# Building a multilingual parallel corpus for human users

**Alexandr Rosen, Martin Vavřín**

Charles University in Prague, Faculty of Arts
alexandr.rosen@ff.cuni.cz, martin.vavrin@ff.cuni.cz

**Abstract**

We present the architecture and the current state of *InterCorp*, a multilingual parallel corpus centered around Czech, intended primarily for human users and consisting of written texts with a focus on fiction. Following an outline of its recent development and a comparison with some other multilingual parallel corpora we give an overview of the data collection procedure that covers text selection criteria, data format, conversion, alignment, lemmatization and tagging. Finally, we discuss challenges and prospects of the project.

**Keywords:** parallel corpora, multilingual, Czech

## 1. Introduction

Building any parallel corpus, and especially one including many languages, requires an opportunistic approach, because a balanced mix of texts is hard to achieve even for written language. The balance may be a problem especially when the primary users are humans (teachers and students of foreign languages, translators or linguists), rather than applications (e.g., in machine learning). In comparison to monolingual corpora, some types of texts are rarely available as translations (e.g., spontaneous non-literary writing such as personal correspondence). We try to solve this problem by including a substantial share of original and translated fiction, even though it is a more costly alternative to genres freely available on the web. Additionally, languages with fewer native speakers may end up as underrepresented mainly due to a smaller pool of original texts, which also results in a skewed overall balance between originals and translations. A higher share of "institutional" translations in such languages (often provided as mandatory within an organisation or service) makes up for the quantitative loss, but does not make the corpus more balanced genre-wise.

An opportunistic approach is needed also because taggers and other available language-specific tools are not tailored to the needs of a multilingual environment and the development of a single consistent suite covering all our languages is not a realistic target. The issue of a heterogeneous set of linguistic tools can be ignored as long as they produce useful results, but in the long run we have to resolve their conceptual and formal incompatibility.

To produce more reliable corpus data, most of our automatically pre-processed texts are manually checked for typos, segmentation and alignment errors within a dedicated parallel text editor/aligner. Morphosyntactic annotation is provided for languages where taggers or lemmatizers are available. The result is available for on-line searches, and options to provide entire texts in a scrambled format are available to satisfy some application-driven demand.

After an overview of the corpus in the context of some other multilingual parallel corpora in §2. we present a sketch of the pre-processing workflow in §3., before going into some detail about the linguistic mark-up in §4., presenting a sample query via our user interface in §5. and concluding by an outlook in §6.

## 2. An overview of *InterCorp*

*InterCorp* is a parallel synchronous corpus of written language, built since 2005 at the Faculty of Arts of Charles University in Prague. Currently, in its release 4 at the end of the project's first phase (December 2011), it includes 92 mil. tokens in 22 languages plus 46 mil. tokens in Czech. Czech serves as the *pivot*: all 'foreign' texts have their Czech counterparts, while a foreign text may have no counterpart in another foreign language. More languages, more data and more sophisticated and reliable annotation, together with improved user interface and export options, are previewed for the second phase of the project, due to end in 2016.[1]

Unlike other corpora involving many languages but built from open-source data, such as *Opus* (Tiedemann, 2009) and *JRC-Acquis* (Steinberger et al., 2006), but like some other multilingual corpora oriented primarily towards linguists, students and translators as direct users, such as *ParaSol* (von Waldenfels, 2006), *InterCorp* is built around fiction, a genre best approximating the needs of the project participants and most of the users. For a short overview of comparable projects see Table 1 – for each corpus, the Tokens column gives the number of word tokens in Czech texts aligned with English counterparts. Other columns give the maximum number of languages in the corpus, a rough indication of the text types, the presence of morphosyntactic descriptions (MSD tags) or even syntactic annotation, and the option to search the corpus on-line via a web-based user interface.

There are at least two additional projects worth mentioning but not quite fitting the pattern. *Multext-East* in its version 4 (Erjavec, 2010)[2] is still primarily a corpus consisting of a single novel (George Orwell's 1984), but it has paved the way for multilingual corpora including some less common languages – currently it includes 17 languages together with harmonised morphosyntactic specifications and lexica, syntactic annotation, and a small parallel speech corpus. Although not multilingual, *CzEng* – Czech-English Parallel Corpus[3] currently includes a fairly balanced mix of texts of

[1] See http://www.korpus.cz/intercorp/?lang=en for more details about the project and a link to the corpus search interface.

[2] http://nl.ijs.si/ME/

[3] http://ufal.mff.cuni.cz/czeng/czeng10/

a substantial size (206 million Czech tokens), together with syntactic annotation (Bojar and Žabokrtský, 2009). Unfortunately, a representative parallel corpus of such size is not possible for most language pairs including Czech. A comparable corpus could be the way out, but it is not clear to what extent it would satisfy human users, who usually expect parallel concordances.

The corpus is built in a distributed fashion – coordinators for specific languages, mostly members of the participating departments are responsible for the choice and acquisition of texts and provide some pre-processing tasks, including proofreading and checking of the results of automatic sentential segmentation and alignment. The head coordinator – the Institute of the Czech National Corpus (ICNC) – is responsible for overall management and infrastructure, including the central data repository, software tools, automatic alignment, linguistic mark-up and availability of the results. Additionally, some texts are acquired for multiple languages and processed fully automatically. Their current share in the number of tokens is 22.5%, including news and political commentaries from Project Syndicate[6] and Presseurop.[7]

Except for poorly represented languages and several classical authors of 20th century Czech literature, there is a recommendation concerning original texts: titles produced (or first published) after 1945 should be preferred. The inclusion of contemporary translations of older texts is decided ad hoc, depending on the interests of the language coordinator and appreciation of the title by contemporary readers. The initial nearly exclusive focus on novels has been extended to drama and non-fiction, such as journalism, user manuals, legal texts and essays.

The fact that only texts available also in Czech can be included restricts the choice in general, but with the growing size of a truly multilingual core, including texts available in more than very few languages, the corpus is becoming more attractive even for users not interested in Czech.

The figures in Table 2 represent the part of the corpus which is at the moment publicly available for queries through a web-based interface.[8]

The columns Syndicate and Presseurop indicate whether the text resource is available for the language. The number of tokens for Project Syndicate ranges between 2.3 to 3 million (except for a much smaller Italian part, a recent extension) and is about 0.8 million for Presseurop. The last two columns indicate whether tags or lemmas are available.

In addition to access via web-based interface, we are going expand the options of exploiting the corpus by offering entire parallel texts, e.g., for machine learning applications, but in a form that could not be used in violation of copyright laws. The solution is based on a requirement that the licensed texts are used only for non-commercial research and – perhaps more importantly – on a modification of texts that makes the original sequence of sentences inaccessible. Bilingual text units consisting of one or more entire sentences up to a limit corresponding to concordance window (100 words) are sorted in a random order (shuffled).[9] Although this procedure makes the data less useful for the investigation of discourse markers, supra-sentential anaphora, or text cohesion in general, we believe it is still a bearable price for a safe dissemination of aligned segments without any chance to contravene the copyright.

## 3. Pre-processing

An electronic version of a specific text may already be available in the archives of ICNC or elsewhere, including archives of a publishing house. If not, the text is scanned, OCR'd and proofread. Proofread texts (as .doc or .rtf files) are exported from MS Word using a Visual Basic macro into a quasi-XML format. In the first step, paragraph boundaries and typeface marks present in the text are translated into XML tags. Special mark-up characters (&, <, >) are rendered as character entities. In the second step, sentence boundaries are identified. For Czech, we use a rule-based splitter,[10] for other languages a tool based on an unsupervised learning algorithm.[11]

The text is then uploaded to a server, checked for formal consistency and read into *InterText*, a parallel text editor, designed and developed as a part of the project, but useful in other contexts.[12] If the editor already has access to a parallel version of the text, an aligner[13] is used to align the two texts automatically. The alignment can then be checked and corrected, together with any remaining typos and segmentation errors.

Thus, in the standard case of a printed book, a text is checked and corrected at three stages: (i) after OCR by an advanced student of the respective language, (ii) after sentential segmentation and alignment by a coordinator and expert in the language, and, finally (iii) by the chief coordinator who may not know the language but can recognize remaining faults in technical detail and alignment.

Before a new release of the corpus is due, all aligned and approved texts are exported from *InterText*.[14] For each aligned pair of texts a file including the alignment information is generated, referring to sentence IDs within the texts.[15] All along, a database is used for tracking the passage of each text through the pre-processing stages, for recording its status, and for supplying its bibliographical

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

[7] http://www.presseurop.eu/

[8] The parallel search interface *Park*, developed by Michal Štourač, is available after registration at http://korpus.cz/Park.

[9] A similar solution was used by (Varga et al., 2005) in a Hungarian-English parallel corpus (http://mokk.bme.hu/resources/hunglishcorpus/).

[10] Program *tokenize* by Pavel Květoň.

[11] The Punkt sentence tokenizer, in an implementation from http://nltk.org/. See (Kiss and Strunk, 2006, p. 485–525).

[12] See (Vondřička, 2010) and http://wanthalf.saga.cz/intertext.

[13] *Hunalign*, see (Varga et al., 2005) and http://mokk.bme.hu/en/resources/hunalign/.

[14] Texts within *InterText* may be corrected and otherwise modified even after they have been exported. Thus, a text stored in *InterText* need not be identical with its version in the current release of the corpus.

[15] The stand-off alignment format, together with the use of the on-line alignment editor, is the major difference from the workflow used previously, described in (Vavřín and Rosen, 2008). Some texts in the corpus processed in the old way, may include errors due to the previously used technology. Such texts are gradually corrected within *InterText*.

| Name | Languages | Tokens | Text types | Annotation | Web search | Link |
|---|---|---|---|---|---|---|
| Opus | 92 | 32,060 | medical, legal, financial, software, subtitles | syntax | yes | http://opus.lingfil.uu.se/ |
| Glosbe | 29 | 6,465[4] | varia | – | yes | http://glosbe.com/tmem/ |
| JRC-Acquis | 22 | 22,843 | legal | – | no | http://optima.jrc.it/Acquis/ |
| EuroParl v.6 | 21 | 10,574 | parliament proceedings | – | no | http://www.statmt.org/europarl/ |
| ParaSol | 31 | 1,679[5] | fiction | MSD tags | yes | http://parasol.unibe.ch/ |
| InterCorp | 23 | 7,297 | fiction, journalism | MSD tags | yes | http://korpus.cz/intercorp/ |

Table 1: An overview of some parallel multilingual corpora

| **Language** | **Tokens (×1000)** | **Texts** | **Syndicate** | **Press-europ** | **Tags** | **Lemmas** |
|---|---|---|---|---|---|---|
| Bulgarian | 1,135 | 15 | | | ✓ | |
| Croatian | 6,735 | 96 | | | | |
| Danish | 190 | 5 | | | | |
| Dutch | 5,203 | 62 | | ✓ | ✓ | |
| English | 7,297 | 49 | ✓ | ✓ | ✓ | ✓ |
| Finnish | 1,435 | 22 | | | | |
| French | 5,234 | 24 | ✓ | ✓ | ✓ | ✓ |
| German | 12,167 | 125 | ✓ | ✓ | ✓ | ✓ |
| Hungarian | 1,123 | 17 | | | ✓ | |
| Italian | 4,028 | 31 | ✓ | ✓ | ✓ | ✓ |
| Lithuanian | 358 | 17 | | | ✓ | ✓ |
| Latvian | 1,075 | 32 | | | | |
| Norwegian | 2,158 | 21 | | | ✓ | ✓ |
| Polish | 6,173 | 92 | | ✓ | ✓ | ✓ |
| Portuguese | 2,503 | 20 | | ✓ | | |
| Rumanian | 1,697 | 9 | | ✓ | | |
| Russian | 3,619 | 25 | ✓ | | ✓ | ✓ |
| Slovak | 6,961 | 139 | | | ✓ | ✓ |
| Slovene | 992 | 16 | | | | |
| Serbian | 2,736 | 38 | | | | |
| Spanish | 14,237 | 126 | ✓ | ✓ | ✓ | ✓ |
| Swedish | 5,234 | 64 | | | | |
| **TOTAL** | **92,290** | **1,045** | **6** | **9** | **13** | **10** |
| Czech | 46,196 | 703 | ✓ | ✓ | ✓ | ✓ |

Table 2: Figures for the part of *InterCorp* available on-line as of September 2011

data before entering the indexed corpus. The metadata are inserted manually at the start of pre-processing, or – for batch acquisitions – supplied by scripts directly from the source texts.

For languages where the tools are available, the texts can be morphologically tagged and/or lemmatized (see §4.). Finally, the texts are indexed by the corpus manager.[16] The alignment files are processed separately to fit the parallel search interface.

The texts acquired from the web as digital files for multiple languages do not follow the track described above until the tagging phase; their clean-up, segmentation and alignment is fully automatic. Although the result is not manually checked, an evaluation of the tools shows that it is reliable enough to be included in the corpus (see §6. below). Anyway the user can always exclude such texts from searches.

## 4. Linguistic mark-up

Both human users and applications benefit even from a minimal linguistic analysis of the corpus data, although it is not without problems and may even complicate access to the raw text.

At the time of writing, word forms in 14 languages (including Czech) are assigned morphosyntactic tags while 11 of them are also lemmatized (see Table 2 again). The numbers are due to rise in near future.

The application of language-specific tools (tokenizers, morphological analyzers, taggers, lemmatizers) can be seen as an additional example of our opportunistic approach – all of them have been acquired ready-made, trained elsewhere on monolingual data using a language-specific tagset.[17] Each of the language-specific tools may thus represent a different conceptual and practical solution to a number of issues: tokenization, 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. Even closely related languages, such as Polish and Czech, may have very different tagsets and tokeniza-

[16] *Manatee*, see (Rychlý, 2007).

[17] See http://www.korpus.cz/intercorp/?lang=en, the section on morphosyntactic annotation, for more details about the tools.

tion rules: 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 the eye of a novice.

Even a very basic search for a form such as *can't* in the English text ends up in failure, because the tokenizer preceding the English tagger splits the contraction in two words: *ca* nad *n't*. So far, our concordancer has no way of storing both, which would be the optimal solution. This is an issue that must be solved by distinguishing the level of graphical and morphological words, e.g., as in Poliqarp, a concordancer designed for Polish (Przepiórkowski, 2004).

There could also be a way to harmonise the mismatching tagsets, possibly by translating the language-specific tags into a single tagset, as in multilingual projects such as *Multext-East* (Erjavec, 2010) or via a shared taxonomy (Zeman, 2010; Chiarcos et al., 2008; Rosen, 2010). Ideally, the task of dealing with multiple tagsets should be delegated to an abstract interlingual representation of linguistic categories, with mismatches between tags properly represented. This would allow for a principled mapping strategy between languages-specific tagsets, and for intuitive and underspecified queries. Such a solution is previewed for a future version of the corpus and the concordancer, together with syntactic annotation. While incompatible structures across languages would be even more problematic in a parallel treebank, the results of different language-specific parsers must be translated into a common format. On the other hand, this is an easier task than translating MSD tags, because the cross-lingual differences are smaller in syntax than in morphology.

## 5. A sample query

Figures 1–3 show the user interface of our parallel concordancer using a sample query and its result. The screenshot in Fig. 1 shows the process of filtering languages and texts half way through. The user has already ticked three languages (Czech, English and Italian) and the list of available texts in this combination shrunk accordingly to two novels (Milan Kundera's *Immortality* and Joanne Rowling's *Harry Potter and the Philosopher's Stone*) and three collections of news and political commentaries. After an additional ticking next to Bulgarian, the list will shrink even further to a single novel (Kundera's – not shown).

Fig. 2 shows the second step: specifying the query itself. The user is interested in concordances including the specified forms in Czech, Bulgarian and English.

Fig. 3 gives the result – there are four concordances that satisfy the query. The user also changed the default viewing options, choosing to see MSD tags for the keywords.

## 6. Outlook

Even though our primary focus is on human users, they still have some wishes that remain unsatisfied so far, mostly related to the search interface (sorting, statistics). Since a parallel corpus is by its very nature a good match to similar resources, a closer integration with the Czech National Corpus or even with other parallel corpora of Slavic or other languages is an interesting prospect.

As for the content of the corpus, the next release coming soon will include, i.a., all texts from the JRC-Acquis corpus, tagged and lemmatized for all languages where the tools are available. As an additional improvement of balance between languages we plan to include some additional freely available multilingual resources untapped so far, such as the proceedings of the European Parliament. Also, a few additional languages are in the pipeline: Albanian, Arabic, Belorussian, Chinese, Hindi and Romani.

Despite the focus on manual checking, segmentation and alignment will never be perfect. Optimizing the automatic tools (e.g., by providing the aligner with lemmatized texts and bilingual lexica) may bring some improvement, but we also plan to let the users mark errors when they are found, directly in the search interface. As a benefit available in a corpus used mainly by humans, crowd-sourcing may serve not only to correct individual instances, but also as a feedback to help improve tools used in preprocessing.

So far, our assessment of the quality of automatic segmentation and alignment rests on a previous evaluation of the aligner, using a sample of similar texts (Rosen, 2005; Singh and Husain, 2005), logs of corrections in the parallel text editor, and the users' feedback, which is overall positive. However, we plan to perform a more detailed study by comparing results of our automatic segmentation and alignment procedure with manually corrected texts. The results will help to further improve the preprocessing tools.

As a preliminary study, we have made a summary of corrections as logged by the parallel text editor for a sample of 46 titles, covering 11 languages, automatically segmented and aligned with Czech. The sample includes 182,614 sentences (in the "foreign" languages) and a slightly smaller number of alignments (segments paired with their Czech equivalents): 172,376. In all of these texts, the users of *InterText* made the total of 13,960 changes to alignment, 5635 changes to sentence segmentation, and 5719 editing changes within sentences. The numbers are not uniformly distributed, depending on the difficulty of the texts as perceived by the segmentation and alignment tool. Moreover, the numbers include repeated clicks in the case of more complex corrections, and even additional clicks concerning the same sentence or segment in the case of repeated corrections. Nevertheless we can conclude that the percentage of automatically misaligned segments in our sample was at most 8.1% and the percentage of wrongly assigned sentence boundaries at most 2.9%, while some cases of wrongly identified sentence boundaries actually lead to additional corrections in segmentation. Considering all the above factors, the figures do not contradict results reported previously.

The number of editing changes is more difficult to interpret, because some editing might have preceded the import to *InterText* and more changes could have been made within a single sentence. Yet we can estimate the percentage of sentences including typos and similar errors at max. 3.1%.

In comparison to alignment and segmentation the evaluation of linguistic annotation in a multilingual environment is prohibitively costly. Therefore, we rely on reported figures, which are available at least for some tools (Spoustová et al., 2007).

The quality of sentential segmentation and alignment is crucial for an additional step towards the identification and alignment of sub-sentential units: words, multi-word units or even constituents. Word-to-word alignment is on the top of our list, because it will be useful immediately in the search interface to suggest potential equivalents of keywords in the query.

Sub-sentential alignment is related to a more sophisticated linguistic mark-up. After resolving the issue of incompatible tokenization and tagsets, and extending morphosyntactic annotation to as many languages as possible, syntactic annotation will be an interesting addition.

## 7. Acknowledgements

The authors are grateful to František Čermák who initiated and guided the project, to many colleagues who contributed their expertise and efforts: Michal Křen, Michal Štourač, Pavel Vondřička, Hana Skoumalová, Pavel Procházka, Saša Marková and Petr Marek – to name just a few, to all project participants in cooperating departments and institutions, and last but not least to all who offered their data and tools, often providing helpful advice to make them run to our satisfaction. We also wish to thank all three anonymous reviewers for important comments.

Our work has been supported financially by the Czech Ministry of Education, grant no. 0021620823.

## 8. References

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92.

Christian Chiarcos, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. 2008. A flexible framework for integrating annotations from different tools and tag sets. *TAL*, pages 217–246.

Tomaž Erjavec. 2010. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

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

Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Alexandr Rosen. 2005. In search of the best method for sentence alignment in parallel texts. In Radovan Garabík, editor, *Computer Treatment of Slavic and East European Languages: Third International Serminar, Bratislava 10–12 November 2005*, pages 174–185, Bratislava. VEDA.

Alexandr Rosen. 2010. Mediating between incompatible tagsets. In Lars Ahrenberg, Joerg Tiedemann, and Martin Volk, editors, *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, volume 10 of *NEALT Proceedings Series*, pages 53–62, Tartu, Estonia. Northern European Association for Language Technology.

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

Anil Kumar Singh and Samar Husain. 2005. Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 99–106, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007*, pages 67–74, Praha, Czechia. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufiş. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147.

Jörg Tiedemann. 2009. News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596.

Martin Vavřín and Alexandr Rosen. 2008. InterCorp: A Multilingual Parallel Corpus Project. In *Proceedings of the International Conference Corpus Linguistics – 2008*, pages 97–104. St. Petersburg State University.

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

Pavel Vondřička. 2010. TCA2 – nástroj pro zpracovávání překladových korpusů [TCA2 – a tool for processing translation corpora]. In František Čermák and Jan Kocek, editors, *Mnohojazyčný korpus InterCorp: Možnosti studia [Multilingual Corpus InterCorp: Research Options]*, pages 225–231. Nakladatelství Lidové noviny.

Daniel Zeman. 2010. Hard Problems of Tagset Conversion. In Alex Fang, Nancy Ide, and Jonathan Webster, editors, *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pages 181–185, Hong Kong, China.

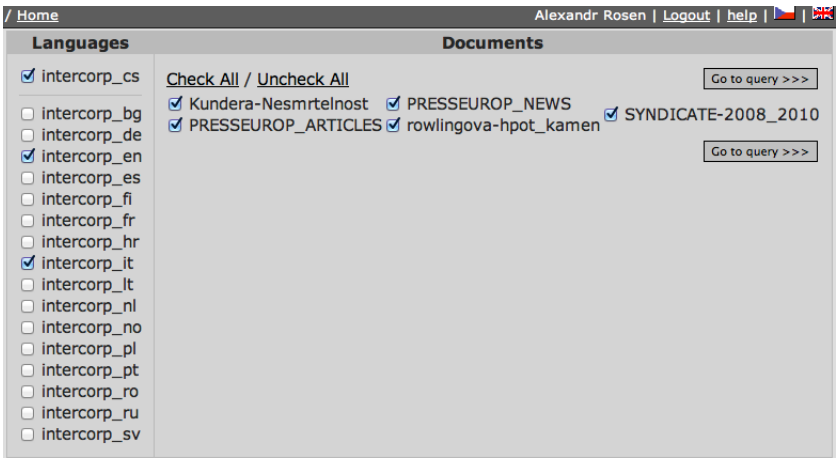


Figure 1: Selecting languages and texts

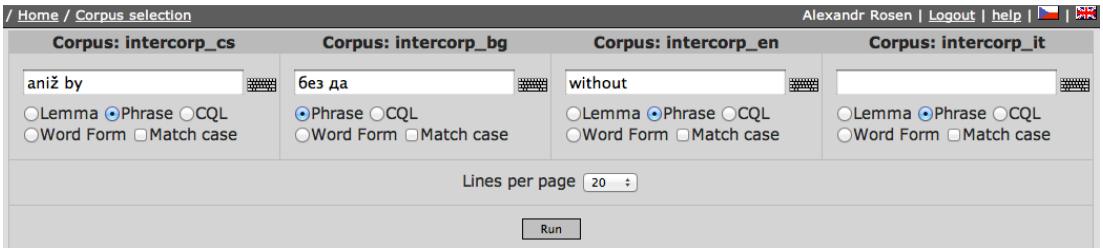


Figure 2: A query

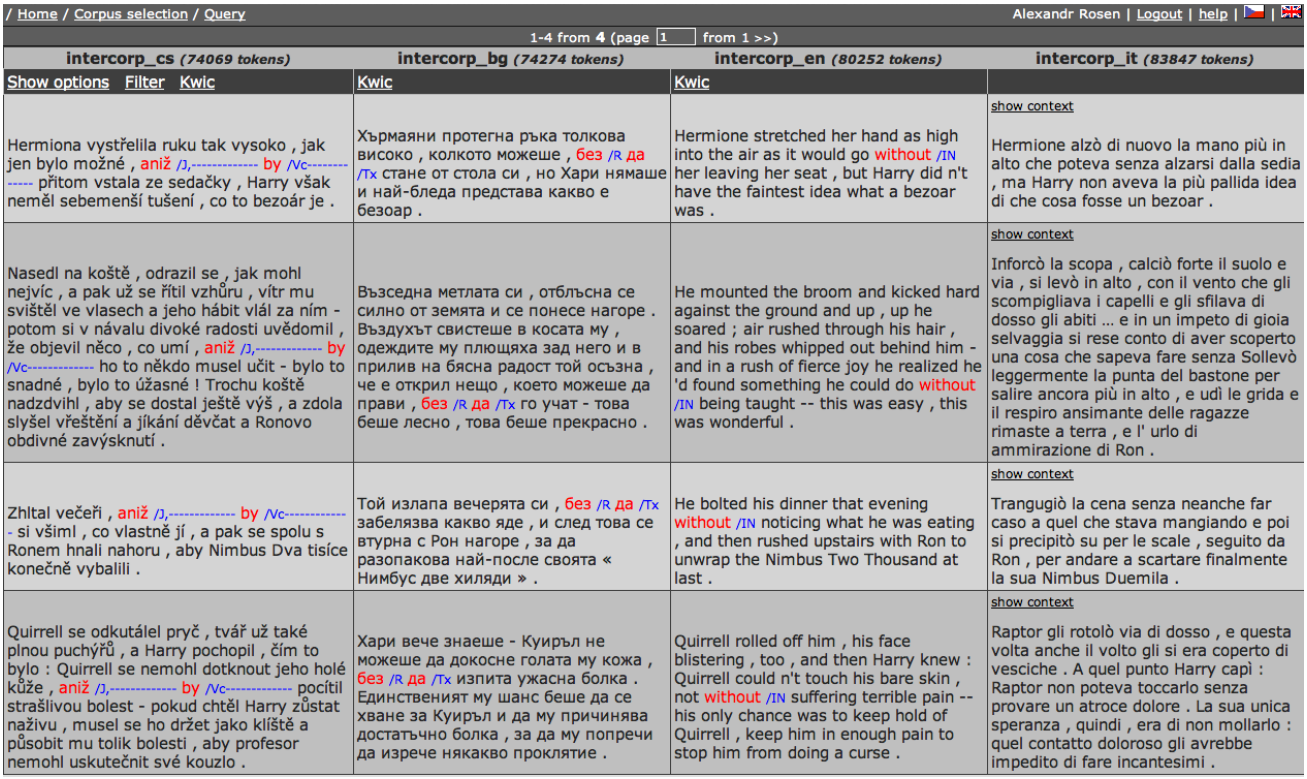


Figure 3: Result of the query