Bulletin of the *Transilvania* University of Braşov • Vol. 5 (54). No.1 - 2012
Series IV: Philology and Cultural Studies

CORPUS BASED TRANSLATION RESEARCH - CORPORA AND THEIR ADVANTAGES FOR TRANSLATORS AND RESEARCHERS

Michaela DURUTTYOVA[1]

Abstract: *The increasing need for translations in many important areas of our lives calls for developing and improving of comparable corpora as effective and efficient aids to professional as well as non-professional translators. Through the comparative analysis of corpora in the individual areas of texts the main differences can be highlighted between text building strategies and approaches. This can help professional and occasional translators achieve high quality translations in such important fields as research projects, applications, calls for tenders, which are widely used today in the context of the European Union. This article presents a look at the advantages of parallel multilingual corpora and briefly presents an example from the OPUS corpus. 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.*

Keywords*: translation, corpus linguistics, OPUS Corpus.*

1. Introduction

Today's society uses many materials, which are in fact translated from English. The number of English Language speakers is increasing in our country, while the amount of text to be translated doesn't decrease, which is connected to the integration into the European Union, requiring a considerable amount of English language materials and documentation. The laws, bank materials, international companies' documentations, ministries as well as educational and cultural institutions use translated materials and translate their own materials. The administration used to be conducted in one language, but this has changed into an administration in two or more languages, which is the case of calls for proposals, research plans, biographies, recommendations, sanctions, business plans, spending, budgets, records, reports and many other documents of the everyday life.

2. Corpus Linguistics and Translation

Descriptive Translation Studies try to look at translations not as garbled text, a lower status text, but also as research objects. The description of translation features in translated texts received a new momentum in the early '90s, when corpus

[1] University of Pavol Jozef Safarik, Kosice, Slovakia.

BDD-A20140 © 2012 Transilvania University Press
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:51:42 UTC)

linguistics tools began to be used for research of translations.

According to Baker (Baker 1993) there are three kinds of important corpora for the translation studies:

(1) The parallel corpora contain the original text (also marked as source language) and the target language translation. Their greatest advantage is that they show how the experienced translators solve the translation problems. They refer to texts that are translations of each other.
(2) The multilingual corpora have the same selection criteria that include texts, which are not translations of each other. Their advantage is that the elements of target language can be examined in the natural environment, and thus the characteristic lexical and structural patterns can be detected in the target language, which would be different in translations and perhaps unnatural.
(3) Comparable corpora refer to texts in two languages that are similar in content, but are not translations. In order to exploit a parallel text, some kind of text alignment, which identifies equivalent text segments (approximately sentences), is a prerequisite for analysis.

2.1. The benefits of translation corpora

According to Klaudy (Klaudy 1994), the grammar handbooks are not adequate in providing enough assistance to the translators:

(1) the differentiation of the grammar rules of the genre,
(2) the lack of statistical approach, and
(3) the lack of approach on the text level.

The problem of texts in different genres including an original corpus and a reverse corpus can be solved on the basis of three criteria:

(1) the occasional or specialist translators should not keep to codified grammar rules based on literary texts, but to the characteristics of a particular text type,
(2) to distinguish features typically held as translation phenomena such as long-noun structures, the adjectival or adverbial structures, the expansive phrases, and find out whether they occur more often in translations than original texts of a particular language, and
(3) determining what role the above mentioned characteristics play in formation of text coherence.

The corpus based approach would verify certain grammar rules, or reject others. Various translation strategies could be proposed based on this approach.

The advantage of corpus based research is that the wrong approach occurring in some translations of the lexical content or sentence errors of translators can be avoided. The corpus based translation research results from the translation of texts as a whole and tries to bring the relevant findings in this context.

The expansion of international relations and the associated enormous quantity of documentation for translation (or creation of parallel texts) is needed more and more in today's globalised context. It is very difficult for professional translators to perform all these tasks. The correspondence between the languages in the future will not be the sole responsibility of professional linguistic mediators, but everybody's daily routine. To achieve this, awareness of the first language should be increased in a contrastive approach in the context of the Indo-European languages. The translation studies can help in exploration, description, organization of the differences in languages thus achieving a higher level of translations.

Corpus linguistics is very important in the translation studies for the reason that knowing the fact that corpora are collections of texts representing the naturally occurring language, it may help translators to produce natural translations, bearing in mind the real usage of the particular language. Translation studies based on corpora empirically approach language description, it shows the real life occurrences, as well as natural language samples, examines syntactic and lexical as well as text features, which are typical for the particular researched language or the language that is worked with.

It is necessary to mention that a translator not only translates a particular text word by word, but

BDD-A20140 © 2012 Transilvania University Press
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:51:42 UTC)

they have to keep in mind the readership or the audience as well as the cultural aspects of the target language. Corpus use can be an irreplaceable aid in the translation process, which is much more flexible and useful than the traditional dictionaries. It proves to be invaluable in the course of translation into a foreign language in a way that it shows the authentic language patterns, which help to overcome the cultural as well as language boundaries. The great advantage in comparison to the traditional dictionary lies in the fact that a word is found as an isolated unit in a dictionary, whereas this word is surrounded by a context in a corpus giving it a much more complex and communicative meaning.

2.2. The significance of effective communication and computerized language processing

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.

2.3. The advantages of executing corpus based analyses

Krieger (2003) states that corpus linguistics provides a more objective view of language than that of introspection, intuition and anecdotes (Krieger, 2003). A corpus-based analysis can explore almost any language patterns, namely lexical, structural, lexico-grammatical, discourse, phonological, morphological, often with very specific aims, such as discovering male versus female usage of tag questions, children's acquisition of irregular past participles (ibid.). With the proper analytical tools, a researcher can determine not only the patterns of language use, but the scope in which they are used, and numerous contextual features (ibid.).

Numerous terminologists and lexicographers are determined on describing as well as investigating word combinations in individual languages, above all combinations that encompass two lexemes bound to one another, which are linked to principles recognized within a given subject matter where a particular lexeme favours the company of another lexeme. According to L'Homme (2000) a lexeme is a unit with special reference within a specialized subject field and the other lexeme is often referred to as co-occurrent. L'Homme states that lexicographers and other specialists call these combinations collocations (ibid.).

Collocations follow a standard usage and they cannot be explained in terms of regular syntactic or semantic rules. It would be impossible for language users to know whether a certain word combination is correct in a specific area without adequate linguistic knowledge. It is a fact that collocations are typical within a given linguistic community and translators as well as learners of a given language should consider them as part of a particular field (L'Homme, 2000).

BDD-A20140 © 2012 Transilvania University Press
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:51:42 UTC)

2.4. Linguistic and non-linguistic associations

It is noteworthy to point out and consider closely two important kinds of associations mentioned by Biber et al (1998): 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), 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.

3. The OPUS Corpus

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

BDD-A20140 © 2012 Transilvania University Press
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:51:42 UTC)

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.

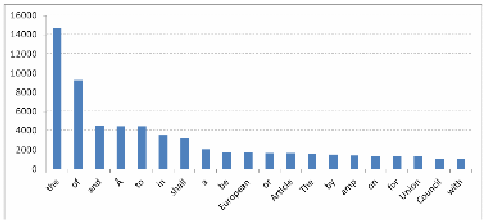
Here one can see the highlighted word "*constitution*" in whole sentences of all the observed languages, namely Czech, English, German, Hungarian and Slovak.

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.

Here the queried words are not just isolated units, but they are surrounded by context, which gives the researchers and investigators the real view of the naturally used language and the associations the particular words might have. The OPUS corpus contains texts in 21 languages, so the possibilities for translators as well as researchers are vast.

Following is a list of the most frequently occurring words in the English Corpus of legal texts based on OPUS.

freq	word
14655	the
9279	of
4545	and
4486	A
4464	to
3570	in
3171	shall
2089	a
1747	be
1727	European
1671	or
1642	Article
1560	The
1500	by
1393	amp
1380	on
1323	for
1308	Union
1089	Council
1070	with



It is clear from this chart which how the most frequent words are distributed within the corpus of English Legal texts.

To investigate lexical associations further, it is uselful to look for collocates associated with a certain word. The researcher can choose e.g. the demonstrative "this" to look for collocates occurring with this particular word within the corpus as well as determine the statistical significance of the found collocates. The investigated collocates and their frequences as well as their statistical significance are limited to the 5 most important ones withhin the corpus or the purposes of this work:

BDD-A20140 © 2012 Transilvania University Press
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:51:42 UTC)

	freq	MI score	T-score
treaty	48	13.63879	6.92764
title	39	13.33923	6.24438
statute	33	13.09822	5.74389
european	31	12.90632	5.56707
explanation	22	12.51326	4.68959

Mi-score – Mutual Information

The Mutual Information Score expresses the extent to which the observed frequency of co-occurrence differs from what is expected. In statistical terms this is a measure of the strength of connection between words x and y. Mi-score is usually calculated based on the number of times one observed the pair together in a given corpus in opposition to the number of times the pair occurred separately. However, Mi-score does not work well with very low frequencies and that is why it is possible to set the lowest limit of frequency in the concordance program and the Mi-score is not calculated for words having the absolute frequency below this limit.

T-score

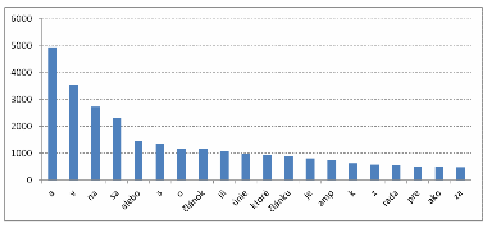
The t-score offers a means to solve this problem as it also takes frequencies into consideration. The t-score is a degree not of the intensity of connection but the certainty with which it is probable that there is an association. Mi-score is more likely to give high scores to completely fixed phrases while t-score produces significant collocates that occur relatively frequently. In most cases, t-score is the most reliable measurement.

T-score in case of collocations tests if the numbers of occurrences of words and their associations are randomly distributed in a corpus. The higher value of t-score gives information about the probability that the word combinations are in fact collocations.

It is useful to look at the collocates of the Slovak translation of the word “this” and its frequency as well as collocates. The investigated corpus is OPUS, which is a corpus of translated Slovak language containing pproximately the same amount of text as the English corpus investigated above.

Following is a list of the most frequently occurring words in the Slovak Corpus of legal texts based on OPUS.

freq	word
4950	a
3567	v
2722	na
2296	sa
1465	alebo
1371	s
1175	o
1165	článok
1070	iii
988	únie
929	ktoré
880	článku
804	je
731	amp
632	k
565	z
558	rada
526	pre
516	ako
490	za



It is clear from this chart which how the most frequent words are distributed within the corpus of English Legal texts.

To investigate lexical associations further, it is uselful to look for collocates associated with a certain word. The researcher can choose e.g. the demonstrative “tento”, which is one of the possible translations of the English “this” into

BDD-A20140 © 2012 Transilvania University Press
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:51:42 UTC)

Slovak, to look for collocates occurring with this particular word within the corpus as well as determine the statistical significance of the found collocates. The investigated collocates and their frequences as well as their statistical significance are limited to the 5 most important ones withhin the corpus or the purposes of this work:

	freq	MI score	T-score
vysvetlivky	17	11.00978	4.12111
komisia	7	10.40775	2.6438
uloží	5	10.24425	2.23422
žalobcu	2	9.92232	1.41276
útvar	2	9.92232	1.41276

The above mentioned approach investigates two of the selected typologically different languages, namely English and Slovak. In this case the investigators have a great opportunity to compare results, significant collocates, compare which words are significant in one sample of language and which ones are significant in another language. It can give a clear idea about how the language works, how it functions and provide more data for understanding.

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.

Acknowledgements

This work was supported by the Slovak Research and Development Agency under the contract No. LPP-0095-09.

References

1. Altenberg, B., & Granger, S. *Lexis in Contrast: Corpus – Based Approaches.* Johns Bejamins Pub. Co., 2002.
2. Baker, M. *Routledge Encyclopedia od Translation Studies*. Routledge, 2001.
3. Bernardini, S., Castagnoli, S. *Corpora for translator education and translation practice. In Yuste Rodrigo, E. (ed.) Topics in Language Resources for Translation and Localisation*. Amsterdam/Philadelphia: John Benjamins, 2008.
4. Biber, D., Conrad, S., & Reppen, R. *Corpus Linguistics: investigating language structure and use.* Cambridge University Press, 1998.
5. Gries, S. T. *What is Corpus Linguistics? Language and Linguistics Compass*, 1225-1241, 2009.
6. Koehn, P. *Europarl: A Parallel Corpus for Statistical Machine Translation*., 2005
7. Krieger, D. Corpus Linguistics: *What It Is and How It Can Be Applied to Teaching*. Retrieved January 19, 2011, from The Internet TESL Journal: http://iteslj.org/Articles/KriegerCorpus.html, (2003, March).
8. Laviosa, S. *Corpus-based Translation Studies*. Amsterdam/New York: Rodopi., 2002.
9. L'Homme, M. C. *Specialized Lexical Combinations: Should they be Described as Collocations or in Terms of Selectional Restrictions*, 2000.
10. Olohan, M. *Introducing Corpora in Translation Studies*. Routledge, 2004.
11. OPUS. (n.d.). Retrieved April 22, 2010, from OPUS an open source parallel corpus: http://opus.lingfil.uu.se/, 2010.
12. Tiedemann, J. (n.d.). EUconst. Retrieved April 22, 2010, from OPUS: http://opus.lingfil.uu.se/, 2010.

BDD-A20140 © 2012 Transilvania University Press
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:51:42 UTC)

13. Tiedemann, J. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*. Retrieved from http://logos.uio.no/opus/, (2009).

14. Toury, G.; Pym, A.; Shlesinger, M.; Simeoni, D. *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury*. Amsterdam/ Philadelphia: John Benjamins, 2008.

BDD-A20140 © 2012 Transilvania University Press
Provided by Diacronia.ro for IP 216.73.217.42 (2026-11-15 19:51:42 UTC)