Milena Hebal-Jezierska
Uniwersytet Warszawski

Elżbieta Kaczmarska
Uniwersytet Warszawski

Alexandr Rosen
Univerzita Karlova, Praha

Between the devil and the deep blue sea or between users' needs and the compilers' powers: An analysis of the Czech-Polish part of the parallel corpus InterCorp

Między młotem a kowadłem, czyli czego potrzebuje użytkownik korpusu równoległego, a jakie są możliwości twórców korpusów (na przykładzie czesko-polskiej części korpusu równoległego InterCorp)

Streszczenie

Celem artykułu jest próba porównania oczekiwań użytkownika korpusu równoległego co do możliwości prowadzenia różnego typu badań, zwłaszcza analiz konfrontatywnych oraz translatologicznych z technicznymi możliwościami twórców korpusu.

Autorzy rozpoczynają rozważania od szczegółowego opisu problemów twórców InterCorp. Wskazują na największe bolączki polegające na braku proporcji pomiędzy liczbą tekstów w poszczególnych językach umieszczonych w korpusie, a także na tym, że teksty reprezentują różne poziomy anotacji i tagowania. Szczegółowo opisana została polska część korpusu InterCorp. Autorzy podają dane statystyczne dotyczące poszczególnych wersji korpusu. Wiele miejsca poświęcono również problemowi anotacji i tokenizacji (znakowania). Zauważono, że dużym utrudnieniem jest brak jednolitego systemu znakowania dla wszystkich obecnych w InterCorpie języków.

Na przedstawione w skrócie problemy twórców korpusu nakładają się trudności, jakie napotykają jego użytkownicy oraz ich oczekiwania względem jego zasobów. Osoby korzystające np. z zasobów polsko-czeskiej części InterCorpu narzekać mogą na zestawienie tekstów. O ile literatura piękna jest opracowywana ręcznie, o tyle tzw. kolekcje tekstów (Acquis, PressEurope, Europarl, Open Subtitles) są opracowywane tylko automatycznie. Paradoksalnie więc teksty, które nie sprawiają kłopotów twórcom korpusu, są dla niektórych

użytkowników mniej przydatne. Nie można na przykład przeprowadzić szeregu badań opartych na materiale korpusowym, jeżeli nie da się ustalić kierunku przekładu albo języka źródłowego. Dotyczy to wszystkich analiz translatologicznych. Również niedostateczna wielkość korpusu stanowi dla użytkowników dużą przeszkodę. Zbyt mała liczba poświadczeń może uniemożliwić całkowicie przeprowadzenie badań nad konkretnym zjawiskiem leksykalnym czy gramatycznym (przykłady podane zostały w artykule).

Użytkownicy sięgają jednak do korpusów paralelnych, ponieważ, mimo wszelkich niedociągnięć, stanowią one niezwykłe narzędzie służące do poszukiwania ekwiwalentów, a także porównywania znaczeń jednostek językowych. Dopasowanie odpowiedniego tematu badania do możliwości korpusu jest w tym przypadku podstawową czynnością poprzedzającą samo badanie, a jednocześnie gwarantem wiarygodności wyników.

Sposób rozbudowywania InterCorpu jest sprawą powodującą prawdopodobnie największe kontrowersje pomiędzy twórcami a użytkownikami korpusu. Korzystającym z części polsko-czeskiej czy czesko-angielskiej zależy na tym, aby twórcy poświęcili jak najwięcej uwagi tej konkretnej parze języków, tę część rozbudowywali i doskonalili. Twórcy natomiast chcą uwzględnić w korpusie jak najwięcej języków. Z puntu widzenia użytkowników to zabieg mniej ważny, z punktu widzenia twórców to działanie przyszłościowe. Zarówno użytkownik korpusu, jak i jego twórca, znajdują się w sytuacji pomiędzy tym, co mogą i tym, co by chcieli – między swoistym młotem i kowadłem.

Keywords: parallel corpus, Polish, Czech, comparative studies, lexical equivalents
Słowa kluczowe: korpus równoległy, język polski, język czeski, badania komparatywne, ekwiwalenty leksykalne

1. Introduction

The aim of this paper is to confront expectations of users of a multilingual parallel corpus with the potential available to corpus compilers. The idea arose from discussions of the first two co-authors as corpus users with several compilers of *InterCorp*,[1] especially with the third co-author. These discussions mainly arise from the fact that the corpus compilers' efforts (aimed, i.a., at a steady growth of text volumes and improvements in corpus search tools) do not quite meet users' specific research needs. Our comments are presented from two points of view: the compilers' perspective (Section 2) and the users' perspective, based on comparative analyses and translatological studies (Section 3).

1 For more details about *InterCorp* see Rosen, this volume.

2. Problems faced by *InterCorp's* compilers

InterCorp was born with the aim to provide software infrastructure, know-how and some managerial and financial support for linguistic departments at Charles University's Faculty of Arts interested in building parallel corpora suited to their needs and preferences. The principle of subsidiarity was at its foundations: at first, the project consisted of a set of unconnected parallel texts in Czech and a foreign language, collected and built to a large extent by the departments, who were responsible for most tasks of the workflow, including the choice of texts to be included in the corpus.

Even after its integration into a single, on-line searchable corpus with shared formats, pre-processing workflow and tools, the birthmarks of *InterCorp* are still visible. In addition to the distributed mode of building the corpus, it represents a general pragmatic approach to corpus design:[2]

- a sub-optimal variety of texts in the corpus, mainly across but also within the individual languages, due to the individual preferences of the coordinators for a specific language, but also to the lack of suitable translations from or into a given language
- large differences in volume, due mainly to the availability of texts for a given language, but also to the availability and research priorities of the coordinators
- an opportunistic approach to the choice of methods and tools used for building the corpus
- preference for fiction as the source of the richest and most diverse language

In the following sub-sections, we focus on the constraints faced by *InterCorp*'s compilers given the (real or expected) complaints of corpus users listed below:

1. content – inadequate representation of texts with certain properties (originals/translations, genres, authors, translators)
2. size – insufficient volumes of texts
3. searching – missing or unintuitive features of the search interface
4. segmentation, alignment and typos – typos and errors in sentence segmentation and alignment
5. annotation – faulty, inconsistent, unintuitive linguistic annotation, incompatible across languages, including tokenization

2 For a discussion concerning the design of *InterCorp*, including the idea that comparisons with other languages, preferably based on a parallel corpus, are very useful even for monolingual research, see Čermák and Rosen (2012).

2.1 Content

The content is largely determined by the project goals, the availability of texts and time/manpower/financial constraints. Another factor to consider is whether to include only copyright-free texts or rather to prevent a misuse of copyrighted texts by technical means. If a parallel corpus is to include contemporary fiction, the answer must be the latter option. Especially for some less common language pairs, a pragmatic – rather than principled – decision is also necessary in the choice of texts. However, some representative mix of genres, periods, originals/translations, authors, or even translators is needed for both contrastive and translatological studies. Facing the elusive ideal of a balanced parallel corpus, the solution could be custom-created, ad-hoc but reproducible subcorpora, drawn from a pool of all available texts, possibly with a few ready-made selections.

Concerns about the contents of *InterCorp* have recently led to a revision of the policy for including new texts. If only experts for a given language decide, the common goal of a single multilingual corpus with a substantial shared and representative core is hard to achieve. Moreover, a text may not be a priority from the perspective of the language of the original, yet it is desirable to have its original in the corpus. On the other hand, the project management lacks the expertise to decide about the specific literature and research needs. So the new policy is a compromise: proposals for new texts by the experts are submitted each year with two priority levels and reviewed by the corpus management. The criteria for the final approval are as follows:

1. The original of the text is present in the corpus or is already included in the plan. If not, the coordinator for the language of the original is encouraged to include texts that are not of her immediate interest. This has recently been the case of texts such as Hemmingway's *Farewell to Arms*, Kerouac's *On the Road*, Styron's *Sophie's Choice* or Pasternak's *Doctor Zhivago*.
2. The text is important for the language, as shown by the assigned priority.
3. The text does not exceed the limit of new texts per year for the language.
4. The text is already included in the corpus in multiple other languages.
5. The text adds to the diversity of the corpus.

In the first round of this selection process, more than 200 texts in 16 languages were proposed, 60% with high priority. About 85% were approved, the rest put on the waiting list, mainly because of the original text missing.

2.2 Size

Even with the rapidly rising volumes of all bi- and multilingual resources on the web, parallel corpora will always be lagging behind monolingual corpora in size. So it seems that "the more the better" is the right approach. Indeed, in *InterCorp* the numbers may still be too low, especially for lexically more specific studies or less frequent syntactic structures. This applies even to the best-represented languages such as German or Spanish. While the Czech version is available for all texts, the situation is much worse for language pairs not including Czech or for more than two languages. The overlap of Polish and English in the core part is 5.2 million tokens, as opposed to 21.7 million tokens (17.5 million words) in the Polish core or 18.3 million tokens (15.5 million words) in the English core alone.[3] On the other hand, there are reasons why the hunger for ever more words should be kept under control, and these are quality concerns. This applies especially to some freely available multilingual sources, which may include texts that are flawed in both formal and content-related ways, such as garbled character encoding, tokenization or segmentation, as well as duplicated texts, pieces of text in a foreign language, suboptimal and/or unidentified choice of translation.[4]

Let us look more closely at the statistics for Polish in *InterCorp*. With 17.5 million words in the core, and 79 million words in total (including all available collections of texts except *Project Syndicate*) it belongs to the best-represented languages in *InterCorp*. The Polish part of the core includes 232 texts, 18% of the total of 1282 texts in the Czech part. Polish originals (50) slightly prevail over Czech originals (38), the rest are translations from a different language. Table 1 shows that a Polish version is always available among the 27 texts present in 15 or more languages. The last row shows that 110 Polish texts have at least 4 counterparts in a different language.

Languages available	**Texts available:**	**Texts including Polish available:**
≥ 20	≥ 9	≥ 9
≥ 15	≥ 27	≥ 27
≥ 10	≥ 55	≥ 47
≥ 5	≥ 186	≥ 110

Table 1. Texts available in many languages in *InterCorp*

3 Unless specified otherwise, all figures here and below are from *InterCorp* release 8.

4 See, e.g., *A massively parallel corpus: the Bible in 100 languages* (http://christos-c.com/bible/), which does not provide metadata about the translations and sometimes picks dated or less widely known translations, such as *Bible kralická* for Czech (http://gospelgo.com/u/czech_bible.htm), or *Biblia Gdańska* for Polish: (http://biblehub.com/pol/).

Table 2 shows the growth of the Polish part of *InterCorp* across the successive versions in the context of other languages. Perhaps most telling is the comparison with the average size for a foreign language. An average foreign language is outnumbered by a rising factor starting from the first release. Figures 1-3 highlight some of the developments in the corpus size.

	Release date	Foreign core	Foreign total	Czech core	Czech total	Polish core	Polish total	Foreign avg core	Foreign avg total
v0	11/08	25.138	25.138	22.924	22.924	2.066	2.066	1.323	1.323
v1	04/09	34.464	34.464	27.427	31.927	2.244	2.244	1.723	1.723
v2[5]	10/09	39.826	49.293	33.503	35.077	2.422	2.422	1.896	2.347
v3	02/11	62.813	72.280	39.766	41.340	4.716	4.716	2.855	3.285
v4	09/11	71.479	92.290	43.207	46.196	5.462	6.173	3.249	4.195
v5	06/12	91.528	542.640	52.651	75.926	8.396	29.571	3.390	20.098
v6	04/13	138.779	867.287	61.962	99.547	12.710	47.640	4.477	27.977
v7	12/14	173.225	1390.105	77.122	165.425	16.009	77.683	4.559	36.582
v8	05/15	194.055	1423.098	84.718	174.364	17.516	79.905	5.107	37.450

Table 2. A history of *InterCorp* in millions of words [5]

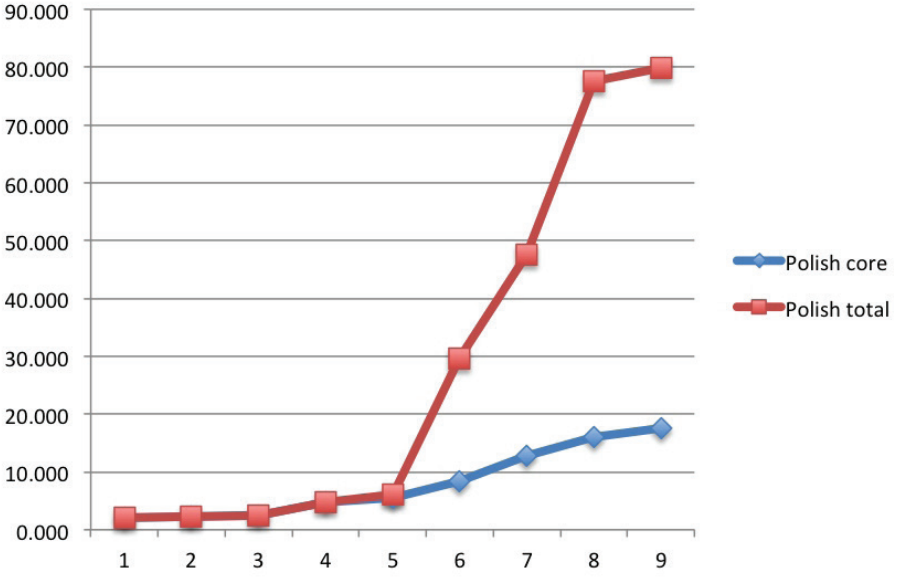

Figure 1. The growth of the Polish part of *InterCorp* from release 0 to 8

5 Some of the figures for v2 are estimated.

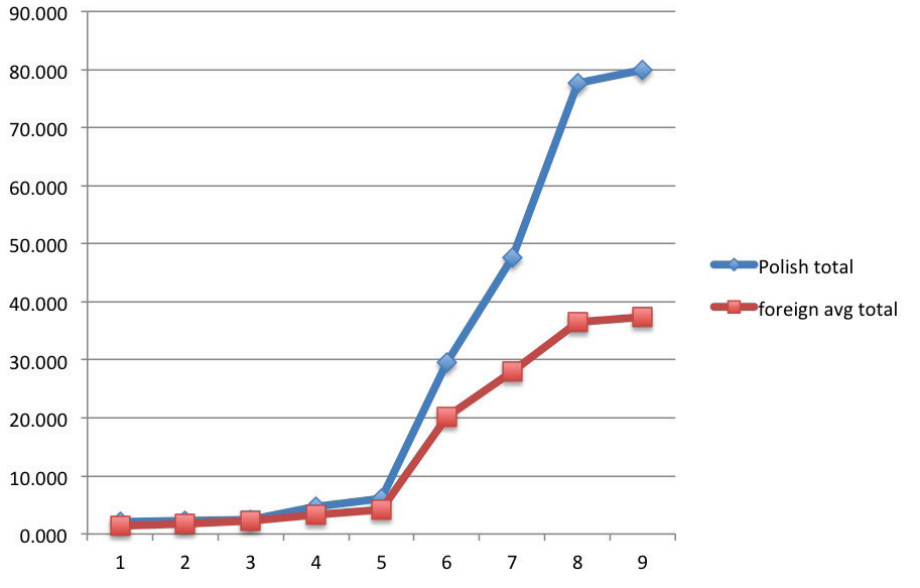

Figure 2. The growth of the Polish core, compared to an average foreign language core

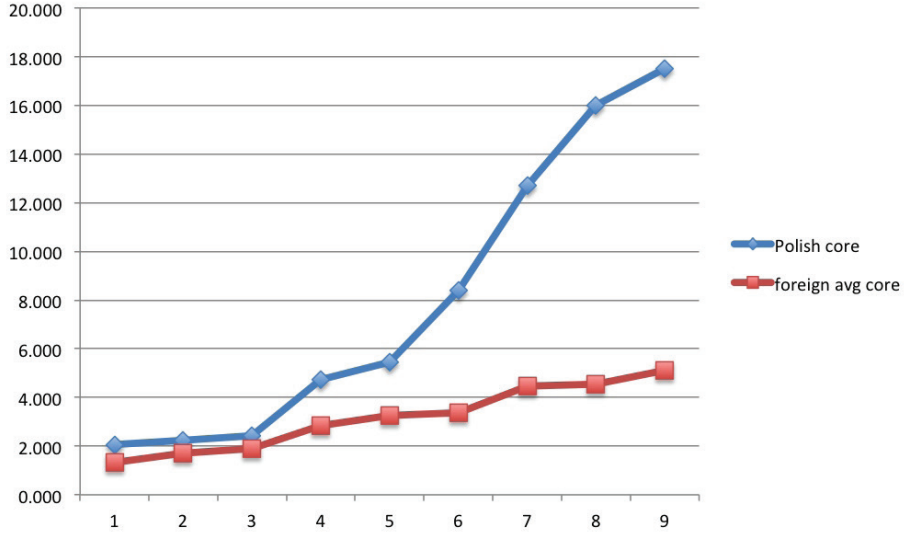

Figure 3. The growth of the Polish total, compared to an average foreign language total

2.3 Searching

There are a number of issues related to the specific concordancer and the search interface that are currently used to search *InterCorp*. Most of the issues will be resolved in their future release, depending on their priority status and the amount of effort necessary to fix them. The following list shows some of the issues waiting for a solution. Hopefully, by the time the list reaches the Reader, some items may no longer be relevant.

- The present *InterCorp* infrastructure cannot accommodate multiple translations in a single language. This is an obvious drawback, especially for users interested in translational research.
- The *biKWiC* feature, highlighting the keyword equivalent in the other language, is missing. Ideally, it could be based on word-to-word alignment, but solutions are available identifying the most likely keyword equivalent from the set of current concordances.
- Although the corpus data include information on the alignment geometry (1:1 / 2:1 / 1:2), neither a flag on whether the alignment has been checked by a human, nor an alignment confidence score, produced by the automatic aligner, can be used and/or displayed while interacting with the corpus using the concordancer.
- It is not possible to create a subcorpus of Czech (the pivot language) including only documents aligned with a specific language. The Czech part of the corpus is restricted to alignments with another language in the query interface, but statistics such as items per million (ipm) relate to the whole corpus of Czech.
- Context-based help on morphosyntactic tags is only available for positional tagsets and currently implemented only for the Czech tagset.
- The user, especially a novice, might appreciate more help or alerts, such as pie charts showing the setup of the selected corpus (users are often unaware of the pitfalls of using a skewed corpus), a list of sample queries, keyboard shortcuts, more context help, including help on text type codes, display of the tag and the lemma of a word below the pointer (mouse hovering), or automatic switching to CQL type query when typing a character such as [.
- Other options beyond mere search would be welcome, such as comparisons across text types, languages and corpora, or collocational profiles, both monolingual and contrastive.[6]

6 See Kilgarriff et al. (2014), Belica (2014), Pęzik (2014), Baisa (2014).

2.4 Segmentation, alignment, typos

The texts in the core part of *InterCorp* are proofread for typos, sentence boundaries and alignment. The results will never be 100% error-free, but errors should be rare and their reporting or flagging in order to crowd-source improvements is now partially implemented in the search interface. On the other hand, the collections are released without human intervention. Rosen and Vavřín (2012) report that in a sample of about 2×180 thousand sentences the number of misaligned segments was at most 8.1% and the percentage of wrongly assigned sentence boundaries was at most 2.9%, while some cases of wrongly identified sentence boundaries actually lead to the misalignments. The percentage of sentences including typos and similar errors was estimated at maximum 3.1%. The figures depend on the type of text, but misalignments, wrong sentence boundaries and typos do not seem to represent a major concern, except in special cases, such as in some novels by Bohumil Hrabal, abounding in long sentences, sometimes spanning over several pages.

2.5 Annotation

Unlike in monolingual corpora, the precision of morphosyntactic tags and lemmas assigned to the tokens is not the main concern in parallel corpora. It is rather the diversity of language-specific tagsets and tokenization rules. Both may be different even for closely related languages, such as Polish and Czech: contractions can be split or left intact, POS classification may be based on morphological or syntactic priorities or represent a parochial view, the format of tags may be very different and confusing to a novice's eye.

The corpus would be limping without another important part of annotation – complete and correct metadata. Omissions and errors hamper filtering of texts for queries and subcorpora as well as providing precise information about concordance sources. Although they are the responsibility of the language coordinators, a bulk of metadata has been corrected and complemented centrally for release 8.

The present priority is to extend morphosyntactic annotation to as many languages as possible. This is the main reason why the corpus does not offer any syntactic annotation at the moment.

2.5.1 Tokenization

Some queries may not return expected results due to language-specific tokenization.[7] Some taggers are based on specific assumptions about contracted

7 For an overview of issues and a solution to conflicting tokenization see Chiarcos et al. (2012).

or hyphenated items[8] such as French |*aux*|, |*dit*|*-il*|, |*cure-dents*|; English |*ca*|*n't*|, |*I*|*'m*|, |*children*|*'s*|, |*parents*|*'*|; German |*zum*|, |*deutsch-französisch*|, |*Jelzin-Ära*|; Polish |*na*|*ń*|, |*że*|*by*|*śmy*|, |*niemiecko*|*-*|*rosyjski*|, |*ty*|*ś*|, |*zrobiłe*|*ś*|; Czech |*padne*|*-*|*li*|, |*Tchaj*|*-*|*wan*|, |*naň*|, |*abychom*|, |*tys*|, |*udělals*| or even about space-separated multi-word items, such as Spanish |*Estados Unidos*|, |*a lo largo de*| or |*al mismo tiempo*|. Note that cognates or similar phenomena often receive different tokenization across the languages. Hyphenated compounds are treated as a single unit in Bulgarian[9] |*Avstro-ungarski*$_{\text{A-pi}}$|, Dutch |*Frans-Duitse*$_{103}$|, English |*Franco-German*$_{\text{NP}}$|, French |*franco-allemande*$_{\text{ADJ}}$|, German |*deutsch-französisch*$_{\text{ADJA}}$|, Italian |*franco-tedesco*$_{\text{ADJ}}$|, and Spanish |*franco-alemana*$_{\text{NC}}$|, but not in Czech |*francouzsko*$_{\text{A2--------A}}$|*-*$_{\text{Z:}}$|*německý*$_{\text{AANS3----1A}}$|, Hungarian|*angol*$_{\text{ADJ}}$|*-*$_{\text{PUNCT}}$|*japán*$_{\text{ADJ}}$|,Polish|*niemiecko*$_{\text{adja}}$|*-*$_{\text{interp}}$|*rosyjski*$_{\text{adj:sg:nom:m1:pos}}$| and Russian | *franko*$_{\text{Ncmsny}}$|*-*|*germanskij*$_{\text{Afpmsnf}}$|.

Within a language, the treatment of hyphenation is fairly consistent. The German and French taggers prefer not to split: |*Jelzin-Ära*$_{\text{NN}}$|, |*gut-ausgearbeiteten*$_{\text{ADJA}}$|, |*cure-dents*$_{\text{NOM}}$|, unlike the Czech tagger: |*padne*$_{\text{VB-S---3P-AA}}$|*-*$_{\text{Z:}}$|*li*$_{\text{TT}}$|, |*Tchaj*$_{\text{AAXXX----1A}}$|*-*$_{\text{Z:}}$|*wanu*$_{\text{NNIS2-----A}}$|. Yet, care must be taken in specific cases, as in the following German and French examples: |*Rechts-*$_{\text{TRUNC}}$| |*und*$_{\text{KON}}$| |*Entwicklungsbewegung*$_{\text{NN}}$|, |*dit*$_{\text{VER:pres}}$|*-il*$_{\text{PRO:PER}}$|.

Tokenization of strings including an apostrophe may not be straightforward either: |*children*$_{\text{NNS}}$|*'s*$_{\text{POS}}$|, |*parents*$_{\text{NNS}}$|*'*$_{\text{POS}}$|, |I$_{\text{PP}}$|*'m*$_{\text{VBP}}$|, |*ca*$_{\text{MD}}$|*n't*$_{\text{RB}}$|.

In some cases, even contiguous strings of alphabetic characters are split and each part is assigned a tag and lemma of its own. This is what happens to Polish (orthographic) words with the agglutinative auxiliary attached, as in *zrobiłeś* '(you) made': |*zrobił*$_{\text{zrobić/praet:sg:m1:perf}}$|*eś*$_{\text{być/aglt:sg:sec:imperf:wok}}$|. A single orthographic word such as *żebyśmy* 'that we would' is split into three parts: |*że*$_{\text{że/conj}}$|*by*$_{\text{by/qub}}$|*śmy*$_{\text{być/aglt:pl:pri:imperf:nwok}}$|.[10]

On the other hand, Czech enclitic *s* as a second person singular auxiliary, spelt together with the preceding form, is treated on a par with inflectional endings. An orthographic concatenation of an l-participle with

8 Vertical bar in the examples indicates token boundaries, as determined by the tokenizers bundled with taggers currently used in *InterCorp* for the given language.

9 The examples are followed by subscripts indicating morphosyntactic tags.

10 A single orthographic word can have different interpretations depending on the way it is tokenized. The form *miałem* can be tagged either as |*miałem*$_{\text{miał/subst:sg:inst:m3}}$| 'dust' or |*miał*$_{\text{mieć/praet:sg:m1:imperf}}$|*em*$_{\text{być/aglt:sg:pri:imperf:wok}}$| 'had'. Similarly with |*gdzieś*$_{\text{gdzieś/qub}}$| 'somewhere' or |*gdzie*$_{\text{gdzie/qub}}$|*ś*$_{\text{być/aglt:sg:sec:imperf:nwok}}$| 'where have (you been)'. Unfortunately, the tagger's choice is not reliable and the present version of the corpus manager cannot see the original orthographic words. This means that searching for such words may involve more than one attempt – a query for its non-split version and another one for its split version.

enclitic auxiliary *udělals* '(you) made' is tagged as a single form of the l-participle |*udělals*$_{\text{udělat/VpYS---2R-AA}}$| (2nd person singular masculine, past tense, affirmative, active voice). The complementizer + enclitic auxiliary *žes* 'that (you) are' is tagged as subordinate conjunction in 2nd person singular |*žes*$_{\text{že/J,-S---2}}$|. However, the second person singular pronoun *ty* is specified for person even without the clitic |*ty*$_{\text{ty/PP-S1--2}}$|, so the form with the clitic attached is distinguished by additional specifications for tense, polarity and voice, irrelevant for either the pronoun or the clitic auxiliary |*tys*$_{\text{ty/PP-S1--2P-AA}}$|. The German and French contractions of preposition and article (*zum*, *aux*) are similar examples of the same phenomenon.

A single token can be searched using any of the query types (Basic, Lemma, Phrase, Word Form, Character or CQL). However, when querying for *Estados* only the Character query type would show all occurrences of *Estados Unidos*. To search only for *Estados Unidos*, the two words should be treated as a single token. The opposite is true about contractions with internal token boundaries: a query for *can't*, *żebyśmy*, *padne-li* etc. must treat the strings as two or more tokens, i.e. as the Phrase or CQL query type, including the internal boundary identified by space in Phrase, i.e. as *can 't*, *że by śmy*, *padne - li*.

This snag is not present in the Poliqarp search engine, used in the *National Corpus of Polish*: the query for *nań* or *na ń* gives the same result.[11] The concordancer currently used for searching *InterCorp* does not allow for this option, i.e. for distinguishing between the two levels of tokenization (orthographical and morphological/syntactical).

2.5.2 Morphosyntactic tags

Currently (in release 8), word forms in 21 languages (including Czech) are assigned morphosyntactic tags while 18 of them are also lemmatized. The language-specific tools (morphological analyzers, taggers, lemmatizers) have been acquired ready-made, trained elsewhere on a language-specific tagset. Each of the tools may thus represent a different conceptual and practical solution to lemmatization, patterning of word classes and morphological categories. While some of the decisions reflect real contrasts between individual languages, other show differences in theoretical backgrounds and formal approaches.

Table 3 below compares the annotation of a sample prepositional phrase such as *in the best apartments* across some of the available languages.

11 See Przepiórkowski et al. (2004). However, it seems that only agglutinative forms of *być* allow for this choice. Contractions such as *żeby*(*śmy*) and *niemiecko-rosyjski* are only found when entered as multiple tokens.

Language	Preposition	Determiner	Adjective	Noun
Bg	R	Pde-os-n	Ansi	Ncnsi
Cs	RR-6	PDXP6	AAFP6---3A	NNFP6---A
De	APPR	ART:Def:Dat:Pl:Fem	ADJA:Sup:Dat:Pl:Fem	N:Reg:Dat:Pl:Fem
En	IN	DT	JJS	NNS
Es	PREP	ART	NC	ADJ
Et	P--s3		A-p-s3	Nc-s3
Fi	Adv:Up		A:Pl:Ine:Foc_kin:Superl	N:Pl:Ine
Fr	PRP	DET:ART	ADJ	NOM
Hu	ART	ADJ	ADJ	NOUN(CAS(ILL))
Is	Aþ	Favfþ	Lvfþvf	Nvfþ
It	PRE	PRO:demo	NOM	ADJ
Lt	Prln	Jvrd	Bdvr	Dktv
Nl	600	370	103	000
No	Prep	Det	Adj	Subst
Pl	prep:loc:nwok	adj:sg:loc:m3:pos	adj:sg:loc:m3:pos	subst:sg:loc:m3
Pt	SPS	DA0	NCFS	AQ0
Ru	Sp-l	P—pl	Afp-plf	Ncmpln
Sk	Eu6	PFfs6	AAfs6x	SSfs6
Sl	Sl	Pd-nsg	Agpfsg	Ncnsl
Sv	PP	DT:UTR:PLU:DEF	JJ:POS:UTR:PLU:DEF:NOM	NN:UTR:PLU:IND:NOM

Table 3. A prepositional phrase annotated by different tagsets

The notational diversity may obscure the fact that even if the tags are translated into a uniform set of labels, some of the seemingly corresponding labels have mismatching denotations. Two corresponding tags can share only a part of their denotations, as in Table 4.

Czech	v	těch	nejodlehlejších	Zástavbách
	RR—6	**PD**XP6	AAFP6----3A	NNFP6-----A
Polish	w	tym	wspaniałym	Apartamencie
	prep:loc:nwok	**adj**:sg:loc:m3:pos	adj:sg:loc:m3:pos	subst:sg:loc:m3

Table 4**.** Partial overlap – Czech PD vs. Polish adj

Czech *těch* 'those' is tagged as a demonstrative pronoun, undistinguished between attributive and substantive use, unlike Polish *tym* 'that', which is tagged as a form of adjectival declension.

In contrast to the Czech tagsets, distinctions in the Polish IPI PAN tagset are based on inflectional classes (Przepiórkowski, Woliński, 2003). Thus the two tagsets, designed for the two closely related languages, have a very different concept of word class, with the Czech tagset closer to the traditional view and mostly more fine-grained and the Polish tagset better defined but lacking some distinctions.[12]

12 The original Polish tagset has been slightly modified for the *National Corpus of Polish* – see Szałkiewicz and Przepiórkowski (2012) or http://nkjp.pl/poliqarp/help/en.html [accessed 21 February 2016].

A Polish adjective (*dziewiąta*$_{\text{adj:sg:nom:f:pos}}$ 'ninth') may correspond to a Czech ordinal numeral (*devátá*$_{\text{CrFS1}}$ 'ninth'), possessive pronoun (*swoje*$_{\text{adj:pl:acc:m3:pos}}$ – *svoje*$_{\text{P8XP4}}$ 'his/her/its/their'), demonstrative pronoun (*temu*$_{\text{adj:sg:dat:m1:pos}}$ – *tomu*$_{\text{PDZS3}}$ 'that'), or relative pronoun (*który*$_{\text{adj:sg:nom:m1:pos}}$ – *který*$_{\text{P4YS1}}$ 'which'). For examples with some context see (1) – (4).

(1) ordinal numeral or adjective?
cs: *devátá*$_{\text{CrFS1}}$ *hodina*$_{\text{NNFS1}}$
pl: *dziewiąta*$_{\text{adj:sg:nom:f:pos}}$ *godzina*$_{\text{subst:sg:nom:f}}$

(2) possessive pronoun or adjective?
cs: *svoje*$_{\text{P8XP4}}$ *rysy*$_{\text{NNIP4}}$
pl: *swoje*$_{\text{adj:pl:acc:m3:pos}}$ *cechy*$_{\text{subst:pl:acc:f}}$

(3) demonstrative pronoun or adjective?
cs: *tomu*$_{\text{PDZS3}}$ *poručíkovi*$_{\text{NNMS3}}$
pl: *temu*$_{\text{adj:sg:dat:m1:pos}}$ *porucznikowi*$_{\text{subst:sg:dat:m1}}$

(4) relative pronoun or adjective?
cs: *který*$_{\text{P4YS1}}$ *vyvěsil*$_{\text{VpYS---XR-AA}}$ *prapor*$_{\text{NNIS4-----A}}$
pl: *który*$_{\text{adj:sg:nom:m1:pos}}$ *wywieszał*$_{\text{praet:sg:m1:imperf}}$ *flage*$_{\text{subst:sg:acc:f}}$

A Polish tag for non-inflected words may correspond to a Czech tag for particles (*nie*$_{\text{qub}}$ *tylko*$_{\text{qub}}$ – *ne*$_{\text{TT}}$ *jen*$_{\text{TT}}$ 'not only'), non-gradable adverbs (*wtedy*$_{\text{qub}}$ – *tenkrát*$_{\text{Db}}$ 'then'), reflexive pronouns (*się*$_{\text{qub}}$ – *se*$_{\text{P7-X4}}$ 'himself/herself/itself/themselves'), subordinating conjunctions (*kiedy*$_{\text{qub}}$ – *když*$_{\text{J,}}$ 'when'), or coordinating conjunctions (*czy*$_{\text{qub}}$ – *nebo*$_{\text{J^}}$ 'or').

Some categorial distinctions are ignored or reflected only implicitly in the tagset. The Prague tagset implicitly marks reflexivity in personal pronouns such as *sobě* 'himself/herself/itself/themselves' (P6-X3) and reflexivity plus possessivity in possessive pronouns such as *svůj* 'his/her/its/their' (P8IS1), while the Polish IPI PAN tagset treats the corresponding forms either as a specific class – siebie:dat for *sobie* 'himself/herself/itself/themselves' – or as a syntactic word class – adj:sg:nom:m1:pos for *swój* 'his/her/its/their'.

Mismatching tagsets could be harmonized by providing a single tagset as in *Multext-East* (Erjavec, 2010), or by using an intermediate taxonomy (Zeman, 2010; Nivre, 2015). Ideally, the task of dealing with multiple tagsets should be delegated to an abstract ontology of linguistic categories (Chiarcos et al., 2012), with mismatches between tags properly represented. This would allow for a principled mapping strategy between languages-specific tagsets, and for intuitive and underspecified queries.

3. Users' problems

For many users, the main problem is a transfer of habits acquired from work with a monolingual corpus to the parallel corpus. This concerns expectations of the users, accustomed to specific software, annotation, research methodology and larger amounts of stylistically more varied material. All of the listed features often result in a disappointment when working with *InterCorp*. This disappointment pertains especially to the low number of corpus occurrences, a restricted choice of research topics, and unsatisfactory research results.

Regarding the corpus research methodology, it is particularly important to be aware of the direction of translation, to realize the potential differences in the notation and linguistic theory behind the tagsets (e.g. Polish adjectives are not the same as Czech adjectives), and to be aware that quantitative methodology cannot be applied, as *InterCorp* is not a reference corpus.[13] Ignoring the direction of translation is one of the problems resulting in incorrect findings and conclusions (cf. Nádvorníková et al., 2010). This is confirmed by recent user access statistics: many users seem to prefer the size of the corpus to an appropriate specification of texts to be queried, including the direction of translation (see Sub-section 3.5 below). Similarly, incorrect identification of a part of speech or a grammatical category or a failure to apply an appropriate methodology may produce results which are misleading or at least not representative.

3.1 Content

From the users' perspective, the content of the Polish-Czech parallel corpus is far from perfect. While the core is mostly hand-corrected fiction, the rest of the corpus consists of collections of automatically processed texts (*Acquis*, *PressEurope*, *Europarl*, *Open Subtitles*). Paradoxically, the texts that are less problematic for the corpus builders are less useful for corpus users.

The automatically processed texts, which allow for rapid extension of the corpus size, are not very useful for the type of research described below in Sub-section . The Polish-Czech parts of the *Acquis*, *PressEurope* and *Open Subtitles* do not include any texts with Polish or Czech specified as the source language. In our translatological studies, where the goal was to find translation equivalents of specific words, multi-word expressions, and selected syntactic constructions from Czech/Polish into Polish/Czech, texts unspecified for the source language cannot be used.

13 Although all *CNC* corpora are now described as reference corpora, a part of them, including *InterCorp*, does not comply with some standard definitions of such corpora, which require that they are representative and balanced.

However, this does not mean that texts where none of the investigated languages is the original cannot be used for other tasks. In an attempt to find how nouns denoting 'the English' and 'the Vietnamese' are translated from English into Polish and Czech, a pilot probe into *Open Subtitles* has shown remarkable results. Polish translations included many more pejorative names for nationalities than Czech translations. An unmarked lexeme denoting a Vietnamese or Japanese person in Czech was often translated into Polish by offensive words.

(5) pl: **Żółtki** będą w was naparzać.
cs: **Japonci** na vás budou střílet.

(6) pl: Nie, lubiła maklerów i **żółtków**.
cs: Ne, jela po makléřích a **Číňanech**.

3.2 Size

Insufficient volumes of available texts are the main problem not only for the corpus compilers, but also for the corpus users. Although Polish belongs to the best-represented languages in *InterCorp*, results obtained from the Polish-Czech part may not be representative enough. The range of topics is limited, so before a real start, the researcher should probe the corpus. Our experience shows that some research topics run into a dead end due to insufficient evidence. Researchers should treat results with caution especially in domains where errors in translations, such as those due to false friends, are more likely. For instance, for cs. *frajer* – pl. *frajer* it is impossible to establish a Czech equivalent (see Table 5).

(pl) *frajer*	12
blbec	2
blbeček	1
chlápek	1
hošánek	1
trouba	1
frajer	3
error	3

Table 5. The equivalents of the Polish lexeme *frajer* in the Czech part of *InterCorp*

In the Polish-Czech part of *InterCorp* (the core), we found 12 examples of the Polish word *frajer* 'a loser'. From their analysis alone appropriate equivalents cannot be identified: the number of occurrences is too small, so the relative frequencies of the equivalent pairs are not conclusive. Moreover, the same word

in Czech has the opposite meaning (an elegant man / boy). Translators of texts included in *InterCorp* did not avoid the trap in three cases, where Polish *frajer* is rendered into Czech as its false friend *frajer.*

A similar problem occurs in translations including the orthographic variants of *džez* vs. *jazz.* Research shows that Czech forms including *dž* occur more often than their parallels including *dż* in Polish. It would be interesting to see how the Czech words including *dž* are translated into Polish. However, the lack of sufficient occurrences does not allow for a conclusive answer. Still it is worth noting that available occurrences show that Czech *džez* is translated into Polish *jazz* (see Hebal-Jezierska, 2013) On the other hand, a similar investigation of *džudo* vs. *judo* stumbled over the problem of insufficient occurrences.

Another example concerning insufficiently representative results is related to translations of the names of nationalities (see (5)(6) pl: Nie, lubiła maklerów i żółtków.(6)Insufficient volumes of available texts are the main problem not only for the corpus compilers, but also for the corpus users. Although Polish belongs to the best-represented languages in InterCorp, results obtained from the Polish-Czech part may not be representative enough. The range of topics is limited, so before a real start, the researcher should probe the corpus. Our experience shows that some research topics run into a dead end due to insufficient evidence. Researchers should treat results with caution especially in domains where errors in translations, such as those due to false friends, are more likely. For instance, for cs. frajer – pl. frajer it is impossible to establish a Czech equivalent (see Table 5).). The question of how the pejorative names for the English and Vietnamese are translated was not answered due to a small number of occurrences. Queries targeting *żółtek* return predominantly homonymous forms denoting genitive plural of 'yolk' rather than the pejorative name for someone of East-Asian origin.

The small number of occurrences also means an increased probability of error. It appears not only in corpus-based translation studies, but in grammar studies as well, e.g. *InterCorp* (release 6) found only 18 occurrences of the structure *toužit* 'to desire' + complement clause.[14] These are not sufficient data for any analysis.

In some cases there is a different situation. For some words the results may be partly sufficient, e.g. establishing equivalents of the Czech verbs *čumět*

14 We analysed the valency of the verb *toužit* and divided the occurrences into groups: *toužit po* + human object (37 occurrances), *toužit po* + abstract object (94), *toužit po* + real object (14), *toužit* + infinitiv (90), *toužit* (*po*) + complement clause (18). The occurrences were excerpted from the Czech-Polish part of *InterCorp* core. (Kaczmarska, Rosen, 2013, 2015; Kaczmarska et al., 2015; Kaczmarska, 2014). In the core of *InterCorp* release 8 restricted to Czech or Polish originals (5,662 thousand tokens), the number of occurrences of the lexeme rose to 27.

and *koukat* (both belong to the semantic field 'to see'). A comparison of their equivalents shows that *čumět* is more often than *koukat* translated by the expressive lexeme *gapić się* 'to stare', while *koukat* is much more often translated by the unmarked lexeme *patrzeć* 'to look'. It is worth noting that the second meaning of *čumět* 'to be stuck' was not distinctive among the obtained equivalents. On the first sight the number of occurrences seems to be sufficient, but the distribution of the Polish equivalents of *čumět* shows that up to 47% of the translation come from Škvorecký's *The Cowards*. After this finding the results were analysed with a greater caution.

Apart from problems with sufficient corpus evidence, various other types of research were successful. The parallel corpus can be helpful for the identification of equivalents of frequent lexemes not only with specific reference to extra-linguistic reality, but also for ambiguous lexemes whose meaning is highly dependent on the context. An example is the Czech word *snad* 'maybe, perhaps', which poses many problems for students of Czech. An analysis of the translations helps to identify the most common meanings, see Table 6.

chyba	29.0%
może	30.0%
pewnie, na pewno	5.0%
przypadkiem	2.0%
zapewne	3.0%
czyżby	2.5%
Other: no equivalent, indeterminacy, syntactic construction	28.5%

Table 6. Polish equivalents of the Czech lexeme *snad*

Establishing equivalents of a selected group of words gives even better results. If the words of choice run into the low frequency problem, the field can be extended. Interesting results were obtained in the analysis of equivalents of expressive words, such as those ending in *-ák* (see Hebal-Jezierska, 2010). The aim of this study was to examine to what extent the translator tries to capture the expressiveness of words ending in *-ák*. Table 7 shows some translations of such lexemes.

It is worth noting that some translations can be simply wrong or the translator's coinages. The Czech word *esesák* 'an SS member'[15] has two Polish equivalents in *InterCorp*: *esesman,* a word well-known to every Pole, and an unexpected form *esesowiec*. Indeed, the corpus shows that *esesowiec* is a nonce word used by a single translator in one text.

15 SS is the abbreviation of *Schutzstaffel*, a powerful paramilitary organization in the former Nazi Germany.

A parallel corpus can help us find suggestions for equivalents of a given word. This is particularly important for ambiguous words. The Czech verb *zdát se*[16] is an opposite example. A traditional dictionary (Siatkowski, Basaj, 2002: 1006–1007) offers four possible Polish equivalents *śnić się, wydawać się, zdawać się*, *podobać się*. The dictionary, however, does not show the context (Kaczmarska, 2012a, 2012b). On the other hand, *InterCorp* found 978 occurrences (release 6, Czech-Polish core, Czech originals)[17] of the verb and its translations into Polish (see Table 8, equivalents related the core meaning of the Czech lexeme are in boldface).

Czech word	Meaning of the Czech word	Polish dictionary translation	Meaning of the Polish translation	Translation found in the corpus
montgomerák	kind of waterproof military coat	*wojskowy płaszcz angielskiego kroju*	military coat of English cut unmarked MWU	
montgomerák	kind of waterproof military coat	*drelich*	denim	change of meaning: an expressive word for a type of material
březňák	kind of wine	*marcowe*	type of wine	change of word form: univerbation by suffix → univerbation by ellipsis
slepák	Appendix	*ślepa kicha*		change of word form: univerbation by suffix → desintegration
blonďák	fair-haired man	*blondynek*	fair-haired (little) man	univerbation by suffix → diminutive suffix
vedlejšák	side job	*chałturka*	diminutive for side job	univerbation by suffix → diminutive suffix
obhroublý dobrák	coarse good man	*dobroduszny grubas*	good fat man	change of meaning (mistake made by the translator): coarse good man → good fat man

Table 7. Equivalents of some Czech expressive nouns ending in *–ák* found in *InterCorp*

Polish equivalents	Number of occurrences	Percentage
wydawać się	**509**	**52.10%**
zdawać się	**190**	**19.42%**
mieć wrażenie	**49**	**5.11%**
wyobrażać sobie	1	
sen / śnić się	**29**	**3.27%**
przyśnić się	1	
przywidzieć się	1	
mieć sny	1	
podobać się	**1**	**0.20%**

16 The tricky Czech verb can be translated into English as: *to seem, to appear, to occur, to dream*.

17 The current version of *InterCorp* (release 8, Czech-Polish core, Czech originals) returns 1433 hits including the lemma *zdát* (*se*).

Polish equivalents	Number of occurrences	Percentage
być zadowolonym	1	
czuć	**3**	**0.92%**
poczuć	2	
doznać uczucia	3	
mieć uczucie	1	
myśleć	**5**	**2.00%**
uznać	1	
mniemać	1	
podejrzewać	1	
pomyśleć	2	
rozumieć	1	
sądzić	4	
uświadamiać sobie	1	
uważać	4	
wyglądać	**34**	**3.99%**
widać	4	
widzieć	1	
okazywać się	**1**	**0.10%**
pewnie	1	
usłyszeć	1	
jakby	1	
Other	**58**	**6.24%**
Error	**21**	**2.15%**
Omitted in translation	**44**	**4.50%**
TOTAL	**978**	**100,00%**

Table 8. Equivalents of the Czech verb *zdát se* in *InterCorp*

The results show that more than half occurrences of the verb *zdát se* are translated into Polish as *wydawać się* (52%), which seems to be the obvious equivalent. Its synonym – *zdawać się* – appears in 19,3% occurrences.[18] The unit *mieć wrażenie* (5%) is semantically close to the two previous Polish verbs, but differs in terms of style. Other possible equivalents found by *InterCorp* can be divided into several groups (see Table 8). These are not straight equivalents of the Czech unit; they emphasize different semantic components, e.g. visual perception (*wyglądać*, *widzieć*, *widać*), intelectual aspect (*myśleć, mniemać, podejrzewać, pomyśleć, rozumieć, sądzić, uświadamiać sobie, uważać, uznać, moim zdaniem*), the emotional element of the meaning (*czuć, poczuć, doznać uczucia, mieć uczucie*), or the component of objectivity and impersonality (*wynikać* i *okazywać się*).[19] As many as 58 occurrences contain other units (*chyba, najwyraźniej,*

18 The verbs *wydawać się* and *zdawać się* constitute 70% occurrences and seem to be absolutely synonymous. It would be worthwhile to consider when (in which contexts) one or the other is chosen. A Polish corpus (http://nkjp.pl) could be used to investigate several factors: the wider context showing the experiencer and the object (name / noun / pronoun [*I / me*]), the type of the text (dialog / narration) and the stylistic layer. The result of such an analysis may be particularly important for translators and foreign learners of Polish.

19 The translators, however, used the verbs only in cases when *zdát se* did not need to be completed by a personal object.

pewnie, prawdopodobnie) as equivalents of the verb *zdát se*, which include elements of epistemic modality[20] – information referring to the way the speakers communicate their judgments, certainties, guesses, doubts. Of course, *InterCorp* also includes evidence of two other meanings of the Czech unit *zdát se*: *śnić się* 'to dream of' (*przyśnić się, przywidzieć się*) and *podobać się* 'to enjoy' (*być zadowolonym*).[21] To conclude, the parallel corpus *InterCorp* is able to present avenues of possibilities for the choice of the proper equivalent in a given context.

Corpus data can be very useful for the identification of the meaning of a structure (or a unit) such as the Czech unit *být líto*. If we use the methods of Pattern Grammar,[22] manual analysis based on *InterCorp* indicated, i.a., two patterns of *být líto* ('to be sorry', 'to regret'), associated with two meanings. If the unit *být líto* is combined with two nominal phrases (Dative and Genitive), it corresponds to the Polish equivalent *żal*. If combined only with the Dative nominal phrase, and possibly with the element *to*, it corresponds to the Polish equivalent (*być*) *przykro*.

żal	*(być) przykro*
Jak mi ho bylo líto!	*Pak mi je líto.*
Jakże mi go było żal!	*Wobec tego, przykro mi!*
Je mi ho samozřejmě líto.	*Potom nám to bylo oběma líto.*
Jest mi go oczywiście żal…	*Potem nam obu było przykro.*
Přišlo mi jí prostě líto.	*…nabídne mi sisinku a já si vezmu, protože by mu bylo líto, kdybych si nevzala…*
Po prostu zrobiło mi się jej żal.	*...zaprasza mnie na cuksa i ja biorę, bo byłoby mu przykro, gdybym nie wzięła…*
být líto **+ NP$_{DAT}$ + NP$_{GEN}$ =** ***żal***	***být líto*** **+ NP$_{DAT}$ + to / Ø = (*****być*****)** ***przykro***

Table 9. The patterns of *být líto* (*żal, być przykro*)

3.3 Searching

The search interface offers the comfort of the same tools, functions, etc., available for searching both monoligual and parallel corpora. The clickable filtering of the texts, based on the metadata, including the translation direction, is also intuitive and useful. There is only one disadvantage. Statistics such as ipm relate to the whole corpus of Czech, rather than to its intersection with Polish.

20 More on the modality: Boniecka (1976), Roszko (1993), Rytel (1982), Wróbel (1991).

21 *Zdát se* as *podobać się* (enjoy) is possible only with the negation:

cs: *Venca se potil, jak ho Fonda nutil, a nutil ho tak, že si musel doladovač trombónu postrčit skoro o decimetr, až už mu to dál nešlo, ale Fondovi se to pořád* ***nezdálo****.*

pl: *Wacek aż się spocił, tak go Fonda piłował, a piłował go tak, że Wacek musiał stroik puzonu przesunąć prawie o dziesięć centymetrów, aż już dalej nie szło, ale Fonda ciągle* ***nie był zadowolony****.*

22 See Ebeling and Ebeling (2013) or Hunston and Francis (2000).

Moreover, a subcorpus of Czech texts aligned with parallel Polish texts cannot be created.

3.4 Segmentation, alignment, typos and annotation

Errors in alignment and sentential segmentation are often related, but they do not pose a significant problem, especially in the proofread core part of *InterCorp*. Most cases of misalignment are easy to spot while parallel concordances are browsed. To read the whole aligned segment is often unavoidable anyway in the language which was not queried. Then the parallel keywords are not highlighted, since there are no word-to-word alignments in *InterCorp* yet.

Misaligned sentences can be recovered in the extended context or even within the same segment, the latter in case of incorrect alignment of multiple sentences within a single segment. Typos are relatively few, except for misplaced pieces of texts in an inappropriate language in collections such as *Acquis*. In comparison to previous releases, metadata are now significantly more reliable, especially in the crucial identification of the language of the original. Unfortunately, the original language is still unknown for many texts in the collections, which is the main reason why some users prefer to query only the carefully annotated core part.

On the other hand, linguistic annotation is not problematic due to its insufficient reliability, but because of the multitude of different tagsets and disparate tokenization rules (see Subsection above). This is clearly one of the main problems facing the user, who is often unaware of the differences in the tags beyond mere superficial notational dissimilarity.

3.5 User access statistics

During the first half of 2015[23] the users of *InterCorp* made 62 thousand queries, including 2 thousand (3.26%) queries with Polish as one of the languages. The most often queried language combination involving Polish was – not surprisingly – Polish and Czech (1.4 thousand queries, 71% of all queries involving Polish). Apart from monolingual queries into the Polish part of *InterCorp* (6.2%), other combinations are far less common: Polish and French (2.8%), Polish and Russian (2.5%), followed by Czech, Polish and Russian (2.2%). Interestingly, most queries (85.6%) target all available texts. Queries restricted to the core account for mere 10%. This is still more than the share of core queries for all languages – 5.7%, compared with queries for all languages unrestricted by the text type – 91.0%. The high numbers of unrestricted queries both for Polish and

23 More precisely within 1 January – 20 July 2015.

for other languages indicate that most users prefer large data to specific text types and that collections play an important role in the corpus. As an additional explanation, at least some users could be suspected of inadvertently ignoring an important methodological aspect, such as distinguishing the direction of translation.

4. Conclusions

The problem of insufficient size and disproportionate representation, felt as important by the corpus users, is quite hard to overcome – translations in the preferred text type may not be available for a given language pair. Could some of the problems be resolved by the use of comparable rather than parallel corpora?

On the other hand, the alignment and annotation problem, which the corpus builders feel is important, does not seem to be a priority for the users, at least not for some users of the Polish-Czech part. This may be different for users interested in multiple languages, posing problems such as less reliable or missing morphosyntactic annotation or incompatible tagsets.

The collections, where the language of the original is very often unknown and very seldom Czech or Polish, do not seem to help much for this kind of research, although *Open Subtitles* was shown to yield interesting results. Could some methods be adapted to the existing resources, even though they are not perfect?

The bottom line points to the importance of user feedback. Even though there is a user forum and an easy way to report problems, to comment, to make wishes, regular users of *InterCorp* have been asked recently to participate in a survey intended to provide a better picture of the users' preferences to guide future steps in the development of the corpus.

References

Baisa, Vít (2014): Parallel corpora in Sketch Engine. Paper presented at the 5th Sketch Engine Workshop. Bolzano, Italy, 14 July, 2014.

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

Boniecka, Barbara (1976): O pojęciu modalności (przegląd problemów badawczych). *Język Polski* LVI(2), 99–110.

Chiarcos, Christian, Ritz, Julia, Stede, Manfred (2012): By all these lovely tokens... merging conflicting tokenizations. *Language Resources and Evaluation* 46(1), 53–74.

Chiarcos, Christian (2012): Ontologies of linguistic annotation: Survey and perspectives. 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), 303–310.

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

Ebeling, Jarle, Ebeling, Signe O. (2013). *Patterns in contrast*. Amsterdam/Philadelphia, PA: John Benjamins.

Erjavec, Tomaž (2010): MULTEXT-East Version 4: Multilingual morphosyntactic specifications, lexicons and corpora. 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), 2544–2547.

Hebal-Jezierska, Milena (2013): Jazz *http://portal.uw.edu.pl/web/approval/jazz_cz* (3 March 2016).

Hebal-Jezierska, Milena: (2010) *Jak se překládají české univerbizáty do polštiny* In: František Čermák, Jan Kocek (eds.) *Mnohojazyčný korpus InterCorp: Možnosti studia*. Praha: Lidové noviny, 261–268.

Hunston, Susan, Francis, Gill. (2000). *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam/Philadelphia, PA: John Benjamins.

Kaczmarska, Elżbieta (2012a): Czeski czasownik „zdát se" w przekładzie na język polski (na podstawie badań z wykorzystaniem czesko-polskiego korpusu równoległego InterCorp). *Studia z Filologii Polskiej i Słowiańskiej* 47, 247–261.

Kaczmarska, Elżbieta (2012b): Searching for equivalents on the basis of a Czech – Polish parallel corpus (the case of the verb „zdát se"). In: Panajot Karagiozov, Kalina Bahneva, Valentin Geshev, Ina Hristova, Margarita Mladenova (eds.): *Време и история в славянските езици, литератури и култури*. Sofia: Езикознание, 238–245.

Kaczmarska, Elżbieta (2014): Czeskie czasowniki oznaczające stany psychiczne – sposoby ustalania polskich ekwiwalentów na podstawie korpusu równoległego InterCorp. In: Anna Stolarczyk-Gembiak, Marta Woźnicka (eds.) *Zbliżenia. Językoznawstwo – Literaturoznawstwo – Translatologia*. Konin: Państwowa Wyższa Szkoła Zawodowa w Koninie, 45–55.

Kaczmarska, Elżbieta, Rosen, Alexandr (2013): Między znaczeniem leksykalnym a walencją – próba opracowania metody ekstrakcji ekwiwalentów na podstawie korpusu równoległego. *Studia z Filologii Polskiej i Słowiańskiej* 48, 103–121.

Kaczmarska, Elżbieta, Rosen, Alexandr (2015): Jak najít optimální překlad polysémních sloves – porovnání metod automatické analýzy paralelních textů. *Časopis pro moderní filologii* 97(2), 157–168.

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 Filologiczne* 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.

Nádvorníková, Olga, Polická, Alena, Šotolová, Jovanka, Vurm, Petr (2010): Využití InterCorpu ve vyskoškolských kurzech francouzské filologie. In: František Čermák, Jan Kocek (eds.) *Mnohojazyčný korpus InterCorp: Možnosti studia*. Praha: Lidové noviny, 232–240.

Nivre, Joakim (2015): Towards a universal grammar for natural language processing. In: Alexander F. Gelbukh (ed.): *Proceedings of Computational Linguistics and Intelligent Text Processing 16th International Conference, CICLing 2015, Cairo, Egypt*, Part I, volume 9041 of Lecture Notes in Computer Science. New York, NY: Springer, 3–16.

Pęzik, Piotr. (2014): Graph-based analysis of collocational profiles. In: Vida Jesenšek, Peter Grzybek (eds.): *Phraseologie im Wörterbuch und Korpus, Proceedings of Europhras 2012.* Maribor: Univerza v Mariboru, 227–243.

Przepiórkowski, Adam, Krynicki, Zygmunt, Dębowski, Łukasz, Woliński, Marcin, Janus, Daniel and Bański, Piotr (2004): A search tool for corpora with positional tagsets and ambiguities. In: Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, Raquel Silva (eds.): *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon: European Language Resources Association (ELRA), 1235–1238.

Przepiórkowski, Adam, Woliński, Marcin (2003): A flexemic tagset for Polish. In: Tomaž Erjavec (ed.): *MorphSlav '03 Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages.* Budapest: Association for Computational Linguistics, 33–40.

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.

Roszko, Roman (1993): *Wykładniki modalności imperceptywnej w języku polskim i litewskim*. Warszawa: Instytut Slawistyki PAN.

Rytel, Danuta (1982): *Leksykalne środki wyrażania modalności w języku czeskim i polskim*. Wrocław: Zakład Narodowy im. Ossolińskich.

Siatkowski Janusz, Basaj Mieczysław (2002): *Słownik czesko-polski*. Warszawa: Wiedza Powszechna.

Szałkiewicz, Łukasz, Przepiórkowski, Adam (2012): Anotacja morfoskładniowa. In: Adam Przepiórkowski, Mirosław Bańko, Rafał Górski, Barbara Lewandowska-Tomaszczyk (eds.): *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN, 59–96.

Wróbel, Henryk (1991): O modalności. *Język Polski* LXXI, 260–270.

Zeman, Daniel (2010): Hard Problems of Tagset Conversion. In: Alex Fang, Nancy Ide, Jonathan Webster (eds.): *Proceedings of the Second International Conference on Global Interoperability for Language Resources.* Hong Kong: City University of Hong Kong, 181–185.