Alexandr Rosen
Univerzita Karlova, Praha

InterCorp – a look behind the façade of a parallel corpus

InterCorp – korpus równoległy od kuchni

Streszczenie

InterCorp to projekt, który powstał na Wydziale Filozoficznym Uniwersytetu Karola w Pradze. Jego celem jest zbudowanie obszernego równoległego korpusu synchronicznego, który obejmowałby jak najwięcej języków. W tworzeniu korpusu uczestniczą pracownicy naukowi i studenci Wydziału Filozoficznego Uniwersytetu Karola, osoby związane z Czeskim Korpusem Narodowym, a także współpracownicy zewnętrzni.

InterCorp to rzeczywiście obszerny i ciągle rozwijający się synchroniczny korpus równoległy, obejmujący teksty w języku czeskim i 38 innych językach – w tym w języku polskim (wersja 8; stan w lutym 2016), dostępny online poprzez interfejs. Trzon korpusu, który stanowi półautomatycznie opracowana beletrystyka, jest uzupełniony automatycznie opracowanymi tekstami z zakresu publicystyki i prawa, a także zapisami debat parlamentarnych i napisami filmowymi. W sumie korpus obejmuje około 1,6 miliarda słów. Wszystkie teksty dysponują wiązaniem segmentów na poziomie zdania i w miarę możliwości są opatrzone lingwistyczną anotacją (z podaniem podstawowych form i kategorii morfologicznych) oraz danymi bibliograficznymi. Po krótkiej prezentacji koncepcji korpusu przedstawiamy jego parametry liczbowe; zwracamy przy tym uwagę na olbrzymią nierównowagę w reprezentacji tekstów z różnych języków, oryginałów i przekładów oraz typów tekstów. Staramy się także dokonać porównania z niektórymi innymi projektami tego typu. W części poświęconej wykorzystaniu korpusu zwracamy uwagę na możliwości i ograniczenia wyszukiwarki KonText (wcześniej wykorzystywane wyszukiwarki Bonito i NoSketch Engine nie są już dostępne) oraz różne sposoby wykorzystania tekstów równoległych takich jak ekscerpcja ekwiwalentów leksykalnych czy analiza zgodnych fragmentów tekstu. Spojrzenie na korpus od strony użytkownika jest uzupełnione komentarzem twórców korpusu. W części przedstawiającej opracowywanie tekstów przed ich włączeniem do korpusu oczekiwania i życzenia użytkowników zostają skonfrontowane z koncepcyjnymi, technicznymi i fizycznymi możliwościami budowy korpusu paralelnego. Końcowa część zawiera wnioski, jakie się nasuwają na podstawie dotychczasowych doświadczeń, a także plany na przyszłość obejmujące zarówno konkretne projekty twórców korpusu, jak i koncepcje dotyczące zmian wymagających dużych technicznych interwencji w samej strukturze korpusu.

Powstały i ciągle rozwijany korpus równoległy InterCorp ma z założenia służyć między innymi jako źródło danych do badań teoretycznych, analiz gramatycznych i leksykograficznych, prac translatorskich, projektów dotyczących nauki języków obcych, a także jako materiał do badań dla studentów.

Keywords: parallel corpus, Czech, multilinguality, user feedback, annotation, balance
Słowa kluczowe: korpus równoległy, język czeski, wielojęzyczność, feedback od użytkowników, anotacja, równowaga

1. **About *InterCorp***

InterCorp,[1] a part of the *Czech National Corpus* (*CNC*),[2] is a multilingual parallel corpus, built since 2005 at Charles University in Prague. Although its original purpose was to serve researchers, teachers and students from the linguistic departments at the Faculty of Arts, it has reached out to users beyond the academic community and national borders. However, its typical users are still humans, with their varied and often challenging needs, rather than computer applications.

New releases of the corpus are published approximately once per year. With each new release the amount of texts is growing, often together with the number of languages and the extent and quality of annotation. Starting with release 6, previous versions remain available on-line. Currently (at release 8) the corpus includes about 1.4 billion words in 38 languages plus 174 million words in Czech.[3] All 'foreign' texts have a Czech counterpart, while a foreign text may have no counterpart in any other foreign language.

There are two main groups of texts included in the corpus: the core, consisting largely of literary texts, and collections as well as a mix of other text

---

[1] For more details about the corpus see http://www.korpus.cz/intercorp/. For a slightly outdated but more theoretically oriented account see Čermák and Rosen (2012), or the more technically focused paper Rosen and Vavřín (2012). The project is supported by the Ministry of Education of the Czech Republic, project no. LM2011023.

[2] https://www.korpus.cz

[3] See Table 2. for more details. Like any other *CNC* corpora published since 2014, *InterCorp* is now officially described as a reference corpus. The reason for using this term is the permanent availability of its previous releases in their entirety. We are aware of the somewhat non-standard usage of this term, cf. Brown (2005: 209): "When a sample corpus claims to be a reasonably reliable repository of all the features of a language, it can be called a reference corpus."

types, obtained from freely available resources. The proportions are very much language-specific. The size of the core part (altogether 194 million words in 28 languages plus 85 million words in Czech) ranges from 3 titles in Arabic to 327 titles in German. The core has a privileged status as the linguistically more interesting and reliable resource, also because it has been proofread for typos, sentence segmentation and alignment errors.

The collections are acquired from other multilingual corpora, web services or databases. The languages of the EU countries have a substantial portion of legal texts and parliament proceedings (approx. 40 million per language from *JRC-Acquis*,[4] the Acquis Communautaire corpus, and about 9–17 million from *Europarl*,[5] the corpus of European Parliament proceedings), and some include journalistic texts (approx. 4 million per language from Project Syndicate,[6] a site of newspaper commentaries, and Voxeurop,[7] a European news site). For most languages the corpus also includes film subtitles (in sizes ranging from 113 thousand words in Japanese to 52 million words in English; obtained from the *Open Subtitles*[8] database).

Texts in all languages are equipped with available bibliographical data, such as translator's name, language of the original or publication year, and are automatically aligned by sentences with a corresponding text in Czech. Czech has the role of the pivot – two foreign languages are aligned via Czech. Depending on the availability of tools, texts in 20 languages are lemmatized and/or tagged.

InterCorp can be accessed via a standard web browser from the integrated search interface of the *CNC*.[9] Upon request and after signing a non-profit license agreement, the texts can also be acquired as bilingual files, including shuffled pairs of sentences as a physical protection against infringement of copyright.

On the organizational front, the Institute of the Czech National Corpus (ICNC) is responsible for the top-level management, financing, technical support, training, consulting, central data repository, automatic alignment, morphosyntactic markup, lemmatization, availability and dissemination of *InterCorp*. The coordinator for a specific language is responsible for the selection and acquisition of texts (pending the Institute's approval), proofreading and alignment checking. While most coordinators are the staff of the Faculty of

---

[4] http://ipsc.jrc.ec.europa.eu/index.php?id=198

[5] http://www.statmt.org/europarl/

[6] http://www.project-syndicate.org/

[7] Formerly Presseurop: http://www.voxeurop.eu

[8] http://www.opensubtitles.org

[9] https://kontext.korpus.cz

Arts, some come from other faculties of Charles University or other institutions: Masaryk University in Brno, Palacký University in Olomouc, the Czech Academy of Sciences, University of Warsaw and the Polish Academy of Sciences. Some texts, mainly the collections but also fiction titles, and many of the tools, such as taggers, have been acquired, processed or developed by researchers from abroad.[10]

2. *InterCorp* in numbers

Table 1 shows the number of words (in millions) for Czech, Polish, all foreign languages and the total, separately for each text group. The more detailed Table 2 shows the number of words (in thousands) for each language and text group. For the core part, the number of texts is also included. There are striking differences between the languages. Some languages of the EU countries are represented in all the text groups, with a correspondingly high total (German, English, Spanish, French, Italian, Dutch, Portuguese), but not all of them also have a high number of core texts. In addition to German, English and Spanish, languages with over 10 million words in the core part include Croatian and Polish. On the other hand, there are languages such as Arabic and Hindi with very few texts in the core, or Hebrew, Icelandic, Japanese, and Albanian with some texts from *Open Subtitles* and nothing else. It is mainly this disproportionate distribution of texts across languages that makes *InterCorp* a somewhat opportunistic corpus (arguably an unavoidable feature of all parallel corpora), suffering from a shortage of suitable texts, or – for some language pairs – of any texts.

	Czech	**Polish**	**All foreign**	**Total**
Core	84.7	17.5	194.1	278.8
Syndicate	3.4	0	20.1	24.1
Voxeurop	2.3	2.4	24.7	27.0
Acquis	20.3	20.6	430.2	450.5
Europarl	12.9	12.8	265.0	278.0
Subtitles	50.7	26.6	488.4	539.1
Total	174.3	79.9	1,423.1	1,597.5
No. of core texts	1,282	232	2,516	3,798

Table 1. The size of *InterCorp* in million words, with details for Czech and Polish

	Language	***Core***		***Syndicate***	***Voxeurop***	***Acquis***	***Europarl***	***Subtitles***	**Total**
		words	**texts**						
ar	Arabic	34	3						34
be	Belarusian	2,153	39						2,153
bg	Bulgarian	5,241	68			13,816	9,083		28,141
		words	**texts**						

---

[10] See http://ucnk.ff.cuni.cz/intercorp/?lang=en for details.

Language		*Core*		*Syndicate*	*Voxeurop*	*Acquis*	*Europarl*	*Subtitles*	Total
ca	Catalan	4,633	46						4,633
da	Danish	3,017	27			21,680	13,916	14,430	53,042
de	German	27,682	327	3,725	2,483	21,724	13,089	8,367	77,070
el	Greek					25,070	15,404	23,715	64,188
en	English	15,488	178	3,818	2,670	24,208	15,580	52,101	113,866
es	Spanish	17,476	214	4,324	2,816	27,001	15,885	36,379	103,882
et	Estonian					15,963	10,900	10,296	37,158
fi	Finnish	3,426	58			16,455	10,175	15,098	45,154
fr	French	9,170	137	4,393	2,928	27,352	17,178	25,962	86,983
he	Hebrew							16,221	16,221
hi	Hindi	409	7						409
hr	Croatian	15,480	215					19,093	34,572
hu	Hungarian	5,388	71			19,177	12,307	21,240	58,110
is	Icelandic							1,585	1,585
it	Italian	7,248	69	652	2,708	24,849	15,489	14,654	65,599
ja	Japanese							113	113
lt	Lithuanian	358	17			18,393	11,213	558	30,522
lv	Latvian	1,337	36			18,745	11,689	280	32,051
mk	Macedonian	3,742	49					1,877	5,619
ms	Malay							3,521	3,521
mt	Maltese					14,133			14,133
nl	Dutch	9,962	119	314	2,956	24,746	15,563	29,363	82,904
no	Norwegian	4,816	54						4,816
pl	Polish	17,516	232		2,378	20,628	12,811	26,572	79,906
pt	Portuguese	2,393	29	369	3,000	28,603	16,485	43,392	94,242
ro	Romanian	3,433	36		2,738	8,200	9,446	34,129	57,945
ru	Russian	3,338	63	3,174				6,886	13,397
sk	Slovak	7,402	140			19,223	12,734	5,134	44,493
sl	Slovene	900	15			19,646	12,241	17,025	49,811
sq	Albanian							2,004	2,004
sr	Serbian	8,824	100					20,777	29,601
sv	Swedish	8,138	100			20,586	13,840	14,694	57,258
tr	Turkish							21,191	21,191
uk	Ukrainian	5,054	67					246	5,300
vi	Vietnamese							1,474	1,474
Total		194,055	2,516	20,770	24,677	430,195	265,029	488,373	1,423,099
cs	Czech	84,718	1,282	3,416	2,315	20,303	12,923	50,688	174,364

Table 2. The size of *Intercorp* by language and text groups in thousands of words and in text units (for core texts)

While the text types and their mix is not a critical factor for some kinds of research and applications, other users are quite discriminating and treat some data, such *InterCorp*'s collections, as the last resort option. This may not be primarily because the linguistic annotation and alignment of these data is of a lower standard compared with the core part. The main complaints concern missing metadata (especially about the source language) and the types of texts included in the collections. This is why many users focus on the core part, despite its limitations in terms of size. However, even in the core part there are issues of disproportionate distribution. The most obvious differences across languages are in terms of size (see the *Core* column in Table 2 again). Yet other differences are not visible at first sight, although some users may perceive them to be as critical as limited size.

As a multilingual corpus, *InterCorp* should offer large amounts of texts in as many languages as possible to provide data for truly cross-lingual types of research. The intersection of texts available in multiple languages in the core part of the corpus is very much dependent on both the languages and the texts. As a rough guide, there are now 9 texts in the core part, which are available in at least 20 languages, 27 texts in at least 15 languages, 55 texts in at least 10 languages and 186 texts in at least 5 languages. A Polish translation is available for all of the texts in 15 and more languages, and there are still 110 texts available in five or more languages including Polish. Table 3 shows 27 texts covered in most languages. The list is hardly a balanced mix – except for six Czech novels and a single novel in French, Italian, Portuguese and Russian, the rest is all English originals. Moreover, there are as many as five novels authored by Joanne Rowling, four by J. R. R. Tolkien and three by Milan Kundera. This is perhaps the best illustration of the thorny path to the elusive ideal of a representative parallel corpus.

Another major concern may be the size of available texts for a specific language pair. Table 4 shows the figures for each pair of the core part, shown separately for each language in the pair. For example, Polish texts aligned with German include 6.0 million words ("pl" column, "de" row), while corresponding German texts aligned with Polish include 6.9 million words ("de" column, "pl" row).[11]

Yet another case where the distribution of texts across languages may not be quite satisfactory is the ratio of originals to translations, and the availability

---

[11] The diagonal shows the total number of words for all texts in the language. The extent and sizes of collections available for a specific pair are easy to determine from Table 2. Another option is to use KonText. After clicking the bottommost button 'Refine selection', KonText shows the number of *tokens* (i.e. words plus punctuation signs) for the texts in the language in focus which are aligned with one or more other specified languages and/or which are subject to some other constraints according to the metadata.

of the original. Table 5 shows only texts which have their original version in one of the languages of the pair. For each language with some texts in the core, the rows indicated by the corresponding language code in the first column show the number of texts according to the language of the original, given in the column heading. For example, the core includes three texts in Arabic (the last but one column, headed **Σ**), one original text (in the column headed "ar"), one text translated from Czech (in the column headed by "cs") and one translated from German (in the column headed "de"). The row with "cs" in the first column has at least one text in each column – each text in a foreign language has a Czech counterpart. Except for the column headed "cs", which shows the number of Czech originals (in the language of the original, i.e. in Czech), the numbers in the "cs" row indicate the number of original texts (in the language indicated in the column heading), which are translated into Czech.

Languages	**Author**	**Title**
26	Rowling	*Harry Potter and the Philosopher's Stone*
26	Saint-Exupéry	*The Little Prince*
23	Carroll	*Alice in Wonderland*
21	Kundera	*The Unbearable Lightness of Being*
21	Rowling	*Harry Potter and the Chamber of Secrets*
21	Tolkien	*The Fellowship of the Ring*
20	Kundera	*The Joke*
20	Adams	*The Hitch Hiker's Guide to the Galaxy*
20	Tolkien	*The Return of the King*
19	Bulgakov	*The Master and Margarita*
19	Rowling	*Harry Potter and the Prisoner of Azkaban*
19	Brown	*The Da Vinci Code*
19	Tolkien	*The Two Towers*
18	Tolkien	*The Hobbit or There and Back Again*
18	Hašek	*The Good Soldier Švejk*
18	Eco	*The Name of the Rose*
18	Milne	*Winnie the Pooh*
17	Orwell	*1984*
17	Kafka	*The Trial*
17	Rowling	*Harry Potter and the Goblet of Fire*
17	Coelho	*The Alchemist*
16	Kundera	*Immortality*
16	Frank	*The Diary of a Young Girl*
16	Hrabal	*I Served the King of England*
16	Kipling	*The Jungle Book*
15	Kundera	*Laughable Loves*
15	Rowling	*Harry Potter and the Order of the Phoenix*

Table 3. The top 27 texts in most languages in the core part of *InterCorp*

The columns show how many original texts in the language specified in the heading have a translation in the other languages, indicated in the first column. A language such as English ("en") has at least one text in nearly each row, which means that translations of English originals occur in almost all languages of the *InterCorp* core. The English column is exceptional for another reason too: there are as many as 242 texts translated into Czech while there are far fewer original English texts (125). This means that the core does not include English originals for 117 texts. In all of these cases, a Czech translation is aligned with one or more translations, while the English original is missing. The last column ("other") shows the number of original texts in languages not included in the core of *InterCorp*.

The diagonal gives the number of original texts for the corresponding language of the row and the column. The best-represented languages are Czech (267), German and Spanish (126), English (125) and French (83). On the other hand, the core does not include any original text in Hungarian or Romanian. There is not even any translated Romanian original. But even in languages with a more representative content, the user may be disappointed to see cases of some very lopsided balance between originals and translations. For a pair such as Polish and Czech, the proportion is 46:36 in favour of Polish originals (2.5 million vs. 2.1 million in the number of words, see Table 6), which is a reasonable balance, similar to that for German and Czech (126:85). On the other hand, foreign originals prevail in the English-Czech (125:25), Spanish-Czech (126:25) and French-Czech pairs (83:36). The opposite applies to Croatian and Czech (26:71) and a few other "smaller" languages. Seen from this angle, the best-represented pair is Slovak and Czech, with the score 56:55.

Table 6 shows similar statistics. This time, the texts are not counted in items, but in thousands of words. For example, according to Table 6 the core of *InterCorp* includes 551 thousand words in German originals for which a Polish translation is available ("de" column, "pl" row). Table 5 shows that there are actually 8 such texts. On the other hand, there are 114 thousand words in Polish originals for which the corpus has a German translation ("pl" column, "de" row) in 3 texts according to Table 5. The following remarks are due here:

There is a reason why the number of words in German originals translated into Czech (10,968 thousand) is lower than in untranslated German originals (11,547 thousand), even though the corpus includes more German originals translated into Czech (134) than those untranslated (126). This is because languages may differ significantly in the number of words within the same parallel texts.

	ar	be	bg	ca	cs	da	de	en	es	fi	fr	hi	hr	hu	It	lt	lv	mk	nl	no	pl	pt	ro	ru	sk	sl	sr	sv	uk
ar	34	28			35		34				8		31					32		35		6					32		
be	28	2,153	910	340	2,221	369	1,283	1,381	627	173	559	17	1,028	524	657	110	54	832	1,055	664	1,290	396	570	407	259	71	1,505	453	756
bg		820	5,241	1,603	5,029	1,146	2,531	2,451	2,303	655	1,729	62	2,290	1,819	2,007	71	54	1,737	2,587	1,996	2,693	1,000	397	1,214	580	365	2,621	1,453	2,186
ca		248	1,238	4,633	3,660	819	2,571	1,461	3,836	625	1,000		2,221	1,154	2,071	214	135	1,110	1,753	1,832	1,593	1,051	856	598	242	289	1,801	796	1,016
cs	34	2,153	5,241	4,633	84,743	3,017	27,656	15,488	17,476	3,426	9,170	409	15,480	5,388	7,248	358	1,337	3,742	9,962	4,816	17,517	2,393	3,433	3,338	7,402	900	8,824	8,138	5,054
da		249	927	838	2,487	3,017	1,675	1,373	1,308	170	884	261	1,047	867	969	60	2	1,241	1,158	927	1,394	936	79	332	76	81	1,660	815	1,366
de	28	1,081	2,270	2,847	23,891	1,813	27,656	6,633	5,761	1,628	2,981	120	6,331	2,199	3,249	228	118	2,605	5,258	3,628	5,992	1,173	1,137	1,523	819	455	4,547	2,774	2,556
en		1,080	2,209	1,648	12,951	1,465	6,692	15,488	3,425	923	2,273	120	4,347	1,767	1,584	259	83	2,214	4,664	2,466	4,191	907	1,047	2,111	433	460	3,995	1,998	2,567
es		505	2,065	4,258	15,140	1,415	5,735	3,464	17,476	874	2,457	62	5,519	1,678	4,576	216	137	1,795	3,860	3,145	4,320	1,230	1,632	697	260	289	3,663	2,558	2,314
fi		197	797	840	3,965	229	2,143	1,167	1,073	3,426	755	45	1,502	688	692	110	100	580	1,458	1,389	1,487	154	438	574	443	265	976	662	667
fr	6	421	1,530	1,038	7,281	831	2,818	2,200	2,253	593	9,170	120	2,122	1,479	1,681	209	154	1,446	2,459	1,629	2,380	596	591	806	227	526	1,804	1,415	1,485
hi		12	46		297	224	92	105	50	26	101	409	43	12	51			13	54	81	43	38	15	40			205	15	155
hr	28	963	2,270	2,734	14,707	1,208	7,069	4,936	6,205	1,252	2,595	62	15,480	1,639	3,209	253	174	1,990	5,271	3,324	4,800	865	1,747	1,274	585	405	4,125	2,686	2,692
hu		496	1,953	1,557	5,473	1,137	2,636	2,089	2,025	614	1,876	17	1,862	5,388	1,766	171	81	1,549	2,314	1,803	2,460	867	578	840	384	374	2,236	1,478	1,590
it		497	1,791	2,337	6,451	1,050	3,326	1,590	4,669	606	1,877	62	2,974	1,467	7,248	113	54	1,347	2,546	2,156	2,804	905	871	444	115	149	2,557	1,862	2,166
lt		127	87	338	418	90	320	380	320	111	327		324	196	161	358	18	219	469	361	346	69	218	199	15	81	268	145	69
lv		65	73	196	1,407	2	157	108	195	99	216		214	86	84	16	1,337	139	105	135	133		220	94	257	92	75	104	65
mk	28	720	1,677	1,379	3,494	1,509	2,844	2,403	1,998	462	1,670	17	1,994	1,403	1,487	171	111	3,742	2,361	1,546	2,390	961	464	912	478	297	2,321	1,316	1,913
nl		872	2,263	1,931	8,093	1,210	5,040	4,542	3,716	1,020	2,456	62	4,607	1,799	2,387	326	83	2,147	9,962	2,705	4,390	835	1,079	1,687	433	374	3,449	2,322	2,540
no	28	494	1,795	1,980	4,052	960	3,587	2,464	3,072	1,071	1,695	103	2,913	1,495	2,078	244	102	1,434	2,880	4,816	2,663	858	993	792	520	395	2,290	1,724	1,553
pl		1,251	2,868	2,031	17,625	1,768	6,942	4,892	5,009	1,305	2,927	62	5,025	2,353	3,202	297	116	2,533	5,257	3,134	17,517	1,099	1,173	2,042	820	519	4,438	2,567	3,587
pt	6	269	805	1,087	1,950	945	1,066	835	1,121	111	631	45	759	664	842	43		796	816	826	861	2,393	105	283	76	28	871	613	644
ro		452	338	872	2,800	88	1,073	1,037	1,542	317	567	17	1,480	438	804	143	135	388	1,053	948	935	111	3,433	202		149	1,659	402	278
ru		420	1,320	785	3,459	443	1,824	2,490	823	521	988	75	1,372	850	497	174	83	997	2,077	914	2,113	372	278	3,338	433	297	1,687	1,236	1,399
sk		266	645	330	7,510	96	986	450	330	383	193		624	377	142	13	244	541	567	555	849	94		436	7,402	237	684	342	499
sl		65	371	328	813	87	483	432	303	211	498		379	326	156	58	74	305	431	355	477	30	171	268	220	900	362	410	339
sr	28	1,326	2,529	2,158	8,165	1,901	4,955	4,379	3,898	807	2,076	265	4,030	1,978	2,691	209	54	2,287	3,877	2,482	4,128	1,000	1,882	1,551	640	368	8,824	1,859	2,970
sv		369	1,338	903	7,116	872	2,803	1,958	2,609	511	1,448	17	2,493	1,248	1,867	101	81	1,174	2,460	1,765	2,226	674	448	1,040	317	405	1,742	8,138	1,605
uk		761	2,453	1,515	5,259	1,901	3,155	3,163	2,908	592	1,888	220	3,003	1,679	2,669	58	54	2,224	3,197	1,941	3,722	961	374	1,413	499	368	3,409	1,929	5,054

Table 4. The size of core bitexts in thousands of words: column headings indicate the language of the text, row labels “the other” language

→ orig text ↓	ar	be	bg	ca	cs	da	de	en	es	fi	fr	hi	hr	hu	it	lt	lv	mk	nl	no	pl	pt	ro	ru	sk	sl	sr	sv	uk	Σ	other
ar	1				1		1																							3	
be		3			8		4	13	1		1		1								3			2	1		1	1		39	
bg			19		9		1	27			4				2						1	1		2				2		68	
ca				1	16		3	12	5	1	2				3							1		1						45	1
cs	1	3	19	1	267	9	134	242	127	24	95	2	26	1	20	1	7	1	30	7	49	21		39	56	3	8	58	6	1257	
da					6	9		12																						27	
de					85		126	65	10	1	4			1	7	1	1		6	3	3	2		3	1		3	5		327	
en					25		4	125			3				1				2		1	1		6			5	4		177	1
es				1	25		8	29	126	1	6				7					1		4		2				3		213	1
Fi					11	1	1	12	2	25					1					1		1						2		57	1
fr					36		1	10			83				2				1			2		2						137	
hi					2			1			1	2										1								7	
hr			1		71		15	52	11	2	4		26		6				7	1	3	4		1		1		8		213	2
hu					16		5	23			9				1							3		14						71	
it					4		4	21	9	1	3				19							3		1				3		68	1
lt					8		2	2								1	1				2				1					17	
lv					22		2	1								1	7				2				1					36	
mk					15		1	16			1		1		1			2	1		3			2			2	4		49	
nl					24		3	33	7		3				3				30	2	2	3		3				6		119	
no					11		5	21	4		1				3					6		2						1		54	
pl					36		8	97	10	2	8				2	1	1		3	1	46	4		6	1			5		231	1
pt					6			8														15								29	
ro					7		5	12	3		1		1		1						1	1					1			33	3
ru					9		1	22			2								1		1			22			1	3		62	1
sk					55		2	5	1								1				2				56					122	18
sl					7		1	2					1													2		2		15	
sr					11		7	33	9		3				7				2		4	3		10	1		5	2		97	3
sv					11		4	23	7		2				1				1									50		99	1
uk					6		1	31	3		5				2						5			3				5	6	67	
Σ	2	6	39	3	810	19	349	950	335	57	241	4	56	2	89	5	18	3	84	22	128	72		119	118	6	26	164	12		

Table 5. The number of texts in *InterCorp* by language of the text and of the original (for core texts)

orig → ↓ text	ar	be	bg	Ca	cs	da	de	en	es	fi	fr	hi	hr	hu	it	lt	lv	mk	nl	no	pl	pt	ru	sk	sl	sr	sv	uk	Total	other
ar	1				6		28																						34	
be		141			317		215	792	116		12		31								153		209	43		104	22		2,153	
bg			1,277		697		71	2,106			347				371						58	34	237				42		5,241	
ca				65	1,038		274	1,435	621	265	202				396							167	48						4,511	122
cs	1	138	1,269	53	13,451	831	10,968	20,583	8,635	1,610	5,102	57	1,482	29	1,452	2	288	73	1,338	624	2,758	1,453	2,872	3,419	201	651	4,347	423	84,109	422
da					207	994		1,816																					3,017	
de					5,263		11,547	6,544	901	275	266			6	873	2	2		337	305	114	172	198	2		335	515		27,656	
en					2,212		263	10,546			251				40				161		67	40	926			503	377		15,387	102
es				61	1,504		587	2,786	9,818	243	438				809					45		297	169				608		17,366	110
fi					587	107	100	706	115	1,397					143					115		26					130		3,426	
fr					2,473		76	926			5,061				233				113			94	194						9,170	
hi					62			203			17	82										45							409	
hr			29		4,131		1,051	4,143	1,094	200	246		1,517		601				366	230	174	246	111		140		928		15,207	272
hu					1,038		286	1,762			811				157							135	1,198						5,388	
it					254		506	2,665	826	224	254				1,482							150	139				573		7,074	174
lt					274		3	73								1	1				3			2					358	
lv					1,052		3	2								2	273				3			2					1,337	
mk					992		32	1,576			13		66		38			109	70		183		256			223	184		3,742	
nl					1,974		219	3,418	815		117				262				1,638	221	169	166	306				656		9,962	
no					826		341	1,656	421		159				439					632		172					171		4,816	
pl					2,093		551	8,765	1,050	316	509				192	2	2		166	140	2,509	234	473	2			511		17,514	49
pt					193			961														1,239							2,393	
ro					545		413	1,144	411		15		87		165						81	142				49			3,050	383
ru					757		66	1,117			68								75		49		914			86	206		3,338	
sk					2,628		26	1,216	40								2				127			3,354					7,393	9
sl					463		77	171					10												68		111		900	
sr					617		606	3,091	766		146				684				134		275	224	1,295	72		339	189		8,438	386
sv					811		295	1,954	399		72				191				100								4,317		8,138	
uk					513		16	2,384	158		195				188						326		293				552	429	5,054	
Total	2	279	2,575	179	46,977	1,933	28,619	84,539	26,186	4,530	14,301	139	3,192	35	8,715	9	567	182	4,500	2,312	7,051	5,034	9,839	6,895	409	2,290	14,439	852	276,579	2,028

Table 6. The size of the corpus by language of the text and of the original (in thousands of words for core texts)

Except for Czech, the table does not actually show the size of texts in a specific language aligned with texts in another specific language, because the cells do not show figures for texts available as translations from a third language.

The size of a language-specific part of the corpus aligned with one or more specific languages can be found in Table 4 (in words for specific language pairs) or from the search interface,[12] where the results are presented in the number of tokens (i.e., including punctuation symbols) rather than words. For instance, the Polish-German pair includes 7,392 thousand Polish tokens. When parallel texts in English are added, the number drops to 4,000 thousand tokens. For a combination of four languages, including additional parallel texts in Spanish, the texts available in Polish include 2,640 thousand tokens.

3. Some other parallel corpora

InterCorp is not the only project of its kind. Table 7 below shows *InterCorp* in comparison with some other resources offering access to parallel texts. For each of the resources the table includes some basic information on the types of texts available, languages included, size (in Billions or Millions of words or sentences), annotation (Morphology, Syntax, Semantics), alignment level (Sentences, Words), human intervention in the text processing (Proofread), on-line Search and Download option, and availability of Metadata.

It is perhaps the combination of features that makes *InterCorp* different from the other corpora. On the one hand, there are some very large, massively multilingual resources such as *Opus*, compiled from as many freely available texts as possible, with the Czech part reaching at least 150 million words. On the other hand, there are much smaller resources including literary texts from specific domains, such as *ParaSol* and *ASPAC*. In *InterCorp*, the user can find texts of either type, processed according to the same methodology and offered within the same search and display interface.

---

[12] Visit https://kontext.korpus.cz, select the appropriate combination of languages, restrict to the Core group and click the button "Refine selection".

Name	Types	Langs	Size	Annot	Aligned	Proofread	Search	Download	Metadata
Linguee[13]	legal	25	?	no	S,W	no	yes	no	yes
Glosbe[14]	varia	100+	1Bs	no	S,W	no	yes	no	yes
SKE[15]	varia	38	cs:217Mw	no	S	no	yes	yes	yes
DGT-TM[16]	legal	22	cs:3.7Mw	no	S	yes	no	yes	no
Pelcra[17]	varia	31	pl:58Mw	no	S,W	part	yes	yes	yes
RNC[18]	varia	6	9Mw	M	S	part	yes	?	yes
SNK[19]	fiction	7	sk:388Mw	M	S	no	yes	part	yes
CzEng[20]	varia	en,cs	en:233Mw	M,Sy	S	no	yes	yes	no
PCEDT[21]	news	en,cs	1.2Mw	M,Sy,Se	S,W	yes	yes	yes	yes
Kačenka[22]	fiction	en,cs	3.3Mw	no	S	yes	no	yes	yes
Opus[23]	varia	100+	4.7Bw	M,Sy	S,W	no	yes	yes	no
ParaSol[24]	fiction	31	27Mw	M	S	part	yes	?	yes
ASPAC[25]	fiction	25	68 texts	no	P	yes	no	?	yes
InterCorp	varia	32	1.6Bw	M	S	part	yes	yes	yes

Table 7. Some other parallel corpora in comparison to *InterCorp*

4. Using *InterCorp*

Most users interact with the corpus data via KonText,[26] the web-based interface built on top of the corpus query engine Manatee.[27] This interface is now used for all *CNC* corpora, superseding Park, a search interface dedicated to parallel corpora.

The interface offers a number of options for pre-selecting texts before making a query according to languages and all available metadata, such as text

---

[13] Online search through bilingual texts – http://www.linguee.com

[14] Translation Memory Online – http://glosbe.com/tmem/

[15] Sketch Engine – http://www.sketchengine.co.uk

[16] Translation Memory of the EC's Directorate-General for Translation – http://ipsc.jrc.ec.europa.eu/?id=197

[17] Polish & English Language Corpora for Research & Applications – http://pelcra.pl/new/. For its new parallel search interface see http://paralela.clarin-pl.eu and Pęzik (this volume).

[18] Russian National Corpus – http://www.ruscorpora.ru

[19] Slovak National Corpus – http://korpus.juls.savba.sk/par.html

[20] Czech-English parallel corpus – http://ufal.mff.cuni.cz/czeng, https://lindat.mff.cuni.cz/services/kontext/run.cgi/first_form?corpname=czeng_10_cs_a

[21] Prague Czech-English Dependency Treebank – http://ufal.mff.cuni.cz/prague-czech-english-dependency-treebank

[22] English-Czech Corpus of the Department of English Studies, Faculty of Arts, Masaryk University Brno – http://www.phil.muni.cz/angl/kacenka/kachna.html

[23] An open source parallel corpus – http://opus.lingfil.uu.se

[24] A Parallel Corpus of Slavic and other languages – http://www.slavist.de

[25] The Amsterdam Slavic Parallel Corpus – http://home.medewerker.uva.nl/a.a.barentsen

[26] See http://kontext.korpus.cz. KonText is developed by the *CNC* team led by Tomáš Machálek.

[27] See Rychlý (2007) and Kilgarriff et al. (2014).

type, source language or publication year. These options can also be used to create custom subcorpora. Queries can be made about a single language or in parallel, using single forms, lemmas, form strings or CQL expressions. In addition to a number of other options, concordances can be filtered, exported, sorted, flagged for further processing, or be used for producing frequency distributions or finding collocations.

Some research tasks require full texts rather than sets of concordances in response to corpus queries. Not even statistics based on a part of the corpus or on the concordances can meet such needs. This applies mainly to the use of corpus data in NLP applications such as machine translation, but also to some studies spanning sentence or even paragraph boundaries. The only solution is some form of access to full texts. After signing a non-profit license agreement,[28] texts from *InterCorp* can be acquired as bilingual files. Each file is extracted from a specific text and includes alignment pairs of sentences in blocks up to 100 words (per language), with the blocks shuffled in random order to prevent the use of texts in violation of copyright, while retaining some text structure. 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.

Parallel texts can be seen as interpreting or even 'annotating' each other through the medium of another natural language. This applies to segments of different sizes: texts, paragraphs, sentences, phrases or words. A practical use of this obvious observation rests on the availability of alignment at the level of such units. Existing methods and tools[29] can align words, producing results with a reasonable error rate, usable for tasks such as the extraction of glossaries of translation equivalents. The *CNC* site now offers lists of such equivalent pairs (lemmas or base forms) in Czech and most other languages, sorted primarily by their frequency in the corpus.[30] This is just one of many possible applications using the parallel corpus and offering the results from the corpus site.[31]

---

[28] The license restricts the use of the data to educational and research purposes and prohibits re-distribution.

[29] E.g., Och, Ney (2003).

[30] See http://treq.korpus.cz. See also Kaczmarska (this volume), Kaczmarska et al. (2015) and Rosen et al. (2014) for examples of research based on these results.

[31] The site shows the following list of top Polish equivalents with frequencies of the Czech noun *bouře* 'storm': *burza* (353), *sztorm* (44), *śnieżyca* (35), *wichura* (16), *szturm* (11), *nawałnica* (9), *huragan* (8), *zamieć* (7), *zawierucha* (7), *wiatr* (6), *burzyć* (5), *zawieja* (4), *wichr* (4), *zamieszka* (4), *bunt* (4), *ulewa* (3), *wicher* (2), *wrzawa* (2), *salwa* (2), *padać* (2), *fala* (2), *sztormowy* (2); a similar list in German for the Czech verb *křičet* 'to cry' is: *schreien* (2145), *rufen* (379), *brüllen* (132), *anschreien* (46), *Schrei* (40), *schreiend* (32), *laut* (17), *kreischen* (17), *aufschreien* (16),*Schreien* (13), *Geschrei* (12), *geschrien* (8), *ausstoßen* (6), *schrein* (5), *zurufen* (5), *brüllend* (4), *ausrufen* (4), *sprechen* (4), *angeschrien* (4), *geschrieen* (3), *losschreien* (3), *grölen* (3), *herumschreien* (3), *lärmen* (3), *Schrein* (3), *anschrien* (3), *zuschreien* (3), *Ruf* (3), *anschreie* (3), *zuschrie* (2), *herrschen* (2), *Lärm* (2), *weinen* (2), *nachrufen* (2), *losbrüllen* (2), *toben* (2), *schriest* (2), *verlangen* (2), *Sie* (2).

5. Pre-processing of texts

Most texts in the core of *InterCorp* pass through the following stages: acquisition, scanning and character recognition, proofreading, segmentation (sentence boundary detection), sentential alignment, proofreading and checking of segmentation and alignment and morphosyntactic markup. Texts acquired in an electronic form, especially texts in the collections, bypass some of these steps.

Each of the steps has some impact on the quality of the corpus. Acquisition as the first step (including the choice of texts) determines the corpus content. It has recently been subjected to a new policy aimed at achieving a more balanced representation of languages and text types and remedying the lack of original texts.[32] A selected text that cannot be acquired in the electronic form is digitized. After OCR the text is proofread in a text editor with a special focus on aspects critical to text processing for the corpus, such as paragraph boundaries, quotes, diacritics, punctuation and spaces, the latter crucial for tokenization and detecting sentence boundaries. A proofread text is then exported as plain text with XML-like markup, and a bibliographical record is stored in the project database. The steps above are the responsibility of the coordinator for the specific language, who usually employs students for tasks such as post-OCR proofreading. Texts in most languages are segmented into sentences using Punkt, a tool based on an unsupervised learning algorithm,[33] followed by language-specific fixes. Automatically detected sentence boundaries are checked and (if necessary) corrected by a set of regular expressions, targeting contexts where automatic tools tend to fail.

Parallel versions of the text are sentence-aligned using Hunalign.[34] The aligned texts are accessible within InterText, a parallel text editor.[35] Segmentation and alignment can then be checked and corrected, together with any remaining typos. Automatic sentence segmentation typically fails because of an unknown abbreviation, a missing space, or a lower quotation mark improperly recognized as comma(s). Alignments may be incorrect as a result, but some texts can be difficult to align even for humans. All corrections, usually done by research assistants, are logged, checked by the coordinator in charge of the specific language and finally by the project coordinator.

---

[32] For details see Hebal-Jezierska et al., this volume.

[33] See Kiss and Strunk (2006: 485–525), the implementation is due to http://nltk.org/. The training data consist of previously processed texts.

[34] See Varga et al. (2005) and http://mokk.bme.hu/en/resources/hunalign/.

[35] See Vondřička (2010) and http://wanthalf.saga.cz/intertext. *Intertext* can edit sentence-level alignment, sentence segmentation, paragraph boundaries and typos, and is integrated with *Hunalign*. Changes of the text structure in Czech are projected to all alignments. Other features include change logs, export, searching, bookmarking and support for user classes with different privileges. There are two versions: server and personal, and both are available under the GNU GPL v3 license.

Throughout the process, all the core text are registered in the project database with links to available Czech texts. The language coordinators are responsible for including the bibliographical data, which are crucial for text filtering in the corpus search interface. A missing or incorrect piece of information can have a negative impact on research results. The database also tracks the passage of each text through the pre-processing stages. The finished texts are matched with the bibliographical data from the project database and indexed by the corpus manager. So far, only team members can access the database, but a subset of the database will be available to all corpus users in the foreseeable future.

Linguistic annotation of the texts is still restricted to lemmatization and tagging of word forms by morphosyntactic and morphological categories. Moreover, not all languages are annotated in this way: in *InterCorp* release 8 there are 20 languages with tags including Czech, of which 17 have lemmas. Once again, we adopt an opportunistic strategy of using available tools (tokenizers, taggers, lemmatizers), including tokenization principles hard-wired into the tool, tagsets designed elsewhere by experts on the given language and annotation models and trained elsewhere.[36] This approach frequently leads to very different language-specific tagsets as well as non-uniform tokenization and lemmatization principles across the languages.[37]

These achievements come at a price. Luckily, the whole *Czech National Corpus* project has enjoyed continuous support from Charles University and the Czech government over an extended period, allowing for a steady development of *InterCorp* since 2005. The costs of text acquisition and processing are approximately 55,000 EUR per year, including the core texts – about 180 EUR on average per text (the sum for both the Czech and a foreign version and all the steps), as well as the processing of packages. However, the total costs are much higher and harder to estimate, because some overheads are shared by all *CNC* teams. In addition to two full-time dedicated positions, *InterCorp* uses the *CNC* infrastructure and managerial facilities and also relies on the work of other *CNC* staff in the development of corpus methods and tools.

6. Wishlists and issues

In this section we sum up the expectations, wishes and complaints of corpus users with regard to the limitations of corpus design and other constraints on the side of the corpus builders. We start with **content**, perhaps the most critical

---

[36] See http://ucnk.ff.cuni.cz/intercorp/?lang=en for an overview, including the tools used.

[37] For more about issues of annotation, see Section 1.5 in Hebal-Jezierska, this volume.

aspect of any corpus and the main reason for users' concerns about whether their research results are well-founded or whether their intended research is possible at all. Indeed, they would like to see a more representative and/or balanced core in terms of languages, text types, the ratio of originals vs. translations, authors, translators – all of it useful for both contrastive and translatological studies. But it is hard to decide in general which is more important: the **proportions** or the **size** of the corpus. The answer depends very much on the type of research being conducted. Assuming that users are able to determine an optimal mix relative to their research goals and can select texts from the corpus accordingly, the optimal strategy is the more the better, even if that means the result is far from balanced. For some research goals, when two relatively well-represented languages such as German or English are studied in a pair, the overlap of texts in the core may be too small.

For many types of research, the distinction between **originals** and **translations** is crucial. Original texts may be the only texts of interest. However, even when only translations from a third language are compared, the original text should still be available. Unfortunately, this is too often not the case (see Table 5). A priority of the new text selection policy is to remedy this situation.

A related issue is the option of including **multiple translations** in a single language, which is available, e.g. in the *ParaSol* corpus.[38] This interesting feature requires some profound changes in the corpus design and its implementation is not envisaged in the near future.

InterCorp's **search interface** is one of the most advanced tools available among those available for the parallel corpora listed in Table 7. Still there are a number of wishlist items concerning the interface. Some of them are actually small things that can boost user experience, but are not top priority for the developers at the moment, such as **charts** to see the setup of the selected corpus and to prevent the frequent shortcoming of significantly skewed data, a list of **sample queries** for inspiration and time saving, a few keyboard **shortcuts** for more advanced users, context **help** on **tags**, text type **codes** etc., and – last but not least – automatic switching to CQL type query when typing a character such as "[" to prevent frequent attempts to search the corpus inadvertently for a string which is actually a CQL expression. Some other missing features may not be so trivial or simple to implement, but still very useful, such as **biKWiC** – highlighting keyword equivalent, information about the **alignment** type (1:1 or other) and quality (manual or automatic with a confidence score), or labeling/annotating concordances. Another missing feature is related to the possibility of building a **subcorpus** from texts in a specific language aligned with texts in

---

[38] See http://www.slavist.de and von Waldenfels (2006, 2011).

another language, or even for a specific language pair. Some features are actually beyond the mere search and display options, such as statistical comparison across text types, languages, corpora, or lexical profiles, preferably adapted to parallel texts (Belica, 2011; Kilgarriff et al., 2014).

Issues of search and display are very much connected with the need for complete, effective and correct **annotation**. So far, languages differ in tagsets and tokenization rules and a number of languages are still without any linguistic annotation.[39] Finally, although the quality of alignment and metadata has improved, it is not 100% reliable.

7. Lessons learned and perspectives

The bottom line of all the lessons is the importance of user feedback and interaction with the community of users in general. Although *InterCorp* started out with the idea of being a general resource, serving the needs of disparate users and research types, ultimately the requirements of each individual type must be considered and properly addressed. The purpose of the corpus matters, even if it is meant to be a resource for many. There are some obvious questions such as who the users are, what are their needs, how many languages should be included, whether "the more the better" or "the best balance" is a better strategy (in languages, text types, authors, translators, originals/translations/translations for a third language). Perhaps a comparable rather than a parallel corpus is the answer to some research goals. And although all languages should be equal, it is very hard to achieve comparable levels in size, annotation, and representativeness. Strict criteria may be applicable only to a small group of languages.

Parallel corpora, including *InterCorp*, have proven to be a very useful resource for many tasks. Still we believe that their full potential, embodied in the meaning links between expressions across languages and useful for theoretical research, linguistic practice and software applications, has yet to be discovered. Users' needs and wishes may be an important stimulus, but further progress may have an independent motivation. In addition to a larger and more representative pool of texts, more precise, complete and sophisticated annotation is a clear priority. We need to advance the quality of alignment and sentence segmentation, also by crowdsourcing (encouraging users to flag errors). Alignment by words, multi-word units, and phrases are all realistic goals. Linguistic markup should bring better quality for as many languages as possible, including consistent tokenization of contractions and multi-word expressions, a method for reconciling disparate language-specific tagsets, and syntactic annotation.

---

[39] See Hebal-Jezierska et al. (this volume) for more details on issues relating to linguistic annotation and takenization in *InterCorp*.

Many plans involving a specific parallel corpus make better sense if pursued as a joint effort with other similar projects due to a high synergy in infrastructure and content: many problems are similar across languages; texts in foreign languages may exist elsewhere and native speakers are the best corpus builders. Cooperation can have many forms and levels, from the exchange of know-how, tools, or texts between centers, through virtual integration of content, a common search interface (federated search), and a common text dissemination policy, and even a single center providing coordination and infrastructure for all languages. We hope that the existing ties between parallel corpora both within and across national borders will thrive and develop towards a network of parallel resources. As a small step in this direction we plan to release Czech from its pivot role and no longer insist on the presence of a Czech version of the text.

References:

BELICA, Cyril (2011): Semantische Nähe als Ähnlichkeit von Kookurenzprofilen. In: Andrea ABEL, Renata ZANIN (eds.): *Korpusinstrumente in Lehre und Forschung.* Bozen-Bolzano: University Press, 155–178.

BROWN, Keith, (ed.) (2005): *Encyclopedia of Language & Linguistics*. 2nd edition. Amsterdam and Philadelphia, PA: Elsevier.

ČERMÁK, František, ROSEN Alexandr (2012): The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 13(3), 411–427.

KACZMARSKA, Elżbieta, ROSEN, Alexandr, HANA, Jirka, HLADKÁ, Barbora (2015): Syntactico-semantic analysis of arguments as a method for establishing equivalents of Czech and Polish verbs expressing mental states. *Prace Filologiczn*e XVII, 151–174.

KILGARRIFF, Adam, BAISA, Vít, BUŠTA, Jan, JAKUBÍČEK, Miloš, KOVÁŘ, Vojtěch, MICHELFEIT, Jan, RYCHLÝ, Pavel, SUCHOMEL, Vít (2014): The Sketch Engine: ten years on. *Lexicography*, 1(1), 7–36.

KISS, Tibor, STRUNK, Jan (2006): Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4), 485–525.

OCH, Franz Josef, NEY, Hermann (2003): A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 19–51.

ROSEN, Alexandr, KACZMARSKA, Elżbieta, ŠKODOVÁ, Svatava (2014). Zdrobnienia jako element kultury i pułapka glottodydaktyczna. Czeskie i polskie deminutiva w ujęciu konfrontatywnym na podstawie badań korpusowych. In: Elżbieta KACZMARSKA, Andrzej ZIENIEWICZ (eds.): *Glottodydaktyka wobec wielokulturowości*. Warszawa: Wydział Polonistyki Uniwersytetu Warszawskiego, 51–66.

Rosen, Alexandr, Vavřín, Martin (2012): Building a multilingual parallel corpus for human users. In: Nicoletta Calzolari, Khalid CHOUKRI, Thierry DECLERCK, Mehmet Uğur DOĞAN, Bente MAEGAARD, Joseph MARIANI, Asuncion MORENO, Jan ODIJK, Stelios PIPERIDIS (eds.): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul: European Language Resources Association (ELRA), 2447–2452.

Rychlý, Pavel (2007): Manatee/Bonito – a modular corpus manager. In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, 65–70.

Varga, Dániel, Halácsy, Péter, Kornai, András, Nagy, Viktor, Németh, László, Trón, Viktor (2005): Parallel Corpora for Medium Density Languages. In: Galia ANGELOVA, Kalina BONTCHEVA, Ruslan MITKOV, Nicolas NICOLOV, Nikolai NIKOLOV (eds.) *Proceedings of International Conference "Recent Advances in Natural Language Processing"* (*RANLP 2005)*, 590–596.

Vondřička, Pavel (2010): TCA2 – nástroj pro zpracovávání překladových korpusů. In: František Čermák, Jan Kocek (eds.): *Mnohojazyčný korpus InterCorp: Možnosti studia.* Praha: Lidové noviny, 225–231.

von Waldenfels, Ruprecht (2006): Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment. In: Bernhard Brehmer, Vladislava Ždanova, Rafał Zimny (eds.), *Beiträge der Europäischen Slavistischen Linguistik* (*POLYSLAV*) 9. München: Verlag Otto Sagner, 123–138.

von Waldenfels, Ruprecht (2011): Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In: Daniela Majchráková, Radovan Garabík (eds.): *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011* Bratislava: Trilbum EU, 156–162.