of lemmas in translations than in the comparable corpus of original texts;
5) The number of specific expressive lemmas produced by lexical derivation[12] (also in comparison to the total number of lemmas) will be lower than in the comparable corpus of original texts;
6) The range of synonyms and near-synonyms will be less varied than in the comparable corpus of original texts.

The criterion for directly observed indication of generalization was the following:

7) When compared to the original, a given passage of a translation will display more instances of semantic loss (generalization) than of semantic specification. These instances will not be directly relatable to typological differences between the languages in question or the influence of target-language stylistic conventions. (Kubáčková 2008: 66-68)

[12] Derivation is the typical procedure for word formation in Czech. Therefore, expressive endings adding stylistic colour to the text do not readily suggest themselves to the translator from English; their absence may result in the generalization of lexical meaning, which may endorse the Unique Items Hypothesis.

Analytical methods and procedures

The analysis was carried out on three levels, with three different types of corpora, starting with largely quantitative observations and gradually increasing the proportion of qualitative research. The selection of texts was guided by the principle of mainstream fiction, since this was the type of material on which Levý had based his theory. At the same time, fiction, due to its aesthetic function, could be expected to reveal instances of noticeable semantic loss or enrichment. The aim was not only to verify Levý's experimental data by using electronic tools, but also to apply corpus-based methods on Czech texts and so contribute to their refinement.

The first analytical level handled a monolingual comparable corpus (in terms of Laviosa 1997a: 292) consisting of three subcorpora of Czech fiction extracted from the SYN2005 corpus, which is part of the freely accessible Czech National Corpus (CNC)[13] and 40% of which comprises fiction texts. The CNC corpus manager Bonito was used to design the subcorpora in line with the criteria of Jantunen's three-phase comparative analysis (Jantunen 2004: 106f) to provide for the control of the influence of English as the source language. By spotlighting interference, this method also helps uncover phenomena that are not the result of the influence of the source language.

The building of the subcorpora had to tackle an imbalance in the book market also encountered by Bernardini and Zanettin (2004) – most of the texts were translations from English, Czech originals ranged second and translations from other languages came last.[14] As the aim was to create the largest subcorpora possible, the smallest subcorpus had to be taken as the benchmark and consequently the size of the other two subcorpora had to be adjusted so as to make them comparable. Three subcorpora were obtained with a total size of some 22 million tokens:

ORIG:	7 201 905 (i.e. Czech original fiction)
T-Engl:	7 207 238 (i.e. translations from English)
T-mix:	7 209 242 (i.e. translations from a mix of languages)

The selection criterion of "contemporary mainstream fiction" being rather vague, the subcorpora were composed of texts published in the period 1960-2004, with the majority published in the 1990s and later. T-mix consists of 37 translations from Germanic languages, 37 from Romance languages, 26 from Slavonic languages and 12 from non-Indo-European languages (Finnish, Japanese, Hebrew and Yiddish), which gives quite a balanced mix.[15]

Comparability of the subcorpora is based on the criteria of their size, genre (prose), period of publication and language. As the T-mix subcorpus was limited by the availability of texts in the CNC[16] the criteria could not have been further fine-tuned.

[13] Accessible at http://ucnk.ff.cuni.cz/

[14] It is worth pointing out that the sources for SYN2005 were selected on the basis of a wide-scale readership survey. See http://ucnk.ff.cuni.cz/.

[15] Cf. Laviosa (2002: 63) who mentions the disadvantage of having a large proportion of source languages from one group. The wide choice offered by SYN2005 is probably due to the Czech translation tradition.

[16] All available texts were used. For their list see Kubáčková (2008).

The Bonito manager made it possible to use lemmatization and tagging provided in the SYN2005 texts.[17] After retrieving all the tokens of each subcorpus (query .*.), the negative filter (N-filter) was used to eliminate all lemmas starting with a capital letter (proper names) and all punctuation marks, numbers and numerals and synsemantic parts of speech including pronouns. Thus, allowing for tagging errors, sets of nouns, adjectives, verbs and adverbs[18] were obtained and frequency lists of lemmas produced for the calculation of the lemma/token ratio for each subcorpus (Fig.1).

FIG. 1

	ORIG	T-Engl	T-mix	difference ORIG - T-Engl	difference ORIG - T-mix
No. of tokens (size of the subcorpora)	7201905	7207238	7209242		
No. of appellative autosemantic tokens	3483594	3431286	3473394		
No. of appellative autosemantic lemmas	95145	72256	68873	22889	26272
lemma/token ratio (%)	2,7312	2,1058	1,9829	0,6254	0,7483

The difference between the respective lemma/token ratios is in percentage points.

Since in Czech each lemma of an inflected word occurs as a number of types, there is a significant disparity between the number of types and different lemmas. Therefore to capture lexical diversity the lemma/token ratio had to be used instead of the usual type/token ratio.

The comparative study of the generalization indicators No. 1– 6 was based on the frequencies[19] of lemmas and affixes.

Interestingly enough, although originally smaller than the translation subcorpora, ORIG turned out to contain the highest number of appellative autosemantic lemmas. The differences between the lemma/token ratios of ORIG and T-Engl/T-mix respectively were not statistically significant, but, together with most of the other indicators (% covered by the most frequent lemmas/types, the numbers of low-frequency lemmas etc. as is evident for example in Fig. 2-5), indicated that there was a difference between ORIG on the one hand and translation subcorpora on the other, suggesting a greater lexical diversity in ORIG.

[17] The risk of error must be allowed for, despite the fact that the methods for lemmatization and tagging of SYN2005 represent a major step forward as compared to preceding corpora. For more information see http://ucnk.ff.cuni.cz/ .

[18] The boundaries between different parts of speech are not always clear-cut; The present approach is tailored to the tools of electronic analysis.

[19] [...] in a field like translation, the best, if not the only way to go about estimating "probabilities for terms in [...] systems" is to proceed from "observed frequencies in [a] corpus" (Toury 2004: 20).

FIG. 2

		No. of the most frequent lemmas (list head)					
	size of corpora (No. of tokens)	The first 200		The first 500		The first 1000	
Subcorpus	(appellative, autosemantic)	sum	%	sum	%	sum	%
ORIG	3483594	1261775	36,221	1656373	47,548	1981174	56,872
T-Engl	3431286	1326524	38,660	1732302	50,486	2066143	60,215
T-mix	3473394	1291135	37,172	1708494	49,188	2055030	59,165

FIG. 3

Subcorpus	Size	No. of types (list head)	No. of lemmas (n)	Part of the subcorpus covered by the first 200 types (p)
ORIG	3483594	200	137	780734 (22,412 %)
T-Engl	3431286	200	139	816267 (23,789 %)
T-mix	3473394	200	141	789657 (22,734 %)

FIG. 4

Subcorpus	Average frequency of the first 200 types (f = p/200)	Average frequency of the first n lemmas (f = p/n)
ORIG	3903,670	5698,788
T-Engl	4081,335	5872,424
T-mix	3948,285	5600,404

FIG. 5

	ORIG	T-Engl	T-mix
Total No. of lemmas	95145	72256	68873
No. of lemmas with a frequency ≤ 10	72298	51801	48511
%	75,99	71,69	70,44

As for the usage of expressive affixes in original texts and translations (see e.g. Fig. 6), the results were quite convincingly in favour of ORIG, suggesting that translators tend to neglect the specific potential of Czech morphology. On the other hand there were hardly any differences in the usage of synonyms. From the point of view of methodology, it may be more worthwhile to focus on affixes as parts of words bearing only limited semantic information than on words as such, e.g. synonyms, the occurrence of which appears to depend much more on the texts in the corpora.

FIG. 6 Expressive suffix *–isko* (augmentative semantic value)

Subcorpus	ORIG	T-Engl	T-mix
No. of expressive lemmas	28	7	9
Without proper names	25	7	9
Total No. of lemmas	95145	72256	68873
Of which expressive lemmas (%)	0,0263	0,00969	0,0131

The three-phase comparable analysis also indicated certain instances of interference from English, but these were negligible against the backdrop of the overall tendency of translations to use less varied vocabulary.

Admittedly, the differences between originals and translations were usually small. In addition, we must allow for a number of limitations, such as the size and composition of the corpora, lemmatization errors etc. However, the results repeatedly pointed to a less varied vocabulary in both types of translation subcorpora.

The second level of analysis aimed at testing the third hypothesis – i.e. the prevalence of generalization in translations with the exclusion of instances caused by systemic or stylistic differences. It was based on a parallel corpus of five books of fiction and their translations into Czech (Kubáčková 2008: 74). The originals were all published after 1950 and the translations after 1989;[20] the books were written by well-known authors and can be considered mainstream fiction; the authors include both men and women from Great Britain, the USA and Canada; each of the books was translated by a different person with Czech as their mother tongue; each author and translator is represented only once.

The corpus was analysed with WordSmith and ParaConc. Three reference corpora were used in addition – the British National Corpus (BNC), the frequency lists of the American National Corpus (ANC), and a CNC reference corpus of original Czech fiction (over 10 million tokens) extracted by the author of the present study.

To get a rough picture of the lexical variety of the English originals, their standardized type/token ration per 1000 words was calculated in WordSmith and the results were compared to the standardized type/token ration of original English fiction in BNC 1995 – a benchmark used by Zanettin (2000: 111). The English part of the corpus as a whole was only slightly above the reference value of 44.44 (also calculated by WordSmith) and the values for individual novels showed no extreme deviations that would indicate a peculiar vocabulary usage.

In order to devise a method that would be as objective and as easy to replicate as possible, a ParaConc frequency list of the original texts was produced first. Since the words in the list head are likely to be translated into Czech in a more specific way due to systemic language differences, the subsequent analysis focused on infrequent types:[21] 100 types were selected which occurred only once in the list and less than 100 times in the BNC or the ANC. Their meanings were checked in dictionaries in order to select semantically rich words. The process of selection was carried out prior to the analysis of the translations so as to not to distort the results by any subjective bias.

Subsequently the translations were analyzed in ParaConc and word pairs then examined in the minimum context necessary. Not surprisingly, numerous English expressions were "spread" over several units in translation, which would be unobservable in a purely quantitative study solely relying on electronic analytical data.

Shifts in translation, based on Popovič's typology (1974: 122f; 130) and lexical stylistics, were identified with reference to a variety of dictionaries (monolingual, bilingual,

[20] The year 1989 is considered a landmark which brought a major change into the social and economic context of Czech translation.

[21] These subcorpora were not lemmatized.

synonymic, etymological). There being no occurrences of generalization caused by pragmatic differences between the readers of the originals and the translations in the 100 words chosen, occurrences of generalization and specification were classified as (a) systemic (language-specific), (b) stylistic and (c) translational. Three more categories were needed to account for the remaining cases: other types of shifts, zero equivalents (omission) and zero or negligible shifts.[22] The analysis of the 100 lexical units and their translations yielded a prevalence of translational generalization:

FIG. 7

100 units	systemic	stylistic	translational	sum		
generalization	11	0	26	37	other shifts	zero / negligible shifts
specification	2	1	7	10	9	44

In addition, shifts were observed within the context sentences – i.e. in other lexical units. Here the occurrence of stylistic specification increased and prevailed over generalization. However, after elimination of the systemic and stylistic types of specification, translational generalization prevailed over specification:

FIG. 8

shifts in context sentences	systemic	stylistic	translational	sum	other shifts
generalization	1	0	22	23	3
specification	5	9	10	24	

The results suggest a significant tendency towards generalization and, with respect to the material analysed, confirm the third hypothesis. At the same time they contradict Leuven-Zwart (1990) and Munday (1998), who found a prevalence of specification. However, it is possible that their material displayed a significant degree of systemic or stylistic specification which was not treated separately from translational phenomena.

However, no shift, be it generalization, specification, or even a zero shift, should be a priori qualified as negative, undesirable, or positive (Popovič 1974: 131). Generalization may deprive the translation of some colour (such as in *to marshal other ranks – odvést, i.e. to "lead away"*, Kubáčková 2008: 102), but specification can also have a negative effect by offering an almost ready-made interpretation. Zero or negligible shifts may include both well-fitting solutions as well as ill-fitting expressions. There are also instances where such a shift is deliberate because appropriate from the aspect of a larger context, or may be introduced by the editor. Such information is inaccessible, but such conditioning has to be accounted for as a possible factor.

The third level represents a deep analysis of two translations of the novel *Foundation and Empire* by Isaac Asimov (1952). Two translations of one original offer a unique possibility to

[22] The classification of shifts into these necessarily rough categories is far from unambiguous since, as pointed out above, there are no clear-cut boundaries between the causes underlying each choice made by the translator. For complete results, see Kubáčková 2008.

focus on a limited number of variables – the personality of the translators, their idiolect, experience and preferences, and the context in which the translations were produced.

Although the first Czech translation of Asimov dates back to 1970,[23] it was the years after 1989 that witnessed an outbreak of publishing frenzy. Between 1991 and 2006, the newly emerging small publishing houses, driven very probably by demand from readers, churned out at least two new Asimov translations almost every year - inevitably, with a negative impact on the quality of the new translations. Translator Richard Podaný (2000) mentions the Czech version of the novel *Foundation* by Jarmila Pravcová (1991) as one of the books that inspired the establishment of translation anti-awards. In the same year, *Foundation and Empire* was published, again in Pravcová's translation. The publishing house, AG Kult, became notorious for negligent editorial work. Podaný does not expressly speak about *Foundation and Empire*, but the context, as a starting point of analysis, certainly does not bode well. The second translation of *Foundation and Empire* was probably a reaction to this "rush" period. Published in 2003, the new Czech version was produced by Viktor Janiš, a young translator who has established a good reputation.

The analysis of the original confirmed rich vocabulary (with a type/token ratio per 1000 words of 45.6 compared to the above-mentioned benchmark of 44.44, Zanettin 2000: 111). Next, WordSmith was used to search for keywords, i.e. the words identified as typical for a single text (here regarded as a small corpus) in contrast with a larger corpus. The Keywords tool was employed to compare the vocabulary of *Foundation and Empire* to the other four English-language novels used in the previous analyses.[24]

Disregarding the names of characters and words related to the content of the novel (*planet, galaxy*), the list of keywords featured many expressions related to speech: first and second person pronouns, most common present tense verbs including their short forms and words that could introduce or describe direct speech (*shrugged, smiled, spoke, nodded, replied, frowned, muttered, whispered* etc; adverbs *dryly, coldly, harshly, sombrely*; and also the interjection *huh* and nouns such as *voice* or *speech*). As these words had been chosen for their relatively high frequency, it could be inferred that the novel built heavily on the dialogue or direct speech and its varieties. There were complex structures qualifying speech (*with a crisp air of finality, with slow meaning,* etc.), an unusually varied usage of *verba dicendi*, a high degree of expressiveness in the speech of certain characters, as well as a significant range of synonyms describing communication (*to prate, to jabber, to babble* etc.). These features were established as dominant for the style of the original and therefore as the focus of subsequent translation analysis.

In *verba dicendi*, the difference found between the two translations was striking. Again, WordSmith's Keywords were used, this time in a rather unusual way. Keywords are normally employed to compare a text with a large reference corpus; however, identification of keywords in two translations of one text may be a promising launch pad as both use certain content-related expressions which will thus not appear in the keyword list. By definition, the list will yield words that are "overused" by one of the translators, thus pointing to their idiolect, approach, etc.

[23] Information provided by the online catalogue of the Czech National Library.

[24] A maximum of 1000 words was searched for. The minimum "key" frequency was set at 3, with p=0,001.

The list revealed a pronounced disparity in the usage of verba dicendi.[25] The Czech for *[he] said – řekl* – tops the list of keywords in Pravcová and the corresponding lemma occurs 334 times in her translation.[26] On the other hand, Janiš takes great pains to avoid what he considers to be the obvious interference, and uses the lemma a mere twelve times. However, in his translation, other verba dicendi are conspicuously frequent – they are often rather formal or even bookish (*opáčit, odtušit* - similar, but not quite synonymous with *retort* or *riposte*). Moreover, some of them seem to be overused in translations in general (as detected in T-Engl and T-mix corpora), in contrast with original Czech texts.

Thus, while Pravcová features a high rate of interference of the English stylistic norm, making the Czech dialogues rather stereotypical, Janiš takes care to respect domestic conventions, but doesn't always keep his own lexical predilections under control. By overusing certain semantically rich verbs, he draws attention to them without need or purpose. Besides, some expressive verbs occur in collocations where they do not fit.

Further analysis of *verba dicendi* and longer stretches of discourse revealed that Janiš's effort to use more varied and colourful vocabulary was also reflected in the higher degree of expressiveness in dialogues. His method is certainly in line with the overall tendency of the original, and a great improvement on the previous translation which substantially deprived the dialogue of its original colour. In places, however, Janiš, carried away by this tendency, disregards the context.

The two translations show diverging tendencies – one a tendency towards generality and stylistic interference, and the other an inclination to (over)use of colourful and semantically specific vocabulary with occasional losses due to the intention to be "different". Thus, generalization occurs in both translations, but cannot be said to be equally prevalent. Social conditions and the policy of the publisher can influence this trend while the idiolect and approach adopted by the translator can go against a "general" tendency, as shown in Janiš's effort to counter stylistic interference.

Conclusion

Generalization is observed as a weak but universal tendency of translated texts in monolingual comparable corpora. In pairs of originals and translations, it was prevalent in particular in semantically complex lexical units, as observed at the second-level analysis; however, as seen in the comparison of two translations of one novel, generalization seems to be largely dependent on the translator's idiolect and ambitions as well as on the social context.

The concept of a "universal tendency" follows the line of thought expressed for example by Toury (2004), but what does it mean in practice? On the basis of the findings of the present study, the following can be proposed as a tentative hypothesis: the universal nature of generalization can be expected to reveal itself when large amounts of data are compared by quantitative methods. However, the more one relies on qualitative analysis and the closer one comes to the individual translator and his/her idiolect, experience and working conditions in a particular social context, the more can one expect discrepancies in findings. In

[25] The list of keywords was useful as it indicated the main tendencies. However, WordSmith seems to have difficulty with treating texts in Czech. It was therefore necessary to verify and correct the data.

[26] The original uses only a few more – 358 instances of *said* as introducing speech.

other words, although generalization may perhaps never be ruled out, it can be overridden by contrary tendencies.[27] Different types of corpora can thus be expected to yield different types of results – small samples point to the differences between translator individualities while larger corpora reveal the general pattern.

Finally, the results obtained suggest that with a flexible approach, the enormous potential of corpus managers may be exploited in large corpora as well as in individual text analyses. In the present article, particular attention is paid to the factors that must be taken into account in a corpus-based analysis of two largely different languages. Contrastive semantics, lexicology, stylistics and morpho-semantic typology provide the groundwork in which the research methods are anchored. Care has been taken to present the necessary methodological adjustments and adaptations needed to meet the challenge of studying a synthetic language with tools designed for a language with an analytical structure. Electronic corpora offer potential for innovation, as e.g. our keyword analysis suggests.

References

BAKER, Mona. 1992. *In Other Words: A Coursebook on Translation*. London and New York: Routledge.

BAKER, Mona. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In Baker, M., Francis, G., Tognini-Bonelli, E. (eds), *Text and Technology. In Honour of John Sinclair*. Amsterdam: J. Benjamins, 233-250.

BAKER, Mona. 1995. Corpora in Translation Studies. An Overview and Some Suggestions for Future Research. *Target* 7: 2, 223-243.

BARLOW, Michael. 2003. *ParaConc: A Concordancer for Parallel Texts*. (Draft 3/03). [online] [cit. 1-9-2008]. Accessible at http://www.athel.com/paraconc.pdf

BEČKA, Josef V.1948. Úvod do české stylistiky. Praha: Knižnice Kruhu přátel českého jazyka.

BERNARDINI, Silvia – Zanettin, Frederico. 2004. When Is a Universal Not a Universal? In MAURANEN, A. - Kujamäki, P. (eds), *Translation Universals. Do They Exist?* Amsterdam: J. Benjamins, 51-62.

BLUM-KULKA, S. – Levenston, E. 1983. Universals of lexical simplification. In Færch, C. –Kasper, G. (eds.), *Strategies in Interlanguage Communication*. London: Longman, 119-139.

CHESTERMAN, Andrew. 2004. Beyond the Particular. In Mauranen, A. - Kujamäki, P. (eds), *Translation Universals. Do They Exist?* Amwsterdam: J. Benjamins, 33-49.

[27] Clearly, this is one of the challenges faced by translator training – to warn against undesirable tendencies, while avoiding simplifying judgements.

ČERMÁK, František. 2004. *Jazyk a jazykověda. Přehled a slovníky*. Praha: Karolinum.

ČERMÁK, František – Kocek, Jan. 2008. *Co je korpus?* [online] [cit. 30-8-2008]. Accessible at http://ucnk.ff.cuni.cz/co_je_korpus.html.

FILIPEC, Josef. 1961. *Česká synonyma z hlediska stylistiky a lexikologie*. Praha: Nakladatelství Československé akademie věd.

FILIPEC, Josef – Čermák, František. 1985. *Česká lexikologie*. Praha: Academia.

HAJIČ, Jan. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech),* Vol. 1. Praha: Karolinum.

HAJIČ, Jan – Krbec, Pavel – Květoň, Pavel – Spoustová, Drahomíra – Votrubec, Jan. 2007. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. ACL 2007, Prague*. Praha: Karolinum,. 67-74.

HALVERSON, Sandra. 1998. Translation Studies and Representative Corpora: Establishing Links between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study. *Meta* 43: 4, 494-514.

HALVERSON, Sandra. 2003. The Cognitive Basis of Translation Universals. *Target* 15: 2, 197-241.

JANTUNEN, Jarmo Harri. 2004. Untypical Patterns in Translation. In Mauranen, A. - Kujamäki, P. (eds), *Translation Universals. Do They Exist?* Amsterdam: J. Benjamins, 101-124.

KENNY, Dorothy. 1998. Creatures of Habit? What Translators Usually Do with Words. *Meta* 43: 4, 515-523.

KLAUDY, Kinga. 1996. Concretization and Generalization of Meaning in Translation. In Thelen, M., Lewandowska-Tomaszczyk, B. (eds), *Translation and Meaning, Part 3*. Maastricht: Hogeschool Maastricht, 140-152.

KLAUDY, Kinga. 2003. *Languages in Translation. Lectures on the Theory, Teaching and Practice of Translation*. Budapest: Scholastica, 321-327.

KNITTLOVÁ, Dagmar. 2003. *K teorii i praxi překladu*. Olomouc: Univerzita palackého v Olomouci.

KRÁLOVÁ, Jana – Jettmarová, Zuzana et al. 2008. *Tradition versus Modernity. From the Classic Period of the Prague School to Translation Studies at the Beginning of the 21st Century*. Praha: Karlova univerzita – TOGGA.

KUBÁČKOVÁ, Jana. 2008. *Generalizace a specifikace lexikálního významu v překladu*. [MA Thesis]. Praha: Karlova Univerzita v Praze.

LAVIOSA, Sara. 1997a. How Comparable Can "Comparable Corpora" Be? *Target* 9: 2, 289-319.

LAVIOSA, Sara. 1997b. Investigating Simplification in an English Comparable Corpus of Newspaper Articles. In Klaudy, K. – Kohn, J. (eds), *Transferre necesse est*. Budapest: Scholastica, 531-540.

LAVIOSA, Sara. 2002. *Corpus-Based Translation Studies. Theory, Findings, Applications*. Amsterdam: Rodopi.

LAVIOSA, Sara. 2003. Corpus and simplification in translation. In S. Pertilli (ed.), *Translation Translation*. Amsterdam: Rodopi, 153 - 162.

LEVÝ, Jiří. 1955. Překladatelský proces – jeho objektivní podmínky a psychologie. *Slovo a slovesnost* 16, 65-87.

LEVÝ, Jiří. 1971a. Bude teorie překladu užitečná překladatelům? In *Bude literární věda exaktní vědou?* Praha: Československý spisovatel, 147-157.

LEVÝ, Jiří. 1971b. Geneze a recepce literárního díla. In *Bude literární věda exaktní vědou?* Praha: Československý spisovatel, 71-143.

LEVÝ, Jiří. 1983. *Umění překladu*. Praha: Panorama.

LEUWEN-ZWART, Kitty M., van. 1989. Translation and Original. Similarities and Dissimilarities I, II. *Target* 1: 2, 151-183; 2: 1, 69-95.

MATHESIUS, Vilém. 1975. *Obsahový rozbor současné angličtiny na základě obecně lingvistickém*. Praha: Nakladatelství Československé akademie věd.

MUNDAY, Jeremy. 1998. A Computer-Assisted Approach to the Analysis of Translation Shifts. *Meta* 43: 4, 542-556. [online] [cit. 2008-08-14]. Accessible at http://id.erudit.org/iderudit/003680ar

ØVERÅS, Linn. 1998. In Search of the Third Code: An Investigation of Norms in Literary Translation. *Meta* 43: 4, 571-588.

PODANÝ, Richard. 2000. Koniášovská retrospektiva. In *Interkom,* 9-10-2000. [online] [cit. 2008-08-24]. Accessible at http://www.scifi.cz/ik/2000/20000908.htm

POPOVIČ, Anton. 1974. *Teória umeleckého prekladu*. Bratislava: Tatran.

PYM, Anthony. 2007. *On Toury's laws of how translators translate*. [online] [cit. 2008-07-05]. Accessible at http://www.tinet.org/~apym/

PYM, Anthony – Shlesinger, Miriam – Simeoni, Daniel. (eds). 2008. *Beyond Descriptive Translation Studies. Investigations in homage to Gideon Toury*. Amsterdam: J. Benjamins.

SCOTT, Mike. 1998. *WordSmith Tools Manual. Version 3.0*. Oxford: Mike Scott & Oxford University Press.

STUBBS, Michael. 2002. *Words and Phrases. Corpus Studies of Lexical Semantics*. London: Blackwell.

TIRKKONEN-CONDIT, Sonja. 2004. Unique items – over- or under-represented in translated language? In Mauranen, A. - Kujamäki, P. (eds), *Translation Universals. Do They Exist?* Amsterdam: J. Benjamins, 177-184.

TOURY, Gideon. 1995. *Descriptive Translation Studies and Beyond*. Amsterdam: J. Benjamins.

TOURY, Gideon. 2004. Probabilistic Explanations in Translation Studies. In Mauranen, A. and Kujamäki, P., *Translation Universals. Do They Exist?* Amsterdam: J. Benjamins, 15-32.

VACHEK, Josef. 1974. *Chapters from Modern English Lexicology and Stylistics*. Praha: Státní pedagogické nakladatelství.

VINAY, Jean-Paul; Darbelnet, Jean 1995. *Comparative Stylistics of French and English: a Methodology for Translation*. Transl. and ed. by Juan C. Sager and M. J. Hamel. Amsterdam: J. Benjamins.

ZANETTIN, Federico. 2000. Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis. In Olohan, M. (ed.), *Intercultural Faultiness*. Manchester: St. Jerome, 105-118.

Jana Kubáčková

In *SKASE Journal of Translation and Interpretation* [online]. 2009, vol. 4, no. 1 [cit. 2009-09-07]. Available on web page <http://www.skase.sk/Volumes/JTI4/pdf_doc/04.pdf>. ISSN 1336-7811.