PARALLEL MULTILINGUAL CORPORA AND THEIR ADVANTAGES FOR TRANSLATORS AND RESEARCHERS

Duruttyová M.
University of Pavol Jozef Šafárik

A growing demand for multilingual and cross-cultural expertise, competence in translation, interpreting and foreign language teaching has been brought by internationalization and the gradual integration of Europe. The significance of accurate and proficient communication across languages has become the focus for linguists and translators along with teachers, as well as for governments, trade and international organizations, education institutions and public authorities.

According to Granger [Altenberg et al 2002], the computer revolution and the possibility of analyzing natural language on the basis of large text corpora has opened up new possibilities of research on the basis of multilingual corpora and experiments in natural language processing, e.g. in the field of machine translation, information retrieval and computational lexicography.

Corpora provide pragmatic information for linguistic theories and practical applications or serve as testing grounds for linguistic and computational models as stated by Granger [Altenberg et al 2002]. Linguistic analyses traditionally focus on a particular linguistic characteristic, which can be a word or a grammatical construction. The use of such features can be further examined by taking into account their associations with other features. By observing and understanding the linguistic associations in particular languages, which can be attained with help of computerized corpora, translators can be trained and translations can be produced more effectively and promptly.

It is noteworthy to point out two important kinds of associations mentioned by Biber et al [1998:6]: linguistic associations and non-linguistic associations.
Investigating the use of a linguistic feature (lexical or grammatical)

1. Linguistic associations:
a. Lexical associations (associations with particular words)
b. Grammatical associations (associations with particular grammatical constructions)
2. Non-linguistic associations
a. Distribution across registers
b. Distribution across dialects
c. Distribution across time periods

For the purposes of this paper, we are focusing on the linguistic associations. According to Biber et al [1998:6], linguistic associations fall into two major categories:

1. Lexical associations – investigating how the linguistic feature is systematically associated with particular words;
2. Grammatical associations – investigating how the linguistic feature is systematically associated with grammatical features in immediate context

Many different kinds of association patterns can be explored with corpus-based studies and it is necessary to point out that these patterns occur to differing extents [Biber et al 1998].

"Almost any area of linguistics can be studied from a use perspective – and the corpus based approach provides a suite of tools and methods that are particularly effective for such investigations. " [Biber et al 1998].

Aligned parallel corpora can be used to extract translation templates, they can be of great help to translators as well as language learners or teachers.

Let us look at an example of the OPUS corpus. OPUS is a growing collection of translated texts from the web [Tiedemann 2009]. In the OPUS project there are converted and aligned online data, with added linguistic annotation with a publicly available parallel corpus that uses standard encoding formats. OPUS is based on open source products and the corpus is also delivered as an open content package. It has been created by Jorg Tiedemann and it is a public collection of parallel corpora.

The following is an example of a result extracted from OPUS corpus in Czech, English, German, Hungarian and Slovak languages for the word "*constitution*".

- *2268:* The union shall respect the equality of Member States before the **constitution** as well as their national identities, inherent in their fundamental structures , political and constitutional, inclusive of regional and local self-government. It shall respect their essential State functions , including ensuring the territorial integrity of the State, maintaining law and order and safeguarding national security.

-->sk: Únia rešpektuje rovnosť členských štátov pred **ústavou**, ako aj ich národnú identitu obsiahnutú v ich základných politických a ústavných systémoch, vrátane regionálnych a miestnych samospráv. Rešpektuje ich základné štátne funkcie, najmä zabezpečovanie územnej celistvosti štátu, udržiavanie verejného poriadku a zabezpečovanie národnej bezpečnosti.

-->hu: (1) Az Unió tiszteletben tartja a tagállamok **Alkotmány** előtti egyenlőségét, valamint nemzeti identitását , amely elválaszthatatlan része azok alapvető politikai és alkotmányos berendezkedésének , ideértve a regionális és helyi önkormányzatokat is . Tiszteletben tartja az alapvető állami funkciókat , köztük az állam területi integritásának biztosítását , a közrend fenntartását és a nemzeti biztonság védelmét.

-->de: (1) Die Union achtet die Gleichheit der Mitgliedstaaten vor der **Verfassung** sowie die nationale Identität der Mitgliedstaaten , die in deren grundlegender politischer und verfassungsrechtlicher Struktur einschließlich der regionalen und kommunalen Selbstverwaltung zum Ausdruck kommt . Sie achtet die grundlegenden Funktionen des Staates, insbesondere die Wahrung der territorialen Unversehrtheit, die Aufrechterhaltung der öffentlichen Ordnung und den Schutz der nationalen Sicherheit.

-->cs: Unie ctí rovnost členských států před **Ústavou** a jejich národní identitu, která spočívá v jejich základních politických a ústavních systémech, včetně místní a regionální samosprávy. Respektuje základní funkce státu , zejména ty, které souvisejí se zajištěním územní celistvosti, udržením veřejného pořádku a ochranou národní bezpečnosti.

Here we can see the highlighted word "*constitution*" in whole sentences of all the observed languages, namely Czech, English, German, Hungarian and Slovak. The texts of the OPUS corpus used for the purposes of this query originate from the European Constitution. The presented result is just one example of how the parallel corpus can be effectively used by translators or in translator trainings in this particular area, namely the European Constitution. The OPUS corpus contains texts in 21 languages, so the possibilities for translators as well as researchers are vast.

Naturally, the demonstrated result can be used by linguists, researchers, lexicographers as well as teachers and learners to observe and discover linguistic associations and patterns in the particular area of interest. The fast response and a variety of possibilities in handling of the computerized corpus is a great advantage to all language researchers.

Literature

1. Altenberg B. & Granger S. Lexis in Contrast: Corpus – Based Approaches. - Johns Bejamins Pub. Co., 2002.
2. Biber D. Conrad S. & Reppen, R. Corpus Linguistics: investigating language structure and use. Cambridge University Press. 1998. pp. 3-10
3. Černý M. Lakota Language Revitalization: Past, Present, and Future Prospects // In Globalisation and Its Impact on Localities. Ostrava: University of Ostrava, 2008, Chapter 26.pp. 55–61.
4. Tiedemann Jörg. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces // In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing. - John Benjamins, Amsterdam/Philadelphia, 2009, volume V, pp. 237-248.

Summary

This short study presents a look at the advantages of parallel multilingual corpora and briefly presents an example from the OPUS corpus. To sum up, computerized multilingual corpora signify a great help in the area of language study as well as translation. The influence of contrastive linguistics has been remarkably dominant in issues concerned with natural language processing, for instance machine translation and computational lexicography. It is a fact that processing, improving and developing computerized corpora represent the future in translation and the study of language.