Izvorni znanstveni članak

Cultural and Scientific Transfer through Translation – a Corpus-Based Study of Term Formation in the Period 1848-1919

Špela Vintar
Department of Translation, Faculty of Arts
Aškerčeva 2, SI – 1000 Ljubljana
spela.vintar@guest.arnes.si

Summary

Contacts between cultures are a driving force of technological, scientific and linguistic development, where a culturally or economically more advanced region "feeds" its neighbouring regions. The Austro-Hungarian Empire was a multi-cultural environment where this transfer can be observed through – among other processes – translation. The study focuses on the development of Slovene technical and scientific terminology under the influence of German and is based upon a sample from a recently built digital library of translations from German into Slovene from the period 1848-1919, which was created within two nationally-funded projects in Austria and Slovenia.

The period between 1848 and 1919 was particularly crucial for the Slovene linguistic development, and German was at the time an influential language from which many concepts, expressions, phrases and names were borrowed. Apart from the level of the lexicon, significant changes can also be observed on other linguistic levels, from orthography, morphology, phonology, to syntax and pragmatics.

This study analyses several term formation processes such as borrowings, loan translations, term variation etc.

The corpus-based method allows us quantify the observed linguistic phenomena and compare some generic traits of texts from various domains, such as the type/token ratio or the ratio of borrowings compared to native vocabulary.

Key words: Term formation, Digital library, Corpus-based study

1. Introduction

Much of the world's knowledge is already available in electronic format, however the same is not true for cultural heritage, of which texts in various formats are a significant part. From the point of view of intercultural contacts, texts are also powerful vehicles of cultural influence and thus also carriers of technical and scientific progress.

The field of digital libraries and new access modes to scientific and cultural heritage is developing rapidly on the EU level, with several large initiatives under way, e.g. **DELOS**[1] **“Network of Excellence on Digital Libraries”**, which conducts a joint program of activities aimed at integrating and coordinating the ongoing research activities of the major European teams working in Digital Library-related areas with the goal of developing the next generation Digital Library technologies, **ERPANET (Electronic Resource Preservation and Network)** (IST-2001-32706) which is establishing an expandable European Consortium, which will make viable and visible information, best practice and skills development in the area of digital preservation of cultural heritage and scientific objects. Both of these are also continuing the work of IST SM **DigiCULT**[2] **“Digital Culture”** (2002-2004), which established a regular technology watch for cultural and scientific heritage.

The aim of our study is to show that cultural and technical developments go hand in hand with linguistic development, whereby translation is in effect the bridge between the “feeder” and the “receiver” cultures. The Austro-Hungarian monarchy was extremely diverse in terms of national and linguistic structure, yet the processes of cultural transfer between Austria and Slovenia in the 19th century were by no means always fluent. Especially as far as literary production was concerned, the general opinion among the Slovene literary elite was that only original production was worthwhile and valuable, while translations – especially from German – would contaminate the language and jeopardize the Slovene cultural identity (Hladnik 1992, Orel 2005). This attitude was not completely mirrored in translations of technical and scientific texts from German into Slovene, particularly in the domains that were considered ideologically unproblematic (agriculture, housekeeping, medicine, natural sciences etc.).

As an attempt to create a digital resource that would enable both quantitative and qualitative research of phenomena related to translation, two parallel national projects were launched, namely: the Austrian Government funded **2004-2006** project “**German-Slovene/Croatian translation 1848-1918 (FWF P17465)**”, with Graz Academy of Sciences as the coordinator, and the Slovene Government funded 2004-2007 project “**Slovene Translations of German Texts in the period 1848-1919 – Cultural and Linguistic Impacts (J6-6078)**”, coordinated by University of Maribor (Teržan-Kopecky 2004). The result of the joint efforts of all project partners is the AHlib digital library consisting of the following parts:

[1] http://www.delos.info/

[2] http://www.digicult.info/

- TraDok – a comprehensive bibliography and database of Slovene, Croatian and other translations from German from the period 1848-1919, with their German counterparts, containing over 6,000 bibliographical units and equipped with a multi-function search interface[3],
- digitised and processed texts constituting the AHlib digital library, where each text has undergone scanning, OCR, manual correction, semi-automatic linguistic annotation (part-of-speech tagging and lemmatization), analysis of historical wordforms, and finally conversion into TEI (cf. Erjavec/Ogrin 2005).

Although part-of-speech tagging and lemmatization of contemporary Slovene are procedures for which reliable tools have been developed, linguistic annotation of archaic texts is an entirely different story. Many words are not recognized either because they are no longer in use or because they were spellt differently. Furthermore, orthographic variation was a common phenomenon both with general language words (eg. rujavo, rjavo, rjujavo [brown]) and specialized terms (operacija, operacia [operation]). The number of lemma types compared to word form types in a text can help us estimate variation in a text (see Figure x). Lemmatization of the AHlib digital library is being performed as an iterative process where each round of automatic lemmatization is followed by a manual correction phase. After each manual correction of unknown lemmata the lexicon for automatic annotation is expanded and the results consequently better.

Final versions of the above processing steps (in PDF, RTF and XML formats) are being uploaded into TraDok, while an online interface for the above conversion is available at the Jozef Stefan Institute[4].

Since the processing of the texts – especially the time-consuming manual correction phase – is still underway, the study presented in this paper is based upon a small sample from the target digital library, namely on 7 technical or scientific books from different domains and time periods. Another small corpus of literary works from the same period was compiled for the purposes of comparison. The listing of all books used can be found below.

2. Methods of corpus-based analysis

For the quantitative and qualitative analysis of the historical text collection we are using Wordsmith Tools (Scott 1998), a suite of programs and utilities for statistical text processing and concordancing.

As a measure of lexical density, which can give insight into the size of vocabulary used, the standardised type/token ratio (std. TTR) is used. This measure compares the number of different words (types) found in a corpus or text to the number of running words (tokens). The ratio between the two typically de-

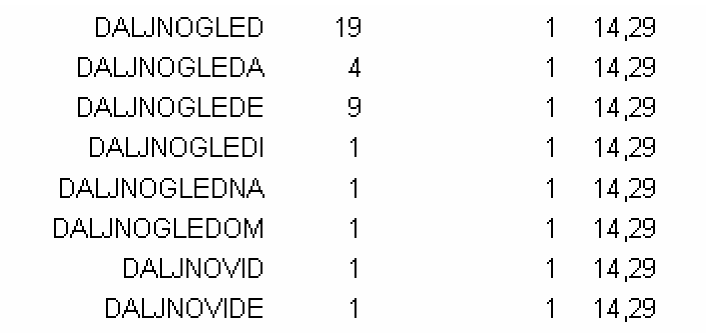
[3] https://buedo22.uni-graz.at/pub/tradok/

[4] http://nl.ijs.si/ahlib

creases with corpus size, therefore comparison between different texts is only possible if samples of equal size are drawn from each text (standardised TTR). In all below experiments we use sample size of 5,000 words.

As an indicator of the complexity of vocabulary we can observe word length, more specifically the amount of extremely long words in a text – usually these words are found to be terms. An indicator of syntactical complexity can be average sentence length, although this may be also related to the text type.

By constructing word lists ordered alphabetically, by frequency or backwords we can analyse orthographic shifts, term variation and term formation from a single root. The example below shows a terminological nest around **daljnogled** (binocular), with the term variant **daljnovid**.



DALJNOGLED	19	1	14,29
DALJNOGLEDA	4	1	14,29
DALJNOGLEDE	9	1	14,29
DALJNOGLEDI	1	1	14,29
DALJNOGLEDNA	1	1	14,29
DALJNOGLEDOM	1	1	14,29
DALJNOVID	1	1	14,29
DALJNOVIDE	1	1	14,29

Comparing different word lists with the Keywords function helps us identify words that occur in a specific text with a higher relative freqnency than in a larger reference corpus. These words most often represent the terminological or technical inventory of the text, however they may include other domain- or author-specific words.

3. Text analysis

3.1 General corpus characteristics

As described above, two subcorpora were compiled for the purposes of this study, one consisting of 7 technical or scientific works spanning the years 1847-1908 (Tech), and the other containing 5 literary works from the same period (Lit). All books were translated into Slovene from German. The sizes of both subcorpora and some general statistics are given in Table 1.

	Tech	Lit
Size in bytes	1,251,569	737,815
Tokens (running words)	212,814	128,669
Types (different words)	25,759	19,385
St. type/token ratio	30,73	37,28
Av. sentence length	16	17
Av. word length	5,2	4,7

Table 1: General corpus characteristics

There appears to be a significant difference in TTR, with the corpus of literary texts exhibiting a higher lexical density than the corpus of technical texts. This could be an indicator of the fact that literary texts indeed use a richer vocabulary than technical texts, and indeed one must acknowledge that literary texts were normally translated by language professionals, either writers or dedicated translators. In contrast, technical texts were often translated by domain experts or clerical people.

Another variable that may influence TTR is the amount of orthographic variation. If the same word is spellt in several different ways, each form is counted as a separate word type. We do not however believe that this had a major impact on the observed difference between the corpora.

The slightly higher average word length in the Tech corpus is upon nearer examination the result of term usage.

3.2 Lexical density over time

In the rest of the study we are focusing mainly on lexical properties of non-literary texts in our sample. An intuitive hypothesis was that older texts would exhibit lower TTRs than newer ones because the term inventory of a language expands over time. There is however no empirical evidence of that, on the contrary – there is a notable tendency of decreasing TTR with texts produced at a later date (see Figure 1). Figure 2 shows the TTRs of literary texts for comparison.

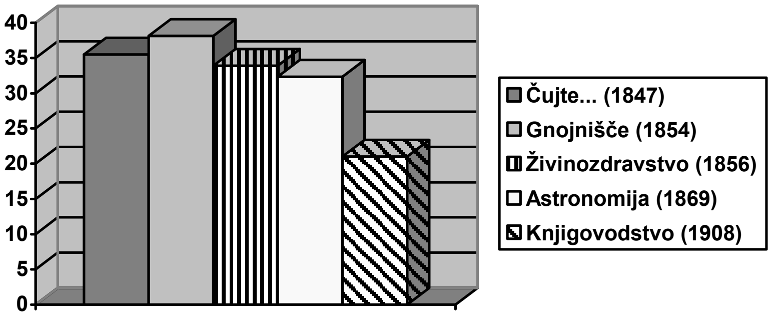


Figure 1: Type-token ratios of technical texts

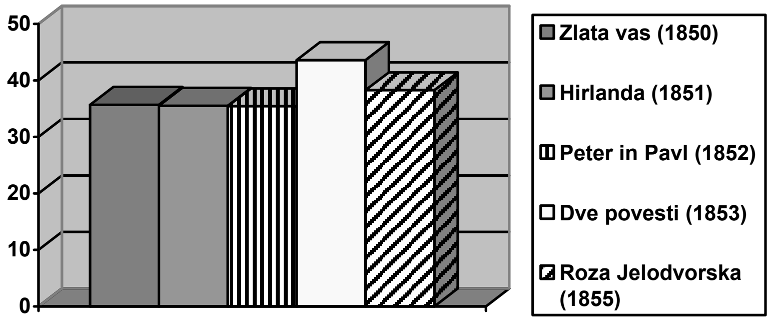


Figure 2: Type-token ratios of literary texts

3.3 Foreign citation as a sign of term pre-formation

A common practice still at work today is to introduce a new term in language accompanied by an original term in brackets. The citation of the original term serves as reference for the readers familiar with the source language, and on the other hand eases the translators dilemma in introducing a new expression. There are numerous examples of this practice in our text collection:

> *13. Pasja kuga ali mor* ***(Staupe)*** *207 14. Poglavje. Ploščnatna glista* ***(Bandwurm)*** *208 15. Poglavje. Gliste v želodcu in črevah 209 (Živinozdravstvo)*
>
> *Tako n. pr. zaznamavamo narazje nebeskih teles z zvezdno daljavo, z zemeljskimi poloméri; zemeljsko površje merimo z miljo, s protom* ***(Ruthe)****, s sežnjem, z metrom, in reči manjše raztege s čevljem, s palcem in črto. (Astronomija)*
>
> *Tudi ne sme konj imeti kraka* ***(Spath)****, ne podplatnih otisk* ***(Steingallen)****, ne nadkosti na zadnji strani skoknega člena* ***(Hasenhacke)****, ne pipe* ***(Piphacke)****, ne lupine* ***(Schale)****. Obedva jajca se morata dobro viditi. (Živinozdravstvo)*

It is clear from the above examples that the citations are used as a reference to another concept system from which the translated work is drawing knowledge, but we can also see the translators’ efforts of using and using Slovene terminology whenever possible. Although citations occasionally reveal instability in term usage (see example below from Živinozdravstvo, where *Bandwurm* is translated first as *ploščnata glista* and second as *trakulja*) and are generally taken to be signs of an extremely passive attitude to term formation, the analysed works show quite the opposite.

...14. Poglavje. Ploščnatna glista ***(Bandwurm)****....*
....ampak le gosenica trakulje ***(Bandwurm).*** *.....*

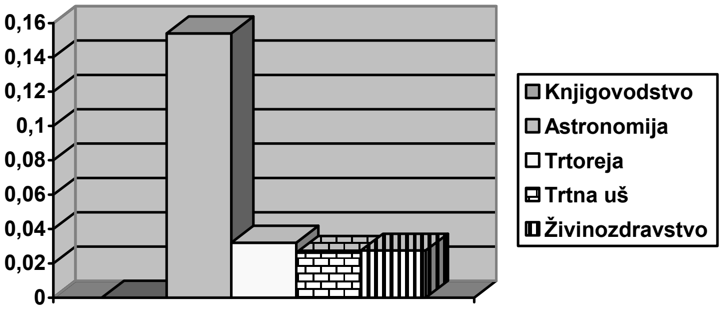


Figure 3: Percentage of German cited wordforms

Figure 3 shows the ratio of German cited words compared to the number of all types.

3.4 Term variation

As mentioned above, there is considerable variation in spelling in our text collection, and this naturally applies to term spelling too. Several variations are systematic and indicate an overall orthographic shift in Slovene, such as *serce - srce, smert – smrt; rudeč – rdeč, rujav – rjav; kteri – kateri, kedar – kadar*. Other variations typically occur with borrowed words (from German or classical languages) at phoneme boundary I and A: *salmiak – salmijak – salmjak, operacija – operacia – operacja*.

Cases of terminological inconsistency where several expressions are used to refer to the same concept are difficult to detect automatically, moreover we do not regard them as variations but as alternative representations of the same concept (e.g. *polutnik – ravnik – aequator – ekvator*; all synonyms from Astronomija). Yet another set of variations occur on the level of grammar, such as the ending –i in the locative case: *na obrežji* – [contemporary Slovene] *na obrežju*.

Since the process of lemmatization usually reduces these variants and assigns them a common lemma, we thought it would be interesting to compare the ratios of lemma types and word form types. The lemma/wordform ratio is thus the number of different lemmata found in a text divided by the number of different wordforms (see Figure 4).

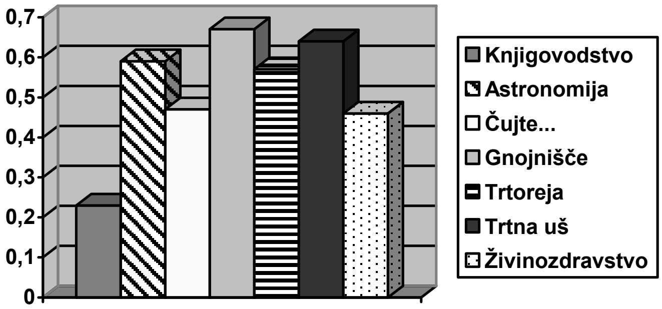


Figure 4: Lemma/wordform ratios

The underlying hypothesis is that texts with a higher lemma/wordform ratio actually contain more lexical diversity and less orthographic variation. Perhaps Knjigovodstvo should be excluded from this comparison because it contains many numbers that had been counted as different wordforms but all have the same lemma. In other texts the hypothesis might be confirmed, especially if we see that the lowest lemma/wordform ratios (apart from Knjigovodstvo) are the ones of the oldest books in the sample.

3.5 Loan translations

A highly productive term formation strategy is direct or loan translation, meaning that the original term is transferred into the target language by translating the original bits (Cabré 1998: 94). There are numerous examples of loan translations in our text collection, especially from Astronomija:

Brennpunkt – gorišče, ognjišče
Hundstage – pasji dnevi
Kegelschnitt – kegeljosečnica
Nachtbogen – ponočni lok

Upon detailed examination of the texts however we find that a more frequent strategy of translators was not to translate directly but to create genuinely Slovene expressions according to word formation principles. Whether neologisms were at that time recieved as reluctantly as today, we can only speculate, but as the general linguistic climate was very much against borrowings, especially from German, we believe that Slovene neologisms were definitely more welcome than cited foreign words.

4. Conclusions

The study explored some aspects of term formation in Slovene in the period 1848-1919. Some intuitive hypotheses have not been confirmed by corpus evidence, for example that the type-token ratio would grow with time as a reflection of the expanding vocabulary. It seems that the vocabulary expansion that had certainly taken place through technological and scientific development was on the other hand balanced by more orthographic standardization and less variation. A comparison of technical and literary texts showed that lexical density was slightly higher in the latter.

Several volumes we explored showed a high percentage of cited German expressions. These were however rarely used as borrowings without any attempt of introducing a Slovene term, rather they were used as additional reference points for the informed reader who may not yet be familiar with Slovene terminology. On the whole, translation was in the selected time period still regarded as a creative process of restructuring and adapting the text to the target audience, while the principle of equivalence only gained importance from 20th century onwards.

A more extensive and detailed study of the above phenomena will be possible when the AHlib digital library is finished, and several studies using AHlib materials are already underway. However, to gain a thorough insight into linguistic and cultural processes at work within the chosed time frame, one should have available a comparable digital library of Slovene original production as well as German originals of the analysed works.

To conclude in a visionary tone, digital libraries of the future will offer entirely new access modes to the knowledge and cultural achievements of past generations, and we may look forward to the applications still to be developed in this field.

References

Cabre, T. (1998). Terminology: Theory, methods and applications. Amsterdam/Philadelphia, PA:John Benjamins.

Erjavec, Tomaž and Matija Ogrin (2005). Digitalisation of literary heritage using open standards. In Paul Cunningham, Miriam Cunningham (eds.). Innovation and knowledge economy: issues, applications, case studies, (Information and communication technologies and the knowledge economy). Amsterdam [etc.]: IOS Press, 2005, str. 999-1006.

Honzak-Jahić, Jasna (2002) O vplivu češkega in nemškega jezika na razvoj slovenskega strokovnega izrazja. In: Slavica comparativa, Prague: Euroslavica, 2002, pp. [33]-39.

Hladnik, Miran (1992) "Vloga prevoda v slovensko-nemški literarni tekmi", in: Hladnik, Miran / Pečaj-Rus, Darinka eds. XXVIII. seminar slovenskega jezika, literature in kulture. Zbornik predavanj. Ljubljana: Center za slovenscino kot drugi tuji jezik pri Oddelku za slovanske jezike in knjizevnosti Filozofske fakultete, 109-119 [http://www.ijs.si/lit/prevodsl.html-l2].

Orel, Irena (ed.) (2005) Razvoj slovenskega strokovnega jezika. Proceedings of an international symposium, Ljubljana, 17-19 November 2005, Ljubljana: Center za slovenščino.

Scott, Mike (1998) WordSmith Tools Version 3. Oxford University Press, Oxford, England.

Teržan-Kopecky, Karmen.(2004) Slovenski prevodi nemških besedil v obdobju 1848-1918 : jezikovni in kulturni vplivi : vloga za (so)financiranje raziskovalnega projekta za leto 2004 = Slovene translations of German texts in the period 1848-1918 : linguistic and cultural impact : (research proposal). Maribor: Pedagoška fakulteta, 2004.

Bibliography of the sample corpus

Gasteiner, Josip (1908) *Knjigovodstvo za dvorazredne trgovske šole* [übers. v. Volc Ivan]. Ljubljana: Slov. trgovsko društvo "Merkur".

Goethe, Arminij (1881) *Trtna uš. Poljudno razlaganje o tem kakšne lastnosti ima in kako živi ovi najnevarnejši sovražnik vinske trte in kaj moramo storiti v obrambo zoper ta mrčes*. V Gradci: Štajersko društvo za omikanje ljudstva.

Goethe, Hermann (1891) *Iz biologičnega trtorejskega poskušališča od Hermana Goethe-ja v Badenu pri Dunaju* [übers. v. C. kr. kmet. društvo v Gorici]. Gorica: Tisk. Paternolli.

N.N. (1852) *Peter in Pavl, ali Bóg ubózih sirót najboljši oče. Povest za otroke in mladénče, za odrašene, kakor tudi za starše in učitelje*. V Celovcu: Pri založniku Joanu Leonu, bukvarju.

Rohlwes, Johann Nikolaus (1856) *Domače živinozdravstvo v boleznih konj, govedja, ovac, prešičev, koz in psov, ali nauk, kako mora kmetovavec svojo živino rediti, ji streči, jo kermiti in ozdravljati* [übers. v. Robida Karl]. V Celovcu: Janez Leon.

Schödler, Friedrich Karl Ludwig (1869) "Astronomija" [übers. v. Ogrinec Vil.[jem]], in: *Knjiga prirode. I. del. Fizika, Astronomija in Kemija*. Band: 1, v Ljubljani: Matica Slovenska, 229-339.

Schmid, Christoph (^{2}1853) *Dve povesti iz pisem Kristofa Šmida. A. Golobčik. B. Kanarčik* [übers. v. P.[intar] A.[nton]]. V Ljubljani: Jožef Blaznik.

Schmid, Christoph (1855) *Roza Jelodvorska. Lepa podučivna in kratkočasna pripovest za mlade in odrašene ljudi* [übers. v. Pijelik Dragoslav]. V Ljubljani: Henrik Ničman.

Šmid, Krištof (1851) *Hirlanda bretanjska vojvodnja ali zmaga čednosti in nedolžnosti. Nauka polna povest za starost in mladost*. Ljubljana: J. Giontini.

Wimmer, Anton (1854) *Gnojnišče kmetovavca zlati rudnik* [übers. v. Bleiweis Janez]. Ljubljana: c.k. kmetijska družba.

Zschokke, Heinrich (1847) *Čujte, čujte, kaj žganje dela! Prigodba, žalostna ino vesela za Slovence* [übers. v. Globočnik Felicijan]. V Celovci: Joan Leon.

Zschokke, Heinrich (1848) *Zlata Vas. Podučna in kratkočasna povest* [übers. v. Malavašič Fran]. Ljubljana: Natisnil Jožef Blaznik.