JoSTrans
The Journal of Specialised Translation

www.jostrans.org · ISSN: 1740-367X

Vintar, Š. (2008). Corpora in translation: A Slovene perspective. *The Journal of Specialised Translation, 10*, 40-55. https://doi.org/10.26034/cm.jostrans.2008.660

Corpora in Translation: A Slovene Perspective

Špela Vintar, Department of Translation, University of Ljubljana

ABSTRACT

This paper reviews the use of corpora in translation practice and translator training, focusing on currently available monolingual and multilingual language resources for Slovene. The first part of the paper briefly outlines the state-of-the-art in corpus linguistics and then introduces publicly available corpora for Slovene, including general and special language corpora as well as several parallel corpora. The advantages and potential pitfalls of using the web as a corpus are also discussed. Part two presents some important considerations and guidelines for using corpora in both training situations and, more specifically, real-world translation projects. In many respects, corpora represent a richer source of information for a translator than dictionaries; on the other hand, a corpus user must know how to critically interpret the results obtained via a corpus query. Because corpora may not be readily available for many special domains and/or language pairs, procedures and tools for compiling one's own corpora are also described.

KEYWORDS

Corpora in translation, parallel corpora, Slovene corpora, corpus-based translation.

1. Introduction

A few decades ago, the stereotypical image of a translator would most likely be one of an overworked, slightly grey woman or balding man nailed to a desk under a heap of dictionaries and encyclopedias, leading a rather solitary life. Today, a more realistic picture of a translator at work would inevitably feature a computer with an internet browser minimised on the task bar and the heap of dictionaries similarly replaced by an array of desktop icons.

It is a fact, yet to be acknowledged by many practising translators and translation scholars, that the digital age brought about a revolution in the translation business much more profound than merely switching from paper to a computer screen. The abundance of electronic texts on the web, available in many languages of the world and often multilingual, is just one of the reasons that printed dictionaries – or electronic editions of these same reference works for that matter – can no longer be considered the primary source of translation-relevant information. Another reason, closely related to the main credo of corpus linguistics that the primary element of analysis in language is the *text*, is that translators extremely rarely translate words in isolation. Any reference work that presents words devoid of textual context is thus of limited value in a translation environment.

If in the early years of corpus linguistics electronic text collections were still considered a luxury for various reasons, among them the cost of computer storage and the complexity of processing large amounts of

textual data, in the past decade the situation has changed radically. Indeed, it now seems obsolete to even compile corpora; instead of fixed collections of texts we are entering an era of tools for dynamic corpus creation in accordance with specific and individual requirements.

Naturally, there are still arguments in favour of proper corpora versus *ad hoc* text collections created dynamically by trawling the Web. Reference corpora aim to be representative of the language in its entirety, and to achieve this they include portions of as many language varieties as possible, including dialects, sociolects, spoken material, etc. The criteria for determining which texts to include and what proportion of the entire corpus a certain type of text should represent are carefully designed in order to obtain a resource where the frequencies of lexical items will correspond to overall language use. Therefore, the difference between this portrait of the language as a whole, although no reference corpus could ever claim to truly represent it, and the language of the Web, may best be illustrated by comparing a page of concordances obtained from a site such as WebCorp with one from Wordbanks Online or any other large monolingual corpus (see figures below). It seems that certain language varieties, for example literary language, are virtually non-existent or seriously under-represented on the Internet, while others like commerce, computers, or the informal chatty style of web forums claim a substantial part of bandwidth.

1. REPORTS AUDIONEWS Morning **home** delivery - save up to 65%
2. when the snow flies at **home** Today's Markets Market Change Value
3. adding a deck to your **home**? Google SketchUp makes it faster
4. Register Go to: Guardian Unlimited **home** UK news World news Comment
5. you normally do on your **home** PC from your mobile phone
6. http://images.apple.com/itunes/**home**/images/2007/11/video_ctp_southpark_home_20071113.jpg
7. Flash Home Deliver rich, dynamic **home** screens Adobe® Flash® Home enables
8. with the company. On the **home** page, the site is described
9. an inescapable feeling in the **home** dressing room last night of
10. and restore software for the **home** and home office that can
11. Internet **home** of: SYMBOL LOOK-UP
12. Search **home** news business tech sports entertainment
13. or transfer files between your **home** and office. Learn More >
14. children a loving and supportive **home**. By providing guidance and stability
15. 'Miracle baby' **home** for holidays PATRICK AIR FORCE
16. you're thinking about turning your **home** into an inn, start here
17. XPointer. Visit the Annotea project **home** page. Amaya - Open Source Amaya
18. to secure voting in his **home** state, New Jersey. November 21
19. of Central England , Birmingham 22-Nov-2007 **home** journals A-Z subject areas advanced
20. to Petfinder.com! The temporary **home** of 250,000 adoptable pets
21. with your choice of wireless **home** networking. View offer details
22. easier time in surgery Pigeons **home** in on nooks in tall
23. Page This is W3C 's **home** page for the XHTML2 Working
24. Welcome to IngentaConnect The **home** of scholarly research Search 23
25. The Internet **home** of: SYMBOL LOOK-UP
26. turkeys will be eaten at **home** today Rosie Stiyer Across the
27. Buying, selling and renting a **home** Home ownership schemes Private renting

Figure 1: Concordances from WebCorp

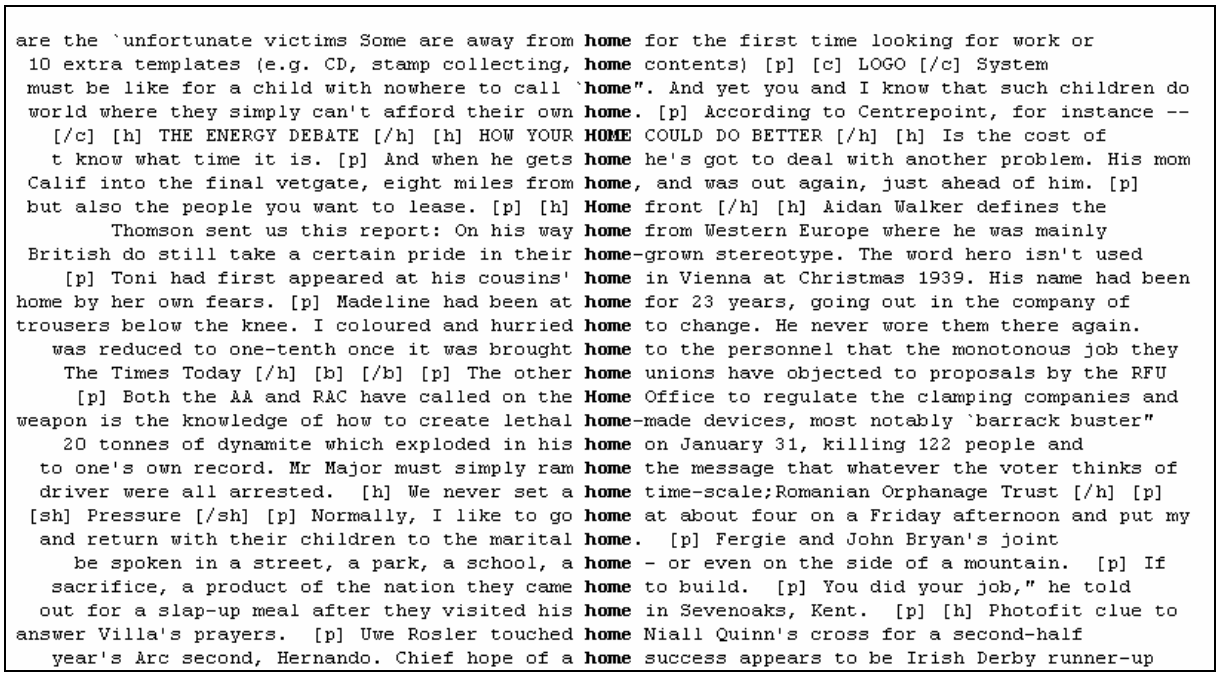

Figure 2: Concordances from Wordbanks Online (formerly known as the Bank of English)

The availability of corpora has not gone unnoticed by technically adept translators. The most valuable corpus type, often designed explicitly for translators, is of course the parallel corpus, giving each language segment in two or more languages. By offering translations of segments instead of equivalents of words, a parallel corpus shifts the translator's attention from a lexical item to an item of meaning. It is impossible to study an expression without its context, and authentic examples of previous translation solutions offer a broader insight into pragmatic equivalence than traditional dictionaries, especially ones intended mainly for decoding. It should be noted here that some dictionaries do provide collocational context and thus attempt to better satisfy the needs of translators.

Another important feature of corpora is the imperfection of language use. Any collection of real texts will contain typos, passages of bad style in the original or translation, and of course translation solutions that run the gamut from excellent to misleading or simply wrong. A corpus user must find a way to critically judge the solutions proposed by the corpus and evaluate them according to the type and contents of the corpus. This point is especially crucial in translator training, not only at the academic level involving translation students, but in all settings where the corpus-using translator works into a non-mother tongue or has a less-than-perfect understanding of the source language.

A monolingual corpus is an equally valuable resource, though usually for different purposes. As monolingual corpora are generally larger and, in some cases, may be considered representative, they are able to offer information about more or less standard language use on the basis of quantitative data. Moreover, a monolingual corpus can be an important

source of translation equivalents for specific expressions, technical terms, or recent borrowings, naturally requiring different search strategies.

Unlike the dictionary, a concordance leaves it to the user to work out how an expression is used from the data. This typically calls for more in-depth processing than does consulting a dictionary, thereby increasing the probability of learning. In more general terms, by drawing attention to the different ways expressions are typically used and with what frequencies, corpora can make learners more sensitive to issues of phraseology, register, and frequency, which are poorly documented by other tools (Aston 1999).

In the remainder of this paper we give an overview of the possibilities of exploiting corpora both in practical translation work and in translation research. We first outline the present state of language resources in Slovenia, focusing on publicly available resources that can be used and re-used for a variety of purposes. Then we give examples of corpus exploitation in translation work and in translator training, in the creation of translation-relevant terminological resources and in translation research. The concluding discussion shows that there is still room for a more systematic approach to corpus-based research of translation phenomena, and that the resources we have available at present literally call for such analyses.

2. An Overview of Mono- and Multilingual Corpora of Slovene

In fairness to various teams and researchers working on Slovene corpora, it should be noted that this section attempts to include only corpora that can be accessed on the Web and that may be considered a translation resource. We therefore will not be concerned with speech databases and collections that have been assembled intended for speech technologies, privately owned mono- and multilingual corpora that have been compiled for the purposes of developing a Machine Translation system, nor with all other text collections that cannot be accessed and are not distributed.

2.1. Monolingual corpora

Among the first Slovene electronic text collections was an online repository of Slovene literature compiled by Miran Hladnik. This digital library was founded in 1995 and is still being updated, however the literary texts today form part of a much larger corpus project, *Nova beseda*, containing 202 million words at the time of writing. *Nova beseda* is being compiled by the Fran Ramovš Institute of the Slovene Language, which is part of the Slovene Academy of Sciences and Arts, and is freely available for online querying. The corpus is composed mostly of the Slovene daily newspaper *Delo* (150 million words), while the rest is taken

from the collection of literature mentioned above, the computer monthly *Monitor* and some other minor text sources.

The interface to *Nova beseda* is rather simplistic and does not offer many advanced options for corpus querying or processing hits. The texts of *Nova beseda* have not undergone any linguistic analysis, hence only word form search is possible, with the wildcard characters * and ?. The corpus does not claim to be balanced or representative in any respect, as it contains a very narrow selection of text sources (see above). However, the user has the possibility of restricting the search to an individually defined subcorpus through an easy-to-use bibliographic taxonomy.

The other large Slovene corpus is *FIDAplus*, a 621-million words reference corpus of Slovene, which was compiled within a government-funded research project launched in 2004. As the main objective of a reference corpus is to be as representative of the language in all its varieties as possible, considerable efforts were invested into building a balanced corpus out of a much larger text collection (Arhar and Gorjanc 2007).

The corpus was morphosyntactically annotated by Amebis, a Slovene language technologies company responsible for most commercially available language tools for Slovene. The annotation includes complex morphosyntactic descriptions, i.e. not just part-of-speech tags but an array of all grammatical categories associated with the word form. Furthermore, the corpus has been lemmatised, meaning that to each word form its uninflected or unconjugated base form (lemma) has been added. In the case of multiple possible lemmata a process of disambiguation was carried out, selecting the correct lemma for the given context.

The corpus can be accessed via the ASP web concordance engine and is free for research purposes upon registration.

2.2. Special language corpora

For mono- and multilingual terminology work, another corpus type is extremely useful: the domain-specific, special language or sublanguage corpus. For this corpus type it is important that it is representative of the domain in terms of the text types contained and the currency of the texts. Such a collection can almost never be entirely bilingual, because a special domain is best represented by a collection of crucial texts in one language. Several criteria should be considered when compiling a sublanguage corpus (Pearson 1998: 56):

- Register. A special domain like genetics will typically be described in texts of various registers, e.g. scientific papers, college textbooks, articles in popular scientific journals etc. Register can have a considerable influence on the terminology used and the style.

- Quality. Although in itself a slippery issue, texts do differ in the amount of effort invested into all stages of their production, from authorship to typesetting and printing. Corpora generally should not impose normative restrictions; for domain-specific corpora however, certain texts might be inappropriate on the grounds of poor quality.

- Original translated or written in a foreign language. Non English-native speaking researchers publish most of their work in languages other than their own; they either write in English or have their texts translated into the target language. Such texts should by no means be considered substandard because they too constitute the language reality in a given domain. We should, however, be aware of the characteristics of such texts and possible inconsistencies resulting from them.

In addition to the two large general language corpora for Slovene mentioned above, within the last few years several projects have yielded a set of specialised monolingual corpora. One covers the domain of information science and includes the proceedings of the largest Slovene IT conference, DSI, from 2003 to 2007 (Days of Slovene Information Science). At the time of writing it has 1.2 million words and is available for online querying at http://nl2.ijs.si/index-mono.html. The corpus is described in more detail by Erjavec and Vintar (2004). The domain of Informatics is a highly productive and terminologically challenging one for all non-English languages, and a monitor corpus is the best way to follow language development fuelled by technology. The DSI corpus was compiled as support for Islovar, the interactive online Slovene–English terminological dictionary of Informatics.

Another such corpus consists of texts from the domain of Public Relations, it contains just under 2 million running words and is available for online searching at http://www.korp.fdv.uni-lj.si/.

2.3. Multilingual corpora

As a small language with close contacts with other linguistic communities, Slovene has a high level of translation activity. Accordingly, the need for and appreciation of multilingual resources have fuelled several projects compiling parallel corpora. The most interesting as well as easiest to obtain is the language pair Slovene–English, which is by now very well served: the total size of freely available Slovene–English parallel corpora amounts to over 35 million words.

Other languages lag far behind, with the notable exception of Evrokorpus, now containing a Slovene-German and a Slovene-French part of the corpus.

2.3.1. MULTEXT-East

The name refers to a large initiative, within which a set of corpora and tools were built or made available, covering a large number of mainly Central and Eastern European languages (Erjavec 2004). The most important component is the linguistically annotated corpus consisting of Orwell's novel *1984* in the English original and translations. The resources are the result of several EU projects: MULTEXT-East (produced linked resources for Romanian, Slovene, Czech, Bulgarian, Estonian, Hungarian, and English), TELRI (added resources for Lithuanian, Croatian, Serbian, and Russian; first release), and CONCEDE (validation, re-encoding; partial re-release). This dataset, unique in terms of languages and the wealth of encoding, is extensively documented (see Multext-East website), and freely available for research purposes, upon signing the licence agreement.

2.3.2. IJS-ELAN

The IJS-ELAN Slovene–English parallel corpus includes 15 texts from various domains; the total size of the corpus is 1 million words (Erjavec 2002). The basic idea behind this project was to build as big a parallel corpus as possible, in the quickest way possible. The already existing MULTEXT-East corpus, consisting of Orwell's *1984*, was expanded through a further 14 texts, ranging from EU legislation and pharmacology to computer manuals and localisation files. As text availability was the main criterion in building this corpus, the selection is quite haphazard. An online concordancer was set up shortly after the texts had been pre-processed and the corpus has since been used for a variety of purposes, including as a translation resource.

2.3.3. Trans

The Trans Slovene-English parallel corpus was compiled in a rather unspectacular manner – as a student project at the Department of Translation, University of Ljubljana. It contains 1 million words and was compiled specifically for translation purposes, which meant that the number of domains covered by the texts was deliberately limited to five: medicine, geology, tourism, nuclear engineering, and public administration. The corpus was made available for online searching at the same address as the IJS-ELAN corpus (see previous section).

2.3.4. Evrokorpus

The largest translation project in Slovenian history was the translation of the *acquis communautaire*, a prerequisite for accession to the European Union and a foundation for all legal and administrative matters concerning EU. As the majority of translation work was performed by the Office of the Government of the Republic of Slovenia for European Affairs using

Translation Memory tools, the resulting databases of bilingual segments could easily be converted into a searchable corpus. The first such collection was made available by Miran Željko in 2002 under the name Evrokorpus (http://evrokorpus.gov.si/ [21 Nov. 07]). The corpus has since grown to the astounding size of 34 million words of Slovene-English parallel materials, 1 million for Slovene-German, and about 200,000 words for Slovene-French. It is being regularly updated with new aligned translations and represents an invaluable resource for translators and terminologists, but also legal experts and others working in the EU domain.

The fact that during the process of EU enlargement most texts produced were made publicly available as a parallel corpus is an unprecedented advantage for translators from and into Slovene. Combined with the terminology database Evroterm, this is a unique infrastructure ensuring consistent translations in all EU-related domains (Željko 2004).

3. Corpora in Translation Practice and Translator Training

For a translator, a corpus is one of the sources of linguistic information, either on the lexical level when searching for translation equivalents or on other levels when seeking to produce a functional translation. As the primary source of lexical information, most translators still rely on dictionaries, although in special domains term banks may be used much more often than general language dictionaries. Bowker and Pearson (2002: 15) list five problems with dictionaries, especially in the context of LSP, for which corpora may provide a remedy:

1. Incompleteness. It takes a long time to compile and publish a dictionary, thus in many cases a dictionary no longer reflects the current state of knowledge or language.

2. Size, especially since most dictionaries are still compiled for a printed version. Large, multi-volume dictionaries may cover a specialized field in its entirety, but people would not want to carry them around. Also, lexicographers have to make choices about which information to include and which to leave out, and their choices do not always meet the needs of LSP language users (or translators for that matter).

3. Lack of contextual or usage information.

4. Lack of frequency information. The choice of a lexical equivalent for a translator can be made easier if they know about the domain-specific usage patterns, including the frequency of lexical items.

5. Even if the dictionary contains the relevant information, users may have difficulties finding it. For example, does one search for HTML or

hypertext markup language? And if only markup needs to be translated, how to locate it in the printed dictionary?

If an item is not found in the dictionary, the next stage is usually Google. Clearly the Internet is a gold mine for translators, as it contains up-to-date documents on almost all subjects in many languages of the world (Fletcher 2004). However, most documents on the Web are not bilingual, and the quest for translation equivalents requires efficient and innovative search strategies.

The Web is in itself a large multilingual corpus, and there are tools available that facilitate its use as a corpus, such as KwicFinder. We might say that between these two extremes, dictionaries as static, normalised, structured data and the Web as unstructured, chaotic, abundant data, there are corpora, some that already exist and some we might build ourselves.

3.1. Using existing corpora

For first time corpus users, it is important to draw attention to some key issues:

1. Corpora are not dictionaries. Texts may contain language usage that does not correspond to what is considered standard or correct.

2. When using non-lemmatised corpora of highly inflectional languages, the search for the base form will return only a small portion of possible hits. The linguistic pattern of the base form may differ from the patterns of inflected forms.

3. Results from a corpus require critical interpretation. Frequency information should be interpreted according to corpus composition.

A monolingual corpus of the mother language will naturally be used for different purposes than a corpus of a foreign language. An interesting feature is the search for translation equivalents in a monolingual corpus. An English word like *spam* will first occur in Slovene as a borrowing, so the search for *spam* in *Nova beseda* returns a considerable number of hits, where, if we examine the context, several possible translation equivalents occur near the borrowed word, e.g. *nezaželena pošta* 'unsolicited mail', *elektronske smeti* 'electronic garbage', *nenaročena oglasna pošta* 'unsubscribed advertising mail', etc.

Translators from and into Slovene working with texts concerning the EU or other political matters can no longer imagine life without Evrokorpus and its associated term bank Evroterm. It is interesting that in this respect Slovenia was pioneering within the EU; Evrokorpus was released in 2002, while a similar multilingual resource containing the *acquis communautaire*

and its translations into all EU languages was only opened for public use in November 2007 (see "The DGT Multilingual Translation Memory of the Acquis Communautaire: DGT-TM.").

An example of how a corpus can provide the correct lexical equivalent while neither a bilingual dictionary nor a term bank contain the established expression is given below:

> However, a number of stocks both in Community and non-Community waters have continued to decline and it is consequently necessary to improve and extend existing conservation measures.

For a translator unfamiliar with EU fishing jargon, *stock* might be a familiar word but difficult to translate in this context. The English-Slovene bilingual dictionary offers over a hundred various equivalents for each of the three meanings of *stock*, however none is the right one for the above sentence. The Evroterm term bank lists two possible translations, *zaloga* [as in *commercial stocks* or *emergency stocks*] and *delež*, '*share*'. An examination of concordances found in Evrokorpus quickly reveals the collocation *fish stocks*, and the correct translation *(ribji) stalež*.

Although we normally think of corpora as synchronic resources portraying language at a certain point in time, some interesting studies into term formation have been made for Slovene and English, for example for the field of mobile communications (Glavan 2004). The *Nova beseda* corpus was used for diachronic research by building several subcorpora according to the year of publication. In this way the frequency of terms like *mobitel*, *WAP*, *wapanje,* etc. could be explored year-wise and the tendencies of terminological development quantified.

An important resource that will enable researchers to perform empirical studies of the influence of translation on language development is the AHlib digital library (Prunč 2005). This large – as yet unfinished – diachronic corpus will contain the majority of translations from German into Slovene and Croatian from the period 1848-1919 and is being created within two parallel national projects, one funded by the Austrian government (FWF P17465) and one by the Slovene government (J6-6078). The digital library consists of the following parts:

- 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;

- 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

lemmatisation), analysis of historical wordforms, and finally conversion into TEI (Erjavec 2007).

The potential of this extensive corpus for translation studies and other disciplines has already been shown in several publications (Lipavic Oštir 2007, Vintar 2007).

3.2. Using self-made corpora

For many languages, special domains, or language pairs, there are no available corpora. On the other hand, the internet is an infinite source of documents and texts on all possible subjects, some available in two or more languages. In addition, most people are in the habit of storing their translation projects on hard drives, and if there were a systematic way of searching through all these files, the process of retrieving previously used items of information might be much faster and easier.

Arguments in favour of compiling one's own corpora are many, although to most people the effort seems too strenuous considering the potential benefits. Especially in view of translation memories and the idea of reusability behind them, it seems that bilingual text collections are gaining ground as key resources in translation. Of course, the purpose of these two types of resources differs to a great extent. While translation memories provide reusability only at the rather rigid level of sentence similarity, (bilingual) corpora provide insight into language or translation solutions on almost any imaginable level.

At the Department of Translation in Ljubljana we have undertaken several student projects for compiling bilingual corpora. Such corpus projects have certain limitations compared to corpora compiled within research projects:

- All tools and methods demonstrated should be available to students inside as well as outside the classroom. The experiment should be completely replicable in any other out-of-the-classroom setting.

- All tools should be free and, if possible, run on Windows.

- Translation students generally cannot program, and all data manipulation must be performed using standard text processing software and non-exotic file formats.

The following sections briefly describe the stages involved in building a corpus and the tools available.

3.2.1. Collecting and pre-processing texts

According to Sinclair (1991: 171), a corpus is "A collection of naturally occurring language text, chosen to characterize a state or variety of a language." The choice of texts should therefore be concerned with the representativeness of a corpus, even if only a small domain is to be represented. Of course, in bilingual corpora it is even more difficult to satisfy this criterion; nevertheless the composition of the corpus should at least be thoroughly discussed. The purpose of this discussion is to clarify issues of corpus size, number of domains included, text types, language(s) and/or the language of the original, possible text sources, copyright, etc.

Once the project has a set of clearly defined objectives in terms of text collection, some technical questions also need to be resolved. Which file formats can be successfully handled? If the main source of texts will be the internet, HTML will need to be handled; if, on the other hand, we expect text donations from translation agencies or private entities, MS Word is likely to be the most common format. Which character encoding should be used? Probably Unicode or UTF-8, although older tools might have problems displaying them. Which encoding should be chosen for the entire corpus? If we are building a resource that should be used and distributed as widely as possible, we should probably choose TEI encoding (Sperberg-McQueen and Burnard 2002). However, without appropriate computational knowledge this standard is not trivial to implement.

3.2.2. Alignment

If we are building a parallel corpus, the texts will need to be sentence-aligned. If we can obtain a licensed copy of SDL Trados WinAlign, alignment is an easy task. A sentence alignment utility is offered by several other translation memory packages (such as ATRIL's DVX), as well as by the parallel concordance tool ParaConc, available for a relatively modest fee.

Sentence alignment is usually a semi-automatic procedure, where the tool proposes sentence pairs, which must be manually corrected in the event of errors. Most commercial alignment utilities can handle various file formats, including HTML, Word, or XML files.

3.2.3. Offline concordancing

A number of tools are available for concordancing at modest prices. A widely known toolkit for monolingual text analyses is Wordsmith Tools by Mike Scott. While perfectly adequate even for advanced corpus linguists working with monolingual corpora, it is of very limited use for querying bilingual corpora. For the latter, the above-mentioned ParaConc is a good option.

Experience gained in this area shows that building bilingual corpora in an educational setting is not only a useful exercise and a corpus-awareness-raising activity, but also an undertaking that produces extremely valuable resources for the entire translation community.

4. Conclusions

A few years ago corpora were unexplored terrain for many practising translators and translation tutors alike. This situation seems to be changing both because translators are required to produce high-quality translations in a shorter time than before and because electronic language resources are more accessible than before. The aim of this article has been to present the situation in Slovenia and regarding the Slovene language, which – with its just over 2 million speakers – counts among the smaller language communities in Europe. Nevertheless, in the field of bilingual freely available language resources, Slovene is considerably well provided for. Not many languages can boast an online parallel corpus of over 34 million words, and corpus-related activities in the context of translator training by now have the status of a well established tradition.

Having corpora available is, however, only a basis for linguistic and translatological research, and in this respect there is plenty of room for future work. In the field of corpus-based translation studies, the properties of translated texts have been studied and compared to original text production within a language (Baker 2004). Such studies can yield interesting insights not only into the differences between translated and original texts, but also into the cognitive processes underlying translation. Thus far no extensive study of this kind has been conducted for Slovene, we do however hope that with the availability of Fidaplus, which contains a large portion of translations into Slovene, this gap will soon be closed.

References

- **Aston, Guy** (1999). "Corpus Use and Learning to Translate." *Textus* 12, 289-314.
- **Arhar, Špela and Gorjanc, Vojko** (2007). "Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa." *Jezik in slovstvo* 52(2), 95-110.
- **Baker, Mona** (2004). "A Corpus-based View of Similarity and Difference in Translation." *International Journal of Corpus Linguistics* 9(2), 167-193.
- **Bowker, Lynne and Pearson, Jennifer** (2002). *Working with Specialized Language: A Practical Guide to Using Corpora.* London: Routledge.
- **Erjavec, Tomaž** (2002). "The IJS-ELAN Slovene–English Parallel Corpus." *International Journal of Corpus Linguistics* 7(1), 1-20.
- **Erjavec**, **Tomaž** (2004). "MULTEXT-East Version 3: Multilingual Morphosyntactic Specification, Lexicons and Corpora." M.T. Lino, M.F. Xavier (eds) (2004) *Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26th, 27th & 28th May 2004. LREC 2004: held in memory of Antonio Zampolli: proceedings (1535-1538).* (2004). Paris: European Language Resources Association.
- **Erjavec, Tomaž and Vintar, Špela** (2004). "Korpus kot podpora slovarju informacijskega izrazja slovenskega jezika." *Uporabna informatika (Applied Informatics*), 12(2), 97-106.
- **Erjavec, Tomaž** (2007). "Architecture for Editing Complex Digital Documents." *InFuture 2007, Digital Information and Heritage*. Zagreb, Dept. of Information Science, 105-114.
- **Fletcher, W.H.** (2004). "Facilitating the Compilation and Dissemination of Ad-hoc Web Corpora." G. Aston, S. Bernardini and D. Stewart (eds) (2004) *Papers from the Fifth International Conference on Teaching and Language Corpora*. Amsterdam: Benjamins. Available at http://kwicfinder.com/Facilitating_Compilation_and_Dissemination_of_Ad-Hoc_Web_Corpora.pdf.
- **Glavan, S**. (2004). *Terminotvorje v slovenščini na primeru izrazja mobilne telefonije*. BA Thesis. Faculty of Arts, University of Ljubljana.
- **Lipavic Oštir, A.** (2007). "Die Reblaus - Trtna uš (1881): prevod v družbenem, socialnem in gospodarskem kontekstu. *Slovenski prevodi nemških besedil v obdobju 1848-1918*. Maribor: Faculty of Arts, Department of Translation, 7-8.
- **Pearson, Jennifer** (1998). *Terms in Context*. Amsterdam/Philadelphia, John Benjamins.
- **Prunč, Erich** (2005). "Hypothesen zum Gattungsprofil deutsch-slowenischer Übersetzungen im Zeitraum 1848-1918." In: Kocijančič Pokorn, Nike et al. (eds) (2005). *Beyond Equivalence*. Graz: Institut für theoretische und angewandte Translationswissenschaft, 19-38.
- **Sinclair, John** (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- **Sperberg-McQueen, C.M. and Burnard, Lou** (eds) (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange.* XML Version. Oxford, Providence, Charlottesville, Bergen: Text Encoding Initiative Consortium.

- **Vintar, Špela.** (2007). "Cultural and Scientific Transfer through Translation – a corpus-based study of term formation in the period 1848-1919." *INFuture 2007, Digital Information and Heritage.* Zagreb, Dept. of Information Science, 289-298.

- **Željko, Miran** (2004). "Evroterm in Evrokorpus – terminološki slovar in korpus prevodov." Humar, M. (ed.) (2004). *Terminologija v času globalizacije*. Ljubljana: Scientific Research Centre SASA, ZRC Publishing, 139-149.

Electronic resources

- "Amebis." On line at http://www.amebis.si (consulted on 21.11.2007)

- "A www concordance service." On line at http://nl2.ijs.si/index-mono.html (consulted on 21.11.2007)

- "Evrokorpus." On line at http://evrokorpus.gov.si/ (consulted on 21.11.2007)

- "Evroterm." On line at (http://evroterm.gov.si/ (consulted on 24.09.2006)

- "FIDAplus." On line at http://www.fidaplus.net (consulted on 21.11.2007) 2007)

- "Islovar." On line at http://www.islovar.org (consulted on 21.11.2007)

- "Korp." On line at http://www.korp.fdv.uni-lj.si/ (consulted on 21.11. 2007)

- "KwicFinder." On line at http://miniappolis.com/KWiCFinder/KWiCFinderHome.html (consulted on 21.11.2007)

- "MULTEXT-East." On line at http://nl.ijs.si/ME/ (consulted on 21.11. 2007)

- "Nova beseda." On line at http://bos.zrc-sazu.si/a_beseda.html (consulted on 21.11. 2007)

- "ParaConc." On line at http://www.athel.com/para.html (consulted on 21.11. 2007)

- "Parallel Concordance Service." On line at http://nl2.ijs.si/index-bi.html (consulted on 21.11. 2007)

- "The DGT Multilingual Translation Memory of the Acquis Communautaire: DGT-TM." On line at http://langtech.jrc.it/DGT-TM.html (consulted on 11.01.2008)

- *Zbirka slovenskih leposlovnih besedi*l (repository of Slovene literature compiled by Miran Hladnik). On line at http://www.ijs.si/lit/leposl.html-l2 (consulted on 21.11. 2007)

- "WebCorp." On line at (http://www.webcorp.org.uk/ (consulted on 21.11. 2007)

- Wordbanks Online (formerly known as Bank of English). Online at http://www.collins.co.uk/corpus/CorpusSearch.aspx (consulted 21.11. 2007)

- "WordSmith Tools." Online at http://www.lexically.net/wordsmith/index.html (consulted 21.11. 2007)

Biography

Špela Vintar is Assistant Professor of Translation at the University of Ljubljana. She has taken part in several language technologies related research projects in Slovenia and abroad, including projects on automatic terminology extraction from bilingual corpora, speech-to-speech translation and semantic text processing. For more information see http://lojze.lugos.si/~spela.

E-mail: spela.vintar@guest.arnes.si