The Case of InterCorp, a multilingual parallel corpus[1]

František Čermák and Alexandr Rosen

Charles University, Prague

Abstract

This paper introduces *InterCorp*, a parallel corpus including texts in Czech and 27 other languages, available for online searches via a web interface. After discussing some issues and merits of a multilingual resource we argue that it has an important role especially for languages with fewer native speakers, supporting both comparative research and studies of the language from the perspective of other languages. We proceed with an overview of the corpus – the strategy and criteria for including new texts, representation of available languages and text types, linguistic annotation, and a sketch of pre-processing issues. Finally, we present the search interface and suggest some research opportunities.

Keywords:

parallel corpora, Czech, European languages, multilingualism, comparative corpus linguistics

1 Introduction

Since one of the basic tenets of corpus linguistics is the demand for an ever-increasing scope of data, the availability of texts in more languages in a multilingual corpus should

represent a clear advantage. However, the problem is acquiring a balanced mix of multilingual texts in quantities usual for monolingual corpora. In a parallel corpus, the issue of insufficient data, scarce both in size and type, becomes ever more central. For languages unable to benefit from a pool of literary translations (from or into the language), or even from a role in an international context (e.g. as one of the official languages of the EU), this bottleneck may become so prohibitive that it prevents any further growth of the corpus. The problem is compounded once we decide to include more than two languages in a corpus.[2]

Today, (bilingual) parallel corpora exist for many language pairs and the technology needed to build, process and exploit parallel texts in general is widely explored – for an overview see Mihalcea & Simard (2005). Apart from the long-standing focus on sentence and word alignment and statistical machine translation, a number of new fields have emerged:[3] syntactic annotation leading to parallel treebanks (e.g. Bojar et al. 2012), extraction of multilingual lexicons and thesauri (Yang & Luk 2003), cross-lingual information retrieval (Cheng et al. 2004), projection of morphological and syntactic annotation onto another language (Bouma et al. 2008) and word-sense disambiguation (Diab & Resnik 2002). In the latter approaches, translation is viewed as a bridge to carry linguistic knowledge across languages rather than simply a set of links between equivalent sentences or word forms. Significant developments also concern the process of acquiring a parallel corpus from the web (Razavian & Vogel 2009). Probably the most striking example of the use of such texts in machine translation is the Google Translate tool, extracting translation equivalents from multilingual sites.[4] It is now also used in combination with a local database of parallel texts – 'translation memory' – to assist many professional

translators.

Some progress has also been made since the days when "parallel" meant "bilingual", when the only substantial sources were restricted to English and French, as in the *Hansard Corpus* (Roukos et al. 1995), or available for a type of language somewhat distant from contemporary or common use – as in the Bible and in some classical authors. Although the latter do represent a real multilingual corpus, they do not seem to attract much research, possibly because of a less keen interest in diachronic studies and the issue of translations dating from different periods. In any case, the choice of texts seems to be far from balanced in all multilingual parallel corpora of substantial size available today (see e.g. Tiedemann & Nygaard 2004, Erjavec et al. 2005, Koehn 2005, von Waldenfels 2006) – perhaps necessarily so.[5]

The current situation shows that it is difficult to materialise the idea of a balanced multilingual corpus, mainly due to the problematic access to representative data, but also due to the complexity of the task of collecting, processing and managing data in a number of languages. Sentence segmentation, tokenization, alignment and concordancing bring some leverage when applied to a number of languages within a single project, but linguistic annotation requires language-specific tools or tagged data. Yet there are ways to achieve at least some of these tasks, even before an appropriate set of tools covering more than a few languages is available.

Despite the problems in building a balanced multilingual parallel corpus, comparison and research of more languages in a single source, including its exploitation by rule-based or stochastic tools, is a goal which should be rather self-evident in today's multilingual Europe, where traditional monolingual and bilingual settings are increasingly giving way to

a multilingual reality (Aronin & Hufeisen 2009). It remains to be seen to what extent concerted corpus-based multingual research can be substituted by a series of bilingual comparisons (see e.g. Johansson 2007), but preliminary observations suggest that a multilingual source of data may be useful at least for tasks such as studying various types of phenomena in a given language from the perspective of comparable phenomena in other languages, ranging from lexical and collocability issues to pragmatics (see e.g. Čermáková & Fárová 2010 or Káňa 2011). From this general perspective, the old dictum saying that language is an instrument of transmission of meaning from thought to form will be complemented by an additional one, namely that languages (if investigated comparatively) are also bridges enabling mutual transfer of meaning. Although many languages create many divides, they help to bridge them more easily at the same time.

Linguistic terms used in this field in the past decades have various connotations, but the best term to use here still seems to be the 'comparison' of languages, as the term 'contrastive' linguistics suggests that the discipline is selective in character, bent on looking for contrasts only (i.e. primarily avoiding statements about agreement). In a sense the notion of exclusively contrastive corpus-based study would go against the all-embracing approach of corpus linguistics. Similarity is much harder to perceive, measure and study than obvious differences. Likewise, the term 'confrontational' linguistics, common in the former Russian and Soviet linguistic tradition, does not seem eligible any more. In this sense, the emerging field of a new kind of 'comparative corpus linguistics' may be given a substantial boost if multilingual corpora are built more extensively and with some concern about representativeness, and researched systematically with the multilingual perspective in mind. The obvious desideratum behind this is to be sure of one's *tertium comparationis* and

to use a broader framework, preferably a typological one. Last but not least, the practical fields of multilingual translation and localization may profit from similarly multilingual resources.

With very few exceptions, the attention paid to bilingual parallel corpora is oriented towards pairs made up of two extensively used languages (such as English and French in the *Hansard Corpus*), or towards pairs where at least one is such a language. On the other hand, a pair of two "small" languages are in a less privileged position. After all, the majority of languages are "small", whatever that might mean. To remedy this situation, comparative studies should be based on as much data from as many relevant languages as possible.

Both bilingual and multilingual parallel corpora are based on translations between languages for which data are available (not necessarily in electronic form, scanning a printed copy is often necessary). In a sense, the sum of available translations from one language into another represents the sum of strands of accumulated interest of one culture in another through its texts. The interest may be historically conditioned (such as the interest in the 'fashionable novel' of the early 19th century) or general and long-lasting over a well-defined period of time. This fluctuating influence of external factors is particularly significant when comparing the sum of what has been translated between two languages with a smaller number of speakers.

Czech, a Slavic language spoken by some 10 million people, can be viewed as a small language. Being typologically inflectional, it has features less prevalent in the "big" European languages, such as rich inflection (cases, personal endings), verbal aspect, free word order, rich verbal prefixation, rich nominal derivation, a high number of particles, etc.

Based in the middle of Europe, Czech has been historically at a crossroads due to the numerous linguistic contacts with its neighbours, one Slavic (Slovak and Polish), one German (Austrian and German German), both of them representing a different type of research challenge. Studies of the long-lasting impact of German might bring more interesting results if one goes deeper, beyond mere loan-words, namely into semantics, calques or influences on the grammar system. The impact of Polish is less pervasive, while that of non-neighbouring Russian lasted only a few decades. Studying contacts with Slovak is insteresting also due to the process of the blurring of differences between two closely related languages.

All these factors have left their traces, and the result is worth researching, in general and from the typological point of view. The language should not be studied only from the viewpoint of native speakers of Czech, but also from other angles, as seen from the perspective of, and in comparison with other languages. Hence the idea of a large multilingual corpus with Czech at its hub. This concept reflects both the traditional geographical and historical contacts and its openness to the influence of a number of non-adjacent languages, including those playing a global role, such as English.

2 *InterCorp*: Goal and strategy

Both theoretical and practical reasons stand behind the idea of a large multilingual corpus with Czech at the centre. *InterCorp* is currently a part of the *Czech National Corpus* (CNC) project. [6,7] The idea at the heart of *InterCorp* is linguistically trivial, yet not very often voiced; having one's own language amply covered by monolingual corpora may not be enough – the language must also be studied from the outside. The project is unique also in its scope, the choice of texts (with a focus on fiction) and a substantial share of manual

work (with a higher quality of alignment, sentence boundary recognition and fewer typos as a result). The project participants, invited in 2005 to join the team headed by the CNC Institute, come from linguistic departments of the Faculty of Arts at Charles University in Prague and a few other institutions. The current number of "active" languages is 30 (plus Czech), with Czech always the other language in a pair. For the time being, 27 languages are available for online searches using a parallel concordancer (free to use after registration as a CNC user).[8,9]

In addition to online queries sent to a web-based concordancer, users may also be granted offline access to sets of parallel bilingual concordances. Each set is extracted from a specific texts and includes only 1:1 alignment pairs in blocks up to 100 words, with the blocks shuffled in a random order. This measure, complementing the terms of a licensing agreement (such as no re-distribution), makes the use of texts in violation of copyright technically impossible. The effect is the same as in results produced by the concordancer – only quotations in a restricted context are available, never a copy of a larger piece of text.

Table 1 gives counts (in thousands of words) for the languages available in the present release of the corpus.[10] The **core** part of the corpus includes texts selected and acquired by expert team members responsible for the given language, following the criteria outlined below. The number of such texts – mostly novels – is given in the Titles column. For some recently added languages there are no texts in the core part, only **collections** of texts acquired from freely available sources and processed nearly without any manual intervention, such as political commentaries published by Project Syndicate, news selected from European press for publication online by *Presseurop*, and the *Acquis Communautaire,* consisting of the EU legislation adopted by the new member states.[11] Due to the large and

potentially misleading share of the *Acquis* part the Subtotal column shows numbers of words for all the other texts. The addition of such large volumes significantly topples the balance for some languages, which are left with a single text type as a result. However, corpus users can search in a subcorpus of their choice, besides having the benefit of a uniform search interface and – at least for some languages – morphosyntactic annotation, unavailable in the original sources.

Table 1.* Size of the corpus according to languages and text types in *InterCorp* available online in release 5

Language	Core	Titles	*Syndicate*	*PressEU*	Subtotal	*Acquis*	Total
Belarusian	68	2	0	0	68	0	68
Bulgarian	1,415	20	0	0	1,415	13,816	15,231
Croatian	8,104	117	0	0	8,104	0	8,104
Danish	190	5	0	0	190	21,680	21,870
Dutch	6,486	79	0	900	7,386	24,746	32,132
English	5,914	72	2,568	799	9,282	24,208	33,489
Estonian	0	0	0	0	0	15,963	15,963
Finnish	2,082	35	0	0	2,082	16,667	18,749
French	3,218	51	2,970	875	7,062	27,352	34,414
German	12,091	170	2,567	753	15,411	21,724	37,135
Greek	0	0	0	0	0	25,070	25,070
Hungarian	1,123	17	0	0	1,123	19,168	20,290
Italian	3,484	34	80	793	4,358	24,850	29,207
Lithuanian	353	17	0	0	353	18,433	18,785
Latvian	1,085	33	0	0	1,085	18,745	19,830
Macedonian	32	1	0	0	32	0	32
Maltese	0	0	0	0	0	14,133	14,133
Norwegian	2,301	22	0	0	2,301	0	2,301
Polish	8,397	124	0	711	9,108	20,464	29,572
Portuguese	2,127	25	0	914	3,041	28,599	31,640
Romanian	1,370	16	0	820	2,190	8,200	10,389
Russian	1,665	30	2,305	0	3,969	0	3,969
Slovak	7,258	139	0	0	7,258	19,222	26,479
Slovene	991	16	0	0	991	19,646	20,637
Serbian	4,295	46	0	0	4,295	0	4,295
Spanish	11,811	141	2,897	861	15,569	27,001	42,570
Swedish	5,888	70	0	0	5,888	20,615	26,503
Total	91,748	1,282	13,386	7,425	112,559	430,300	542,860
Czech	52,896	864	1,574	704	75,460	20,285	97,273

*All figures give the number of word tokens in thousands, except for Titles, showing the number of texts in the core part. Some figures do not add up due to rounding.

Each of the language pairs is different, both in size and content, and the original assumption that there might be a non-trivial common core of titles shared by most if not all languages has not turned out to be true so far – most of the titles are only available as bitexts. Currently the titles available in the highest number of languages are Milan Kundera's novels *The Unbearable Lightness of Being* and *The Joke* (both online in 9 languages including Czech, each with 7 more translations waiting in the pipeline). Three novels of J. K. Rowling's *Harry Potter* series and G. Orwell's *1984* come next, available in 14 and 13 languages respectively, followed by J. R. R. Tolkien's *The Lord of the Rings I*, Kundera's *Immortality*, and J. Hašek's *The Good Soldier Švejk* in 12 languages (not all of them online at the moment). On the other hand, the collections outside the core of fiction make the score slightly more balanced, especially for languages of the European Union.

The general policy and goals behind the acquisition of texts are quite straightforward:

1. We aim for *contemporary* texts, i.e. for those dating no further back than 1945. This time line is set deliberately: except for classical literature, the newer texts are a more representative approximation of language use due to the marginal share of older texts in the input of an average speaker.
2. To make up for the lack of titles shared by more languages, we also admit texts whose original language is not Czech or the other language in the pair. Thus, 6 out of 15 titles in the Czech-Serbian subcorpus currently available online are translations from a *third language*, mostly English, but also Italian, Polish, Portuguese and Russian. The strategy is to have titles with a wide array of translations into other

languages. Titles translated into more languages are actually preferred – based on available bilingual translations, project participants in charge of the individual languages receive a list of titles to consider as candidate additions. The presence of non-original titles on both sides can distort some kinds of analysis while for other purposes it may not be important. Techniques evaluating the relevance of indirect translations from a third language, in comparison to direct equivalents, have yet to be found.

3. *InterCorp* strives to be linguistically *general* so that it might be used for many different purposes. Hence, it is desirable to capture types of language and vocabulary that are as diverse as possible. However, a balanced parallel corpus is much harder to build than a monolingual one: (i) Some types of text and most types of speech are hardly ever translated, including some types of newspaper language. This is the main reason why *InterCorp* consists entirely of *written* texts. (ii) As for *non-fiction* and its most prevalent genre – the language of the press, we have already tapped several sources (see above). The inclusion of the more specific language of parliamentary debates (*Europarl*) is under way, and we also consider various open-source technical and software manuals (as in *OPUS – an open source parallel corpus*), possibly including the very different genre of film subtitles.[12,13] The choice of these texts is largely pragmatic, depending on their availability. In any case, corpus users are free to select a set of texts to be searched and exploited according to their needs and preferences.

Due to this pragmatic approach to the corpus build-up it is difficult to plan the final shape of the corpus to any high degree; it is changing with every new release. Moreover, although

it is an obvious desideratum, it is virtually impossible to achieve any kind of balance between the number of titles translated into Czech and from Czech, and the idea has not been made a criterion (so far).

3 Accessing the data through a search interface

Each language has a coordinator in charge of text acquisition, conversion into a standard electronic format (in case it is needed), text cleanup and proofreading.[14,15] After completing these offline tasks the text is uploaded for formatting checks and automatic detection of sentence boundaries.[16] A pair of sentence-segmented files is then aligned by *hunalign*, one of the best-performing aligners available (Yu et al. 2012). The result is manually checked and corrected in *InterText*, a web-based parallel text editor. [17,18]

In a next step, the aligned texts are exported from *InterText* in a stand-off alignment format (with the alignment links stored in a separate file). Finally, the texts can be morphologically tagged and lemmatized. This option depends on the availability and performance of suitable language-specific tools. The 17 languages currently tagged are listed in Table 2 together with the respective tools.[19] The list is due to be further extended in future releases of the corpus, especially for languages with rich inflection, where the benefit of morphological tagging for users is highest, despite the possibly challenging differences in language-specific tagsets.[20]

Table 2. Morphological tagging and lemmatization

Language	**Tags**	**Lemmas**	**Tool**
Bulgarian	➽		*TreeTagger*
Czech	➽	➽	*Morče*

Dutch	➳		*TreeTagger*
English	➳	➳	*TreeTagger*
Estonian	➳	➳	*TreeTagger*
French	➳	➳	*TreeTagger*
German	➳	➳	*TreeTagger*
Hungarian	➳		*HunPos*
Italian	➳	➳	*TreeTagger*
Lithuanian	➳	➳	V. Daudaravičius
Norwegian	➳	➳	*Oslo-Bergen Tagger*
Polish	➳	➳	*Morfeusz, TaKIPI*
Portuguese	➳	➳	*TreeTagger*
Russian	➳	➳	*TreeTagger*
Slovak	➳	➳	*Morče*
Slovene	➳	➳	*totale*
Spanish	➳	➳	*TreeTagger*

Finally, the texts are matched with bibliographical data from the project database and indexed by the corpus manager (*Manatee*, see Rychlý 2007) to be used with a parallel web-based search interface called *Park* (built by Michal Štourač, cf. note 8). The currently available set of search and display functions of *Park* includes:

- Restrictions on the search scope by languages and texts. Titles in the core part of the corpus (with segmentation and alignment manually checked) are included by default, unlike the collections (processed in a fully automatic way). The default can be overridden by selecting or deselecting titles and collections available for the given combination of languages. The search scope can be further restricted by using other bibliographical data for a specified language, such as the year of publication, the text type (fiction, poetry, drama, legal, etc.), the gender of the author or translator, the source language, or whether the text is the original or a translation.

- Queries into one or more languages by word form, by a string of word forms (a phrase), by a CQL (Corpus Query Language)[21] expression (including regular expressions), for some languages by lemma (base form) and/or morphosyntactic tag; with a virtual keyboard to type in foreign characters; with an option to recall a previous query.
- Displaying parallel concordances side by side or in rows as KWiC or segments; displaying more context; displaying structural tags (paragraphs, sentences, segments), bibliographical data and concordance ID, lemma and/or morphosyntactic tag for the keyword or all displayed words (for some languages); export of concordances as a spreadsheet file; displaying a specified number of randomly sampled lines; filtering the set of concordances by positive or negative restrictions on the keyword or the specified context.

Figure 1 is a screenshot of the search interface after specifying Czech and Russian as the languages to be queried. The list of available titles and text collections shrinks, depending on the choice of languages. After English is selected as an additional language, the list narrows down to five novels, including George Orwell's *1984* and Milan Kundera's *The Unbearable Lightness of Being* and *The Joke* (not shown).

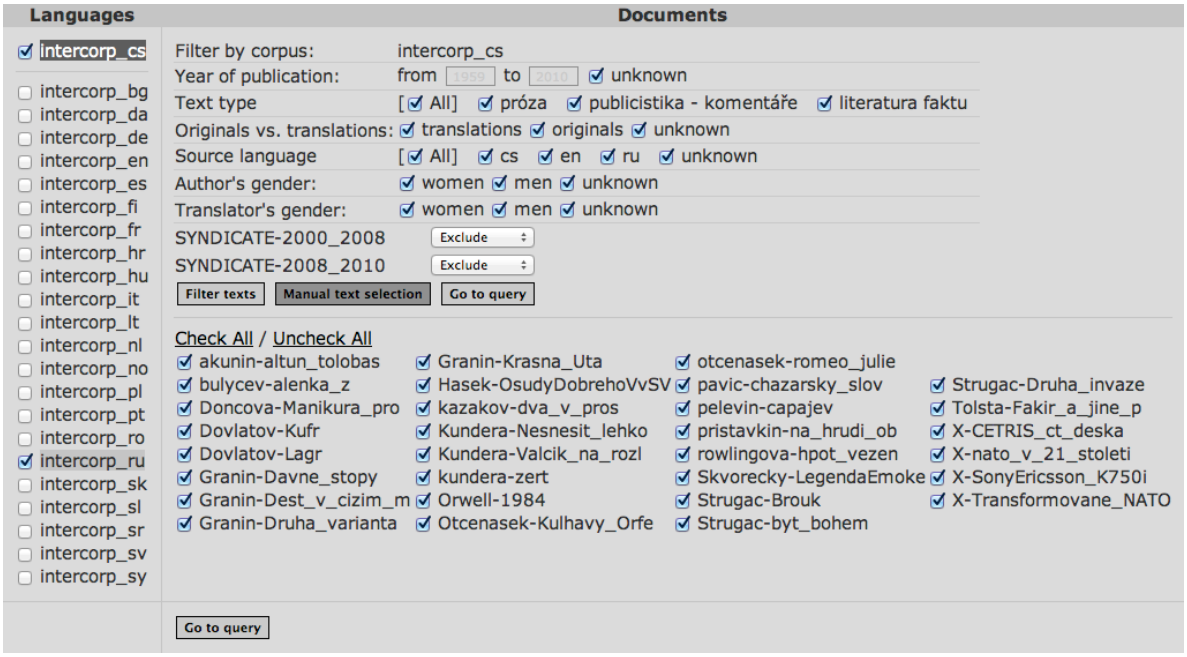


Figure 1. Specifying languages and titles in the search interface *Park*

A query can be specified for any language or any combination of languages. Figure 2 shows a CQL query into the Czech portion. We are looking for negated forms of the verb *věřit* "believe" (the query specifies the lemma and a morphosyntactic tag, using the Corpus Query Language with regular expressions:[22] [lemma="věřit" & tag="V…….N.*"]).

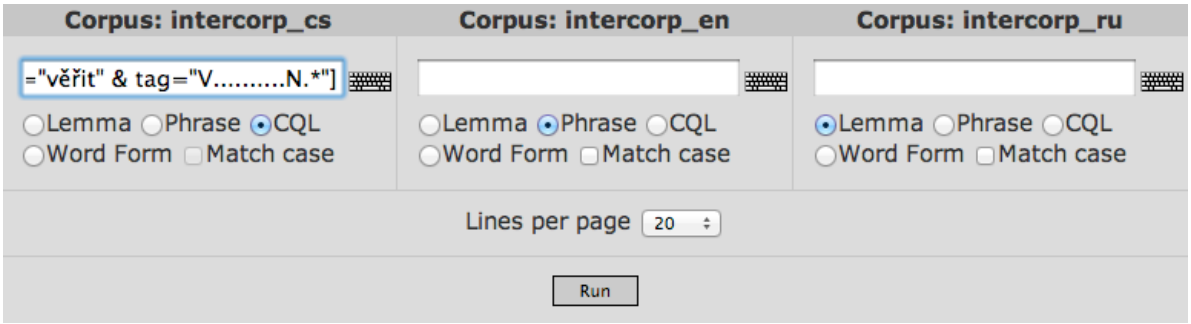
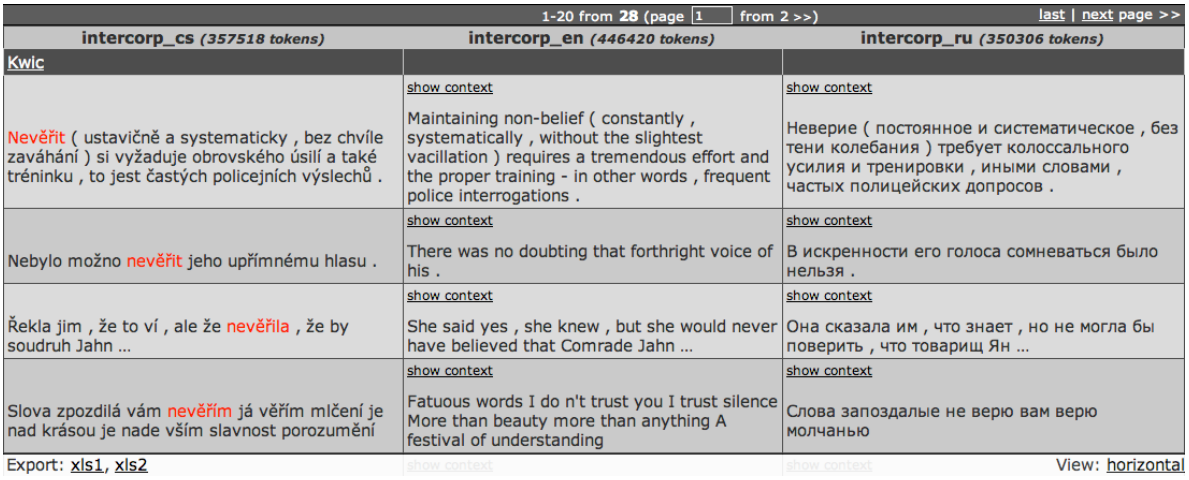


Figure 2. Specifying a query for the negated verb "to believe" in Czech

The first four hits are shown in Figure 3. The number of tokens in the column headings

refers to the total number of word tokens present in the texts selected for the query. In the Czech column, the keyword (the expression specified in the query) is highlighted. For the other languages, no expression corresponding to the keyword can be identified when it is not specified in the query.



1-20 from **28** (page 1 from 2 >>) last | next page >>

intercorp_cs *(357518 tokens)*	intercorp_en *(446420 tokens)*	intercorp_ru *(350306 tokens)*
Kwic		
Nevěřit (ustavičně a systematicky , bez chvíle zaváhání) si vyžaduje obrovského úsilí a také tréninku , to jest častých policejních výslechů .	show context Maintaining non-belief (constantly , systematically , without the slightest vacillation) requires a tremendous effort and the proper training - in other words , frequent police interrogations .	show context Неверие (постоянное и систематическое , без тени колебания) требует колоссального усилия и тренировки , иными словами , частых полицейских допросов .
Nebylo možno nevěřit jeho upřímnému hlasu .	show context There was no doubting that forthright voice of his .	show context В искренности его голоса сомневаться было нельзя .
Řekla jim , že to ví , ale že nevěřila , že by soudruh Jahn ...	show context She said yes , she knew , but she would never have believed that Comrade Jahn ...	show context Она сказала им , что знает , но не могла бы поверить , что товарищ Ян ...
Slova zpozdilá vám nevěřím já věřím mlčení je nad krásou je nade vším slavnost porozumění	show context Fatuous words I do n't trust you I trust silence More than beauty more than anything A festival of understanding	show context Слова запоздалые не верю вам верю молчанью

Export: xls1, xls2 View: horizontal

Figure 3. Results of the query for the negated verb “to believe” in Czech with parallel English and Russian sentences

The result can be displayed with different types of linguistic or metatextual information. Figure 4 shows some bibliographical data (the author’s surname and name, the title of the novel and the publication year) and morphosyntactic annotation for each word (lemma and tag).

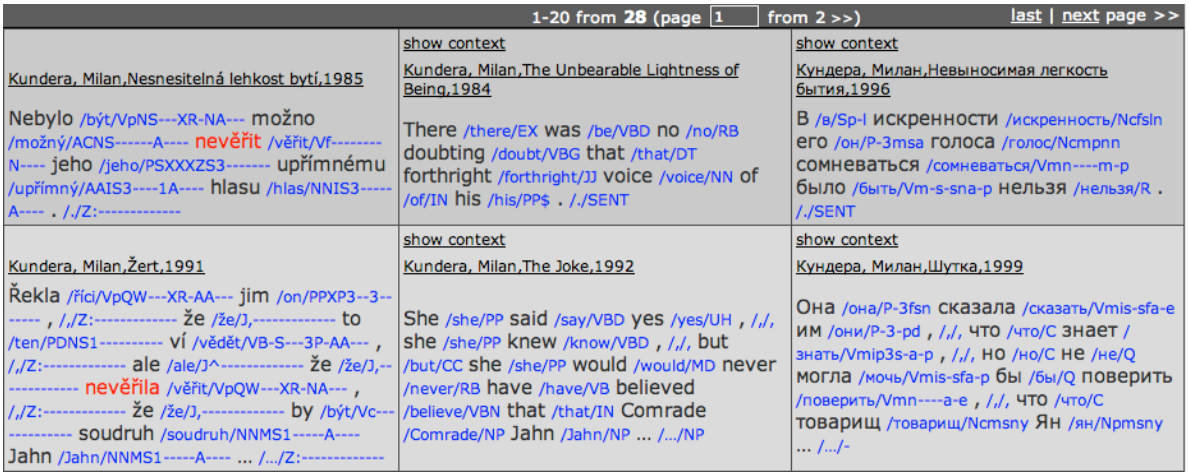


Figure 4. Results of the query with some bibliographical information, lemmas and tags

All *InterCorp* texts available in *Park* can be queried also as a set of monolingual corpora through *NoSketch Engine*, the web-based interface used for the monolingual parts of the Czech National Corpus, offering a more extensive choice of features (sorting, collocations, frequency distribution).[23]

4 Research opportunities with InterCorp and future developments

Results achieved so far (Čermák et al. 2010, Čermák & Kocek 2010, Čermák 2011) support our belief that the corpus is useful in a number of ways. We try to make sure that the way it is implemented supports this practical and open-minded aspect, steering clear from directions restricting it to a mere experiment or academic exercise, an obvious requirement for a project of this size and scope.

Two major lines of research into a multilingual corpus suggest themselves: applied and theoretical (as laid out, for example, in Botley et al. 2000). The former will depend on actual demand and might be related to the traditional fields of translation studies and lexicography (Teubert 2001, 2007; Johansson 2007).

Truly multilingual lexicography does not seem to be very popular at the moment,[24] but that might change. It could be useful, even for people knowing these languages, to have a *dictionary of closely related languages* such as Czech, Polish and Slovak, and similarly for Scandinavian or Romance languages, even for checking only or avoidance of false friends.

The corpus should also be open to a more recent but very active research agenda, extremely voracious for multilingual text data: *machine translation, text-mining, word-sense disambiguation* and other machine learning techniques relying on meaning equivalences implicit in parallel texts to produce both multi- and monolingual applications. The presence of text types rare in other sources of data makes *InterCorp* an attractive partner.

The latter, theoretical line in advanced multilingual comparison, may also open some new vistas hitherto unexplored because of lack of data. The availability of a multilingual corpus means an important new stimulus for *comparative corpus linguistics*. Specifically, it points to general linguistics, typology, pragmatics and discourse studies or at least becomes a challenge for them. However, another basic question will have to be eventually answered, having an uncomfortable implication. While the strong point of any monolingual corpus research has always been in the study of authentic texts and real contexts, bilingual and multilingual corpora are different in that translations are not original, authentic texts (and, for that matter, nor are the contexts that are translated). A methodology will have to be found to evaluate translated counterparts.

Going bottom-up, from lexical items, through collocations and phrases to sentences and their combinations, the value of such divisions and categories, assumed by traditional methods so far, must inevitably become more problematic and prone to various

interpretations. Yet, given meaning, which should ideally be taken as the starting point, it seems that a solution must be sought at the higher levels rather than at the lower ones, such as words. Having a parallel corpus or corpora offering profuse contexts and a variety of equivalents of an item on a scale that can be statistically evaluated means much more than the old-time manual contrastive study based on odd and isolated examples only. Next to this, it seems evident that by using multilingual parallel corpora a lot can be learned about ways of identifying syntactic chunks and collocations. One just does not know to what extent and in what respect collocations in one language correspond to equivalents in other languages.

Linguistically, a number of general issues may be raised in this framework, centered around one language. For example, too general statements about relatedness of languages, both within smaller and larger groups, deserve a more precise formulation. Research into the seemingly endless diversity of non-related languages, covered so far by typology and universals only, would be an open-ended venture where inspiration can be drawn from the data and *typology of the differences*. From more issues of this kind, at least one familiar field can be brought to attention here, namely *internationalisms*, whose research is so much needed,

The present form and content of the corpus data, together with the search infrastructure, are not yet in their last stage of progress. Especially the corpus search has reached its limits in response time and the set of features available in the user interface. The current technical restrictions should be removed with the planned extension of support for parallel corpora in the *Manatee* corpus manager. The parallel user interface could then match the functionalities available in its monolingual counterpart and offer more features relevant for

parallel data.

The size of the corpus will grow further, with the languages and genres that are currently lagging behind catching up at least to some extent, and with more external resources of quality data plugged in. Obviously, the familiar dictum "all languages are created equal" cannot be reflected in size, but we hope to provide a comparable level of linguistic annotation for all languages in the corpus. To assist the concordancer in producing a parallel KWiC formatting of query results, the corpus will receive word-to-word alignment. Finally, to close the gap between words and sentences and to provide parallel structural alignment for collocations and phrases, chunking or some other type of structural annotation is another, more distant perspective.

To conclude, there is some evidence that systematic comparison of texts in more than one language offers inspiration and qualified knowledge unavailable from monolingual resources. Yet in a way, this is a new and refined version of the feeling one started to have when looking systematically into monolingual corpora without any prejudice.

References

Aronin, L. & Hufeisen, B. 2009. *The Exploration of Multilingualism: Development of Research on L3, Multilingualism and Multiple Language Acquisition*. Amsterdam/Philadelphia: John Benjamins.

Botley, S. P., McEnery, A. M., & Wilson, A. (Eds.). 2000. *Multilingual Corpora in Teaching and Research*, volume 22 of *Language and Computers – Studies in Practical Linguistics*. Amsterdam: Rodopi.

Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M., & Tamchyna, A. 2012. "The Joy of Parallelism with CzEng 1.0". In Calzolari, N. et al (Eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul: European Language Resources Association (ELRA), 3921–3928.

Bouma, G., Kuhn, J., Schrader, B., & Spreyer, K. 2008. "Parallel LFG Grammars on Parallel Corpora: A Base for Practical Triangulation". In M. Butt and T. H. King (Eds.), *Proceedings of the LFG08 Conference*, Sydney, Australia. Stanford: CSLI Publications, 169–189.

Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., & Chien, L.-F. 2004. "Translating unknown queries with web corpora for cross-language information retrieval". In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, New York: ACM, 146–153.

Čermák, F. & Kocek, J. (Eds.). 2010. *Mnohojazyčný korpus InterCorp: Možnosti studia* [The *InterCorp* multilingual corpus: Research possibilities]. Praha: Nakladatelství Lidové noviny.

Čermák, F., Klégr, A., & Corness, P. (Eds.). 2010. *InterCorp: Exploring a Multilingual Corpus*. Praha: Nakladatelství Lidové noviny.

Čermáková, A. & Fárová, L. 2010. "Keywords in Harry Potter and their Czech and Finnish translation equivalents." In F. Čermák, A. Klégr & P. Corness (Eds.), *InterCorp: Exploring a Multilingual Corpus*. Praha: Nakladatelství Lidové noviny, 177–188.

Diab, M. & Resnik, P. 2002. "An unsupervised method for word sense tagging using parallel corpora". In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Morristown, NJ: Association for Computational Linguistics, 255–262.

Erjavec, T., Ignat, C., Pouliquen, B., & Steinberger, R. 2005. "Massive multilingual corpus compilation; Acquis Communautaire and Totale". Paper presented at the *2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (L&T'05), Poznań, 21–23 April 2005.*

Johansson, S. 2007. *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies*. Amsterdam/Philadelphia: John Benjamins.

Káňa, T. 2011. "Deminutiva a deminutivní vyjádření v češtině, němčině a angličtině – hledání hranic [Diminutives and diminutival expressions in Czech, German and English – searching for boundaries]". In F. Čermák (Ed.), *Korpusová lingvistika Praha 2011: 1 – InterCorp*, volume 14 of Studie z korpusové lingvistiky. Praha: Nakladatelství Lidové noviny, 168–185.

Kiss, T. & Strunk, J. 2006. "Unsupervised multilingual sentence boundary detection". *Computational Linguistics*, 32 (4), 485–525.

Koehn, P. 2005. "Europarl: A parallel corpus for statistical machine translation". In *Conference Proceedings: the tenth Machine Translation Summit*. Phuket: Asia-Pacific Association for Machine Translation (AAMT), 79–86, available at http://mt-archive.info/MTS-2005-Koehn.pdf

Macken, L., De Clercq, O., & Paulussen, H. 2011. “Dutch parallel corpus: A balanced copyright-cleared parallel corpus”. *Meta*, 56 (2), 374–390.

Mihalcea, R. & Simard, M. 2005. “Parallel texts”. *Natural Language Engineering*, 11 (3), 239–246.

Razavian, N. S. & Vogel, S. 2009. “The web as a platform to build machine translation resources”. In *IWIC ’09: Proceedings of the 2009 International Workshop on Intercultural Collaboration*. New York: ACM, 41–50.

Rosen, A. 2010. “Mediating between incompatible tagsets”. In Ahrenberg et al (Eds.) *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, volume 10 of NEALT Proceedings Series. Tartu: Northern European Association for Language Technology, 53–62.

Rosen, A. & Vavřín, M. 2012. “Building a multilingual parallel corpus for human users”. In Calzolari, N. et al (Eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul: European Language Resources Association (ELRA), 2447–2452.

Roukos, S., Graff, D., & Melamed, D. 1995. *Hansard French/English*. Philadelphia: Linguistic Data Consortium.

Rychlý, P. 2007. Manatee/Bonito – a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno: Masarykova univerzita, 65–70.

Singh, A. K. & Husain, S. 2005. “Comparison, selection and use of sentence alignment algorithms for new language pairs”. In *Proceedings of the ACL Workshop on Building and*

Using Parallel Texts. Ann Arbor, Michigan: Association for Computational Linguistics, 99–106.

Teubert, W. 2001. "Corpus linguistics and lexicography". *International Journal of Corpus Linguistics*, 6 (SI), 125–153.

Teubert, W. (Ed.). 2007. *Text Corpora and Multilingual Lexicography*. Amsterdam/Philadelphia: John Benjamins.

Tiedemann, J. & Nygaard, L. 2004. "The OPUS corpus – parallel & free". In M. T. Lino et al (Eds.) *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon: European Language Resources Association (ELRA), 1183–1186.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., & Nagy, V. 2005. "Parallel corpora for medium density languages". In *Proceedings of RANLP 2005*, 590–596, available at http://www.ldc.upenn.edu/Catalog/docs/LDC2008T01/ranlp05.pdf.

von Waldenfels, R. 2006. "Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment". In B. Brehmer, V. Zdanova, and R. Zimny (Eds.), *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)*, volume 9, München: Verlag Otto Sagner, 123–138.

Yang, C. C. & Luk, J. 2003. "Automatic generation of English/Chinese thesaurus based on a parallel corpus in laws." *Journal of the American Society for Information Science and Technology*, 54 (7), 671–682.

Yu, Q., Max, A., & Yvon, F. 2012. "Revisiting sentence alignment algorithms for align-

ment visualization and evaluation." In R. Rapp et al (Eds.) *Proceedings of the 5th Workshop on Building and Using Comparable Corpora*, Istanbul: European Language Resources Association (ELRA), 10–16.

Authors' addresses

František Čermák
Institute of the Czech National Corpus
Charles University, Prague
ÚČNK FF UK
nám. Jana Palacha 2
116 38 Praha 1
Czech Republic

frantisek.cermak@ff.cuni.cz

Alexandr Rosen
Institute of Theoretical and
Computational Linguistics
Charles University, Prague
ÚTKL FF UK
Celetná 13
110 00 Praha 1
Czech Republic

alexandr.rosen@ff.cuni.cz

---

1 *InterCorp* is part of the *Czech National Corpus*, a project supported by the Czech Ministry of Education within the programme "Large Infrastructures for Science,

2 For some tasks and purposes, comparable rather than parallel texts can be a way out, but in a project like ours, where users expect parallel concordances of translation equivalents, it remains to be seen to what extent a comparable section of the corpus would be useful.

3 For word alignment and statistical machine translation see e.g. Singh & Husain (2005) and "Statistical Machine Translation" – a site dedicated to SMT research: http://www.statmt.org/.

4 *Google Translate* – an on-line translation tool: http://translate.google.com/.

5 A balanced multilingual corpus of a smaller size may be built for some languages – for an example see http://www.kuleuven-kulak.be/DPC (Macken et al. 2011). At the opposite end of the spectrum are very large resources compiled from public domain texts, often presented as translation memories, such as *MyMemory*: http://mymemory.translated.net/, *WeBitext*: http://www.webitext.com/, or *Glosbe*: http://glosbe.com/.

6 *InterCorp* – the project home page: http://korpus.cz/intercorp/?lang=en.

7 *Czech National Corpus* – the project home page: http://www.korpus.cz/english/.

8 The parallel concordancer is *Park* – the web-based search interface of *InterCorp*: http://korpus.cz/Park.

9 The user registration site for the *Czech National Corpus* is available at:

http://korpus.cz/english/dohody.php.

[10] A Czech text is counted only once, even though it occurs in more than one language pair. More texts and languages (Arabic, Chinese, Hindi, Romani, Ukrainian) are in the pipeline, waiting for the next release.

[11] *Project Syndicate* – a site offering political commentaries in 11 languages: http://www.project-syndicate.org/; *Presseurop* – a portal monitoring European daily newspapers in 10 languages: http://www.presseurop.eu/en; *The JRC-Acquis Multilingual Parallel Corpus*: http://langtech.jrc.ec.europa.eu/JRC-Acquis.html. The currently available texts will be followed by more recent additions in future releases, also in languages such as Arabic and Chinese (in *Project Syndicate*) or Croatian (in *Acquis Communautaire*).

[12] *European Parliament Proceedings Parallel Corpus 1996–2011*: http://www.statmt.org/europarl/; *European Parliament* – proceedings search site: http://www.europarl.europa.eu/.

[13] *Opus – the open parallel corpus*: http://opus.lingfil.uu.se/.

[14] Collections of texts, acquired online for several languages by the CNC team, bypass the standard semi-automatic procedure and receive a fully automatic alignment.

[15] For more technical detail about preprocessing and the project infrastructure see Rosen & Vavřín (2012).

[16] For Czech, we use a rule-based splitter by Pavel Květoň, for other languages *Punkt*, based on an unsupervised learning algorithm by Kiss & Strunk (2006), in an implementation available from *Natural Language Toolkit*: http://nltk.org/.

[17] For *hunalign* see Varga et al. (2005). The aligner is available under GNU LGPL version 2.1 or later from http://mokk.bme.hu/resources/hunalign.

[18] *InterText* – a parallel text alignment editor, created by Pavel Vonřička. The editor is available as a server-based or standalone application under GNU GPL version 3 from http://wanthalf.saga.cz/intertext.

[19] *TreeTagger*: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/; *Morče*: http://ufal.mff.cuni.cz/morce/, morphological analysis and training for Slovak by Radovan Garabík; *HunPos* for Hungarian: http://code.google.com/p/hunpos/; *Morfeusz* and *TaKIPI* for Polish: http://nlp.pwr.wroc.pl/takipi/; the tagger for Lithuanian: http://donelaitis.vdu.lt/~vidas/; *totale* for Slovene: http://nl2.ijs.si/analyze/; *OBT* for Norwegian: http://tekstlab.uio.no/obt-ny/.

[20] See Rosen (2010) for an attempt to harmonise existing tagsets by mapping language-specific tags onto an interlingual hierarchy of linguistic categories.

[21] See e.g. http://trac.sketchengine.co.uk/wiki/SkE/CorpusQuerying.

[22] See http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html for a brief description of the positional tagset used for Czech.

[23] The monolingual search interface of *InterCorp*: http://korpus.cz/corpora/intercorp/. For the tool *NoSketch Engine*, available under GNU GPL version 2, see

http://nlp.fi.muni.cz/trac/noske/.

[24] Apart from terminology, see *IATE* – Inter-Active Terminology for Europe, formerly *Eurodicautom*, at http://iate.europa.eu/.