RICHARD KOVÁČ
ANDREEA-NORA POP

BUILDING A SLOVAK-ROMANIAN PARALLEL CORPUS

1. Introduction

One of the major fields of research of the Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences in Bratislava is corpus linguistics and natural language processing. Apart from compiling the Slovak National Corpus, the Corpus of Spoken Slovak, the Slovak Terminology Database etc., researchers are also interested in bilingual resources (combination of Slovak and: English, Czech, Russian, German, French, Bulgarian, Hungarian and Latin). For more details about the issues of the Slovak National Corpus department[1], see Šimková, Garabík (2014).

In the panorama of parallel corpora, the Slovak-Romanian parallel corpus was also compiled. Although a small project, this approach aims to provide language tools for a rare pairing of languages.

The access to the corpora of the Slovak National Corpus department is free after processing the registration[2] and then logging in through the ***NoSketch Engine***[3] tool – see more in section **3.7.**

2. Parallel Corpora

Corpus linguistics is a relatively new discipline, which represents the study of language through large collections of texts in electronic form. Language analysis is done by means of software, known as corpus analysis tools. There are many types of corpora: general/reference *vs* specialized corpora, learner corpora, monolingual *vs* multilingual corpora, synchronic *vs* diachronic corpora, open *vs* closed corpora (Bowker, Pearson 2002, p. 1, 11–13).

In the multilingual field, corpora can be divided into two categories:

Translation corpora: texts stand in a translational relationship to each other; they can be a translation of an absent original or one of them can be the original and the other(s) translation(s);

Comparable corpora: texts are similar samples regarding external criteria (spoken *vs* written language, register); equivalence is established among the main linguistic features of the corpora (Tognini-Bonelli 2001, p. 6–7).

[1] http://korpus.juls.savba.sk/
[2] http://korpus.sk/registration_en.html
[3] https://bonito.korpus.sk

Parallel corpora belong to translation corpora. They generate bilingual or multilingual concordances, revealing cross-linguistic correspondences and differences that are impossible to discover in a monolingual corpus. When entering a key word, the software retrieves all the sentences where the word appears in one language, together with the corresponding translation in the other language (Bowker, Pearson 2002, p. 92–93; Tognini-Bonelli 2001, p. 6–7).

3. The Slovak-Romanian Corpus

In the following section, the process in which a parallel corpus is built, will be described, with examples for the Slovak-Romanian one.

A corpus consists of separate books, articles or other resources, which make up the corpus **archive**. For a parallel corpus, both an original text and its translation are needed and sometimes translations from a third language (mostly English) are also acceptable.

The Slovak-Romanian corpus currently consists of 3 books by Romanian authors and 1 government document. So, it is a minimalistic, experimental corpus.

Author	Romanian title	Slovak title	Translator
Mircea Eliade	*Noaptea de Sânziene*	*Svätojánska noc*	Jana Páleníková
George Călinescu	*Enigma Otiliei*	*Záhadná Otília*	Jana Páleníková
Constantin Chiriță	*Cireșarii I – Cavalerii florii de cireș*	*Tajomstvo Čiernej jaskyne*	Elena Žitná
	Program de colaborare între Ministerul Culturii din Republica Slovacă și Ministerul Culturii din România pentru perioada 2016–2020	*Program spolupráce medzi medzi Ministerstvom kultúry Slovenskej republiky a Ministerstvom kultúry Rumunska na roky 2016–2020*	

Table 1. *Overview of the Slovak-Romanian corpus resources*

3.1. Adding a digital resource

Nowadays many books, articles and documents are available on the Internet. Books are preferable for corpora, since they bring larger amount of text with good stylistic quality; however, because of copyright issues, most of freely available books are of older origin.

Another possibility is to get them directly from authors, translators or publishers via an agreement (license).

Furthermore, an electronic version may be achieved by scanning the paper version, doing OCR (Optical Character Recognition) and correcting the text (the so-called **reading**), but this work is time-consuming.

The amount of available resources for a parallel corpus is strongly reduced due to the fact that versions in both languages are needed.

3.2. Converting source files into text

Source files are in various formats (*doc*, *epub*, *PDF* etc.). Some of them can be easily converted into a pure text. Formats like *PDF* require some post-corrections: removing headers, page numbers, footnotes, joining paragraphs separated by a page boundary etc. Content, index and all parts which are not matching in both languages are also removed and only a simple, paragraph-structured text is left.

3.3.Tagging

In the next step, a text is split into a sequence of **tokens**. A token is any word or a non-word element such as number, punctuation etc. Sentence boundaries are inserted by a segmentation tool. A **tagger** then labels words with their **lemmas** (basic forms) and **morphological tags** (word class and its attributes).

For the Romanian language, we used the freely available ***TreeTagger***[4] with a Romanian parameter file, while for Slovak, the Slovak National Corpus department uses a specialized morphological analysis tool, ***MorphoDiTa***[5], with a Slovak tagging model trained on proofread corrected morphological data (Garabík, Šimková 2012).

A tagged **vertical** file (one token per line) forms a basis to be compiled into a corpus.

3.4. Bibliographical annotation

From the book colophon, we notice a brief bibliographical annotation into a separate file. It contains the author, the title, the publishing house, the year of publishing, the source language, the translator's name and the ISBN, if present. A unique **identifier** code is assigned to each archive resource.

3.5. Aligning

All steps described above are generally taken for each corpus. But **aligning** is what makes a corpus **parallel**. For this, we used an OpenSource tool, ***hunalign***[6]. It aligns texts in two languages, sentence by sentence, trying to maximize the matching score. Sometimes, it happens that to one sentence in a language, none or more sentences in the other language are assigned, depending on the translation. For example, for two Romanian sentences, one Slovak sentence was assigned:

Doctorul continuă: — Mâine, când te duci la școală, spune domnului institutor că n-ai să mai vii.	— Keď zajtra pôjdeš do školy, — **pokračoval doktor,** — povedz pánu učiteľovi, že už viac neprídeš.

Table 2. *Example of a multi-sentence alignment*

Inserting sync marks (we used simply ***) into the texts in both languages (e.g. after each chapter) helps the aligner to keep the track and a better aligning result will we achieved. An optional **dictionary** can also improve this process.

[4] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[5] http://ufal.mff.cuni.cz/morphodita
[6] http://mokk.bme.hu/resources/hunalign/

Since we did not have any Romanian-Slovak dictionary, one was bootstrapped by *hunalign* itself (it has a useful option for creating a dictionary ex post from aligned texts). We first used two books and selected only matching word pairs to avoid mistakes.

Also, it is necessary to check whether the translation corresponds to the original. This is usually found out after alignment checking. Sometimes, translations are shortened too much or they are done from other text versions, which makes them unusable.

3.6. Corpus finalizing

Tagged vertical files are concatenated (for both languages separately) to create the final corpus files. Every sentence in each of the languages has its index number and an alignment file is added, containing matching sentences' index pairs (this is specific for a parallel corpus).

The present version of the Slovak-Romanian corpus consists of **45 524** Romanian and **46 070** Slovak sentences, **688 867** Romanian and **603 111** Slovak tokens.

3.7. Corpus queries

The corpus can be viewed and searched using the ***NoSketch Engine***[7] tool.

The **word list** function creates a list of the words appearing in the corpus, ranked according to their frequency. The total number of items is **37 245** Romanian words/**53 384** Slovak words, **16 380** Romanian lemmas/**21 718** Slovak lemmas.

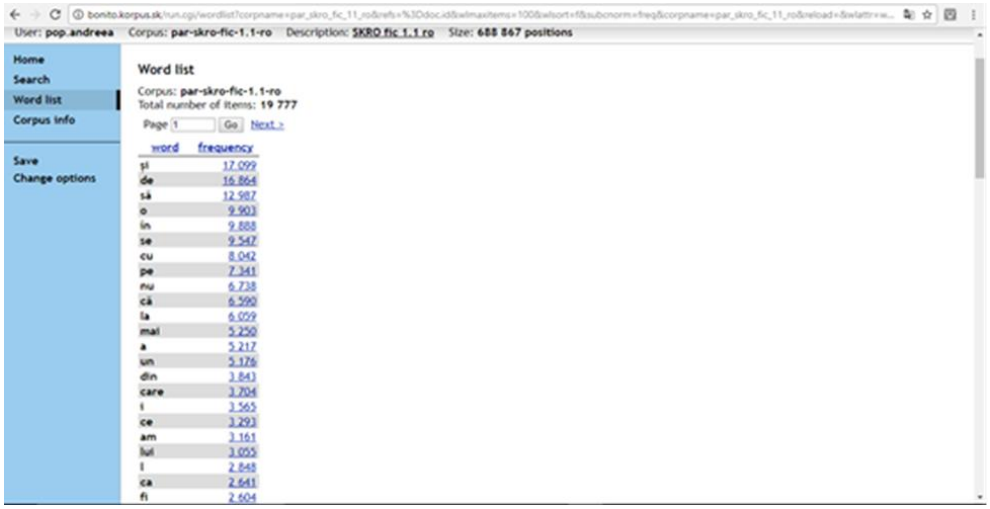

Figure 1. *The Romanian word list*

Queries in the corpus may be simple or filtered through several types, such as lemma, phrase, word, collocation, according to the right and left context or according to the selected subcorpus.

[7]https://bonito.korpus.sk

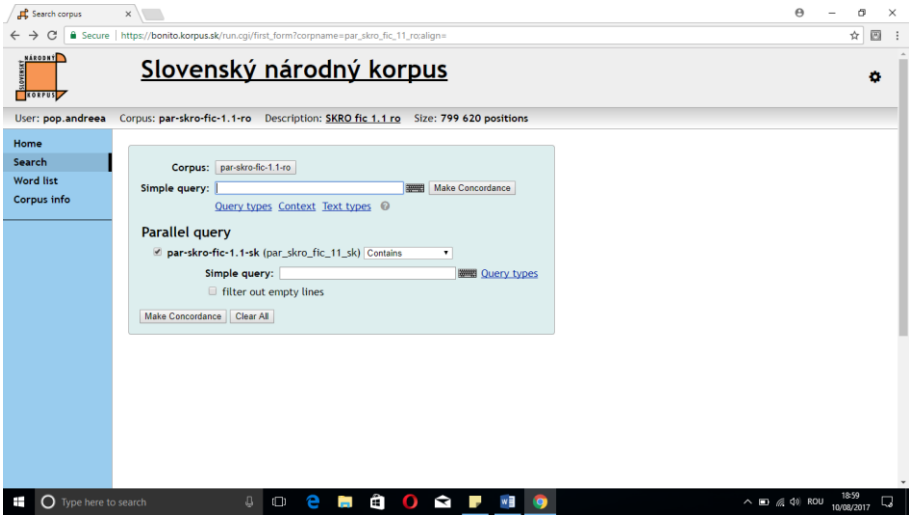

Figure 2. *The query window in the Slovak-Romanian Corpus*

We would like to exemplify the way the corpus can be used. When searching for the word *carte*, the following concordances are displayed:

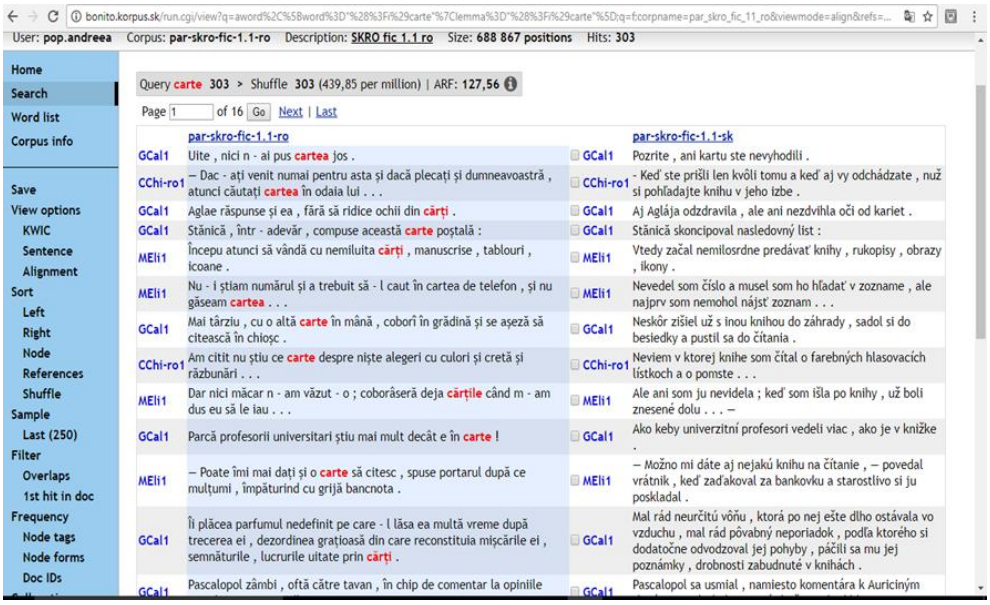

Figure 3. *The concordances of the word* ***carte***

It can be noticed that the word has 303 hits, such as:

GCal1	Ea era o fire bolnăvicioasă şi iritabilă şi stătea mai toată vremea întinsă pe o canapea într-o odaie, citind câte o **carte**, strigând şi poruncind slugilor prin uşa întredeschisă.		**GCal1**	Bola to chorľavá a prchká bytosť, celý čas ležala v izbe na pohovke, čítala knihu a cez odchýlené dvere pokrikovala na sluhov a vydávala im rozkazy.

Figure 4. *An example of a concordance of the word **carte***

Also, one may want to see all the occurrences of the above-mentioned noun joined by the definite or indefinite article or the noun in the plural and to investigate into the translators' approach. The following figure offers a few examples of the occurrences in the corpus in the form of the definite article, with the corresponding sentences:

GCal 1	Orice sugerare a unei alte metode fu respinsă, şi Felix trebui să-i dicteze, privind **cartea** pe deasupra umărului lui, în vreme ce el scriea atent traducerea, făcând să i se repete câte o vorbă.		**GCal1**	Každý návrh na použitie inej metódy zamietol a Félix mu musel diktovať. Díval sa mu cez plece do **knihy** a Titi si pozorne zapisoval preklad, pričom ho tu a tam požiadal, aby mu zopakoval nejaké slovo.
GCal 1	Acesta negă cu capul, luând **cartea** ca pe un album de preţ şi răsfoind-o atent cu degetele uşor apropiate de muchia filelor.		**GCal1**	Ten záporne pokrútil hlavou, vzal **knihu** ako nejaký vzácny album a pozorne v nej listoval, prstami sa zľahka dotýkajúc takmer len okrajov strán.
GCal 1	Atunci Titi scoase din cutie **cartea**, pe care Felix n-o recunoscu la început, fiindcă era îmbrăcată, cu îngrijire, în hârtie albastră.		**GCal1**	Nato Titi vytiahol zo škatule **knihu**, ktorú Félix hneď nespoznal, pretože bola starostlivo zabalená do modrého papiera.
GCal 1	— N-am avut vreme să-l citesc, spuse calm Titi, înapoindu-i definitiv **cartea**, altădată poate ţi-o cer!		**GCal1**	— Nemal som ho kedy čítať, — povedal pokojne Titi a definitívne Félixovi vrátil **knihu**, — možno si ho od teba vypýtam inokedy.

Figure 5. *Some of the occurrences of the noun carte, accompanied by the definite article*

Other examples of queries:

For the Romanian word *slab*, out of the 51 hits, 19 were encountered for the Slovak *slabý* (= "weak") and 18 hits for *chudý* (= "thin"). For example:

Simțeam prea multă ură în mine: ura omului **slab**.		**MEli1**	Cítil som v sebe príliš veľkú nenávisť: nenávisť **slabého** človeka.		
Părea mult mai **slab** și mai palid, și semnul rănii părea mai adânc.		**MEli1**	Vyzeral oveľa **chudší** a bledší a jazva po rane akoby bola hlbšia		

Figure 6. *Some of the occurrences of the adjective* ***slab***

For the Romanian word *băiat*, out of the 181 hits, 109 were encountered for the Slovak *chlapec* (= "boy") and 40 hits for *syn* (= "son"). For example:

CChi-ro1	În alt an, premiul fusese decernat unui **băiat** de treisprezece ani, elev mediocru, liniștit, bolnăvicios, care nu se remarcase prin nimic tot timpul anului.		**CChi-ro1**	Inokedy dostal cenu pokojný a chorľavý trinásťročný **chlapec**, priemerný žiak, ktorý sa počas roka ničím ne vyznamenal.	
MEli1	Ea știa, sărăcuța, că atâta dor avusesem și eu: să am un **băiat**!		**MEli1**	Vedela, chudiatko, že aj ja som túžila po tom istom: mať**syna**!	

Figure 7. *Some of the occurrences of the noun* ***băiat***

Another useful function is searching for the collocation candidates of a word. The user selects the left and right collocation distance range, the frequencies thresholds and computational method(s). The candidates for the word *carte* are:

Cooccurrence count	Candidate count		T-score	MI	logDice		
P | N		**poștală**	15	16	3,871	11,058	10,589
P | N		**poștale**	8	9	2,827	10,981	9,715
P | N		**vizită**	7	36	2,640	8,788	9,402
P | N		**joc**	5	66	2,223	7,428	8,794
P | N		**ilustrată**	4	8	1,998	10,151	8,719
P | N		**ilustrate**	4	9	1,998	9,981	8,715
P | N		**pachetele**	4	11	1,998	9,691	8,705
P | N		**joace**	4	23	1,995	8,627	8,651
P | N		**citit**	4	54	1,988	7,396	8,520

Figure 8. *The collocation candidates of the word* ***carte***

Also, a parallel query can be made. For example, for the Romanian *carte*, the word *kniha* has been searched, being used in 201 out of the 303 initial hits from the Romanian version. It is also possible to check all the rest 102 occurrences automatically, by selecting negative filtering (concordances which do not contain the given word).

4. The usefulness of the corpus

The usefulness of this corpus is very similar to other parallel corpora. It can be especially valuable to language teachers and learners, since the collection of texts can be seen as a resource of Romanian authentic texts and a way of getting meta-linguistic competence. Also, people interested in translations could benefit from the corpus, not only students and teachers, but also theoreticians investigating translation techniques, solutions and errors, missing words, the insertion of extra words; to put it another way, they may want to see "how a message is conveyed from one language to the other", "comparing the linguistic features and their frequencies in translated L2 texts" (McEnery, Xiao 2007, p. 4). For example, information about idioms, proper names, noun phrases, syntactic constructions, temporal or aspectual meanings, word order could be retrieved.

Researchers in the field of contrastive studies and lexicographers could use the Slovak-Romanian parallel corpus for the creation of lexical databases, to compile corpus-based bilingual dictionaries or to undertake quantitative analyses. The corpus could be of interest to computational linguists, namely to identify the features of texts which could be "expressed computationally to facilitate the development of alignment programs" (Bowker, Pearson 2002, p. 94–95) and also to develop "applications like machine translations and computer-assisted translation" (McEnery, Xiao 2007, p. 4).

A specific use of a parallel corpus (translation equivalence of demonstrative pronouns for the Slovak-Bulgarian one) is analysed by Dimitrova, Garabík (2014).

5. Conclusions

In conclusion, the development of corpus linguistics is of paramount importance due to the world of globalization and technology we live in. In particular, the recent approach to parallel corpora creates valuable resources for translation, lexicography, contrastive studies, language teaching and acquisition. Concerning the Slovak-Romanian corpus, the collection of texts could be enriched and diversified; also, it should be developed as a bi-directional parallel corpus. Furthermore, resources should be used in conjunction with L1 target and source monolingual corpora for more accurate results (McEnery, Xiao 2007, p. 9).

REFERENCES

Bowker, Pearson 2002 = Lynne Bowker, Jennifer Pearson, *Working with Specialized Language: A Practical Guide to Using Corpora*, London, Routledge, 2002.

Dimitrova, Garabík 2014 = Ludmila Dimitrova, Radovan Garabík, *Translation equivalence of demonstrative pronouns in Bulgarian-Slovak parallel texts*, in *Études cognitive*, Warsaw, SOW Publishing House, 2014, p. 65–74. (http://korpus.sk/attachments/publications/2014-dimitrova_ garabik _pronouns.pdf)

Garabík, Šimková 2012 = Radovan Garabík, Mária Šimková, *Slovak Morphosyntactic Tagset*, in "Journal of Language Modeling", Institute of Computer Science PAS, vol. 0, 2012, nr. 1, p. 41–63. (http://korpus.sk/attachments/publications/slovak_morpho_2012.pdf)

McEnery, Xiao 2007 = McEnery, A. M., Xiao, R. Z., *Parallel and comparable corpora: What are they up to?* in "Incorporating Corpora: Translation and the Linguist", Clevedon, Multilingual Matters, 2007.

Šimková, Garabík 2014 = Mária Šimková, Radovan Garabík, *Slovenský národný korpus (2002 – 2012): východiská, ciele a výsledky pre výskum a prax*, in Katarína Gajdošová, Adriána Žáková (eds.), *Jazykovedné štúdie XXXI. Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu)*, Bratislava, VEDA, 2014, p. 35–64. (http://korpus.sk /attachments/snkbooks/jsxxxi.pdf)

Tognini-Bonelli 2001 = Elena Tognini-Bonelli, *Corpus Linguistics at Work*, Amsterdam–Philadelphia, John Benjamins Publishing Company, 2001.

REALIZAREA UNUI CORPUS PARALEL SLOVACO-ROMÂN
(*Rezumat*)

Prezenta lucrare își propune să descrie *Corpusul paralel slovaco-român*, elaborat de Institutul de lingvistică „Ľudovít Štúr" al Academiei Slovace de Științe. Sunt prezentate atât etapele generării acestei colecții de texte aflate, deocamdată, într-o etapă experimentală, cât și exemple ale modului în care poate fi utilizată și aplicabilitatea ei în diverse zone ale cercetării (lingvistică contrastivă, traductologie, pedagogie).

Cuvinte-cheie: *lingvistica corpusurilor, corpus paralel, limba română, limba slovacă.*
Keywords: *corpus linguistics, parallel corpus, Romanian, Slovak.*

The Ľ. Štúr Institute of Linguistics
of the Slovak Academy of Sciences
Bratislava, 26 Panská
Slovakia
richardk@korpus.sk

Romanian Academy
The Institute of Linguistics and Literary
History "Sextil Puşcariu"
Cluj-Napoca, 21 Emil Racoviță str.
andreea_nora_pop@yahoo.com