Aligning parallel texts with InterText

Pavel Vondřička

Institute of the Czech National Corpus, Charles University in Prague
Faculty of Arts, nám. J. Palacha 2, CZ-11638 Praha 1, Czech Republic
Pavel.Vondricka@ff.cuni.cz

Abstract

InterText is a flexible manager and editor for alignment of parallel texts aimed both at individual and collaborative creation of parallel corpora of any size or translational memories. It is available in two versions: as a multi-user server application with a web-based interface and as a native desktop application for personal use. Both versions are able to cooperate with each other. InterText can process plain text or custom XML documents, deploy existing automatic aligners and provide a comfortable interface for manual post-alignment correction of both the alignment and the text contents and segmentation of the documents. One language version may be aligned with several other versions (using stand-off alignment) and the application ensures consistency between them. The server version supports different user levels and privileges and it can also track changes made to the texts for easier supervision. It also allows for batch import, alignment and export and can be connected to other tools and scripts for better integration in a more complex project workflow.

Keywords: parallel corpora, alignment, editor

1. What is InterText?

InterText is a post-alignment editor for aligned parallel texts and a user interface to existing automatic aligners *Hunalign* (Varga et al., 2005) and *TCA2*[1] (Hofland, 1996; Hofland and Johansson, 1998; Vondřička, 2010).

InterText has been developed for the project *InterCorp* (Čermák and Rosen, 2012; Rosen and Vavřín, 2012; Křen et al., 2011), a project of translational parallel corpora including more than 30 languages, where nearly 200 external partners cooperate on collecting and aligning thousands of texts for the main, high-quality core of the parallel corpus requiring manual verification. But the tool has been developed with flexibility in mind and thus it can be used for several other purposes, including small personal projects or creation of translational databases. It has already been used in several other research projects world-wide, providing useful feedback.

The name *InterText* today refers to two slightly different and fully independent software applications: a server based text management system with web-based editor interface, called *InterText server*, and a newly developed stand-alone personal desktop application called *InterText editor*. Both applications are also able to cooperate with each other, if required. They are both publicly available as open-source software and include a detailed documentation.

2. The need for a new tool

The extraordinary extend of the *InterCorp* project – both in the number of texts, languages and participators – puts extraordinary demands on its management and therefore on software tools as well. Scalability and robustness are among the most self-evident, but unfortunately not the most common qualities of existing tools. From the beginning, the workflow in the project has suffered from technical issues arising from deficiencies of the technically obsolete aligner tool *ParaConc* (Barlow, 1992), especially when combined with limited technical skills of the collaborators, who are mostly linguists or translation scholars. Although *ParaConc* offers additional value as a concordancer, it cannot treat Unicode and XML files correctly. In addition, it cannot generate stand-off alignment[2] and is limited to the MS Windows platform only. The need for the use of stand-off alignment became quickly evident, as well as the need for separation of the editors from the technical aspects of dealing with XML files directly. At this stage, the texts were exposed to unexpected corruptions leading to complications with batch-processing and automatization of the subsequent processes (tagging and conversion into the internal format of the corpus search engine).

No other equivalent tool has been found, except for the *TCA2* aligner. But *TCA2* did not receive much attention from the *InterCorp* team, because it is an interactive aligner with relatively slow processing of texts with long sentences. The idea of manual post-alignment corrections (as known from *ParaConc*) seemed more user-friendly and effective. In addition, *TCA2* did not address another important deficiency of *ParaConc*: no possibility to fix typos or other defects in the texts, being still frequently discovered in this late phase of text processing.

Development of *InterText* helped to solve all the mentioned problems. Now, the external collaborators only deliver proof-read texts to the project's technical administrator, who proceeds with automatic segmentation of the text into sentences and imports it into *InterText*. *InterText* runs an automatic aligner and makes the aligned pair of texts available to the external editor through the user interface for manual verification of the alignment and for further corrections (see

[1] A modified command-line version of *TCA2* is required.

[2] Stand-off alignment is stored separately from the original text files, in a third file containing only the list of identifiers of sentences from the two texts paired together into semantically corresponding groups (segments of alignment). This is especially useful if one text version is aligned to several other versions, since different alignments may often need different granularity.

figure 1[3]).

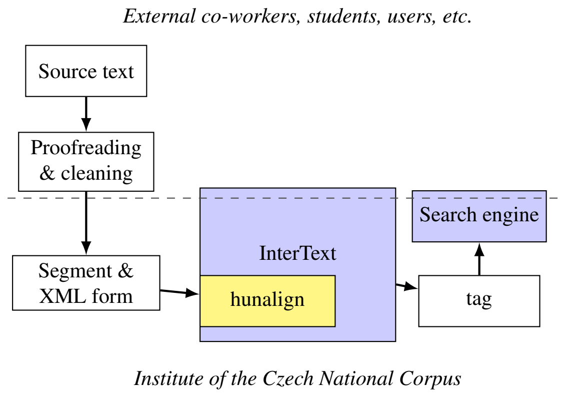


Figure 1: The task of *InterText server* in the current workflow in the *InterCorp* project.

The system ensures text and alignment consistency and keeps a log of all changes made to the text, so that they can be checked and possibly reverted in case of editor's misunderstanding of the process or project standards. *InterText* is also able to keep consistency across several independent alignments of one and the same text, when the editor changes segmentation of the text (in case a wrong segmentation into sentences is discovered).

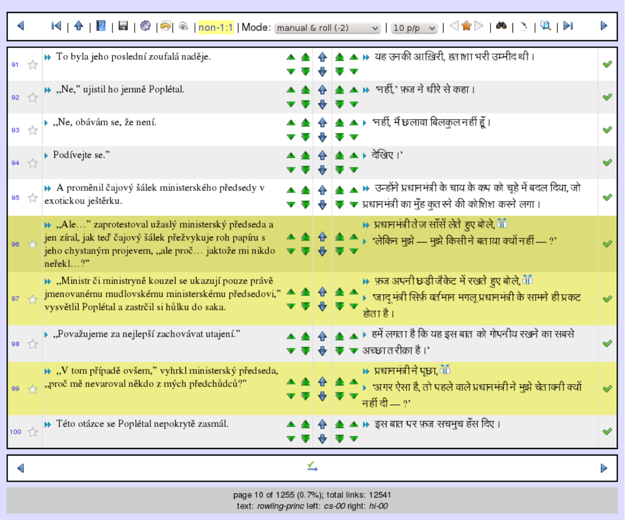
The manual alignment in project *InterCorp* is limited to the level of sentences, but *InterText* is able to align texts on any level. However, the alignment is always limited to one level only and currently – for practical reasons[4] – to linear alignment.

3. InterText server

InterText has been first developed as a web-based tool running in PHP on top of a MySQL database on a central server. Therefore it is called *InterText (web) server*, now. It was connected to a previously existing bibliographical database of texts and database of editors engaged in the project *InterCorp*. Because of the specificity of the bibliographical database, it is not part of the public *InterText* distribution.

InterText is able to make use of three levels of users: project administrators, sub-project (e.g. language) coordinators (supervisors), and editors, and it can limit their access privileges in an appropriate way. But the behaviour of *InterText* can be configured beyond the specifics of the project *InterCorp* in quite many aspects. For example, *InterText* is able to create and edit arbitrary alignments between different language versions of one text, while project *InterCorp* is currently aligning all language versions with the Czech language version, used as the "pivot language". *InterText* also provides a special permission for disabling changes to the segmentation in pivot texts generally, while pivot texts can be defined by virtue of a regular expression matching the version name. Basic editing permissions (editing the texts generally and changing the segmentation of the pivot version specifically) can be individually set for particular alignments as needed.

InterText server can import virtually any custom XML document in any encoding (and thus any language[5]), run an automatic aligner[6] on the textual contents of two arbitrarily selected language versions of one text, and make the alignable text contents[7] available to the selected editor and supervisor for corrections and editing. The automatic aligner can later be (re)applied again to just a part of the alignment, if necessary.[8]



91	To byla jeho poslední zoufalá naděje.	यह उनकी आख़िरी, हताशा भरी उम्मीद थी।
92	„Ne," ujistil ho jemně Poplétal.	'नहीं,' फ़ज ने धीरे से कहा।
93	„Ne, obávám se, že není.	'नहीं, मैं छलावा बिलकुल नहीं हूँ।
94	Podívejte se."	देखिए।'
95	A proměnil čajový šálek ministerského předsedy v exotickou ještěrku.	उन्होंने प्रधानमंत्री के चाय के कप को चूहे में बदल दिया, जो प्रधानमंत्री का मुँह कुतरने की कोशिश करने लगा।
96	„Ale…" zaprotestoval užaslý ministerský předseda a jen zíral, jak teď čajový šálek přežvykuje roh papíru s jeho chystaným projevem, „ale proč… jaktože mi nikdo neřekl…?"	प्रधानमंत्री तेज़ साँसें लेते हुए बोले, 'लेकिन मुझे — मुझे किसी ने बताया क्यों नहीं — ?'
97	„Ministr či ministryně kouzel se ukazují pouze právě jmenovanému mudlovskému ministerskému předsedovi," vysvětlil Poplétal a zastrčil si hůlku do saka.	फ़ज अपनी छड़ी जैकेट में रखते हुए बोले, 'जादू मंत्री सिर्फ़ वर्तमान मगलू प्रधानमंत्री के सामने ही प्रकट होता है।
98	„Považujeme za nejlepší zachovávat utajení."	हमें लगता है कि यह इस बात को गोपनीय रखने का सबसे अच्छा तरीक़ा है।'
99	„V tom případě ovšem," vyhrkl ministerský předseda, „proč mě nevaroval někdo z mých předchůdců?"	प्रधानमंत्री ने पूछा, 'अगर ऐसा है, तो पहले वाले प्रधानमंत्री ने मुझे चेतावनी क्यों नहीं दी — ?'
100	Této otázce se Poplétal nepokrytě zasmál.	इस बात पर फ़ज सचमुच हँस दिए।

page 10 of 1255 (0.7%); total links: 12541
text: *rowling-princ* left: *cs-00* right: *hi-00*

Figure 2: *InterText server* – editor interface showing parallel texts in Czech and Hindi (all control buttons shown).

The interface has a form of a table with two main columns corresponding to the two language versions (see figure 2), and each row corresponding to one segment of aligned sentences. Each table cell then contains any number of sentences belonging to the particular segment. The editor can freely move sentences between table rows (and thus segments), insert new rows or merge existing ones.

In order to help the user to focus on the most problematic places, the interface can highlight rows (segments) containing another amount of sentences than in the most common

[3] The scheme does not show a separate bibliographical database, where all metadata about the texts are stored and retrieved again for the resulting corpus.

[4] Cross-alignment (or out-of-order alignment) of elements does not pose a technical problem for the architecture of *InterText*, but it would put a strain both on the usability of the user interface and the usability of the resulting alignment, which would not be acceptable for most other currently used corpus tools

[5] Web browsers still may have difficulties displaying punctuation correctly in right-to-left scripts.

[6] Currently *hunalign* or *TCA2*. Both aligners may be installed with several configurations (depending on languages aligned) which may be chosen ad-hoc by the user of *InterText* in the form of "profiles".

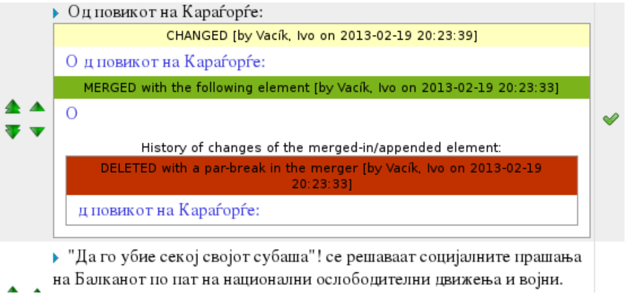
[7] The user can define which XML elements contain alignable text units (for each XML file separately).

[8] For example, when one of the versions is missing a large part of the text, the automatic aligner is confused and the rest of the resulting alignment is very difficult to fix manually. In such case, the editor can manually mark the gap and let the automatic aligner to re-align the rest of the texts much more reliably.

ratio 1:1 – those are the segments containing most errors from the automatic alignment. Segments can also be visually (book)marked for later revision (e.g. by the supervisor).

The system can automatically follow the verification process as the editor is progressively checking and fixing the alignment: when any change is done to the alignment, *InterText* can automatically change the status of all preceding segments to "manually verified" (or "confirmed"), and so the editor can anytime stop his work and later easily find the place where his job was interrupted.

The editor may (if permitted by his supervisor or administrator) open the contents of the single text elements and edit them. It is also possible to split or merge the sentences. All changes to the text contents are tracked and for each sentence a change-log is available for later revision, showing who and when did a change to the contents (see figure 3). This is especially important for larger projects where several editors may change a text (language) version taking part in several alignments (commonly the "pivot version").



Од повикот на Караѓорѓе:

CHANGED [by Vacík, Ivo on 2013-02-19 20:23:39]

О д повикот на Караѓорѓе:

MERGED with the following element [by Vacík, Ivo on 2013-02-19 20:23:33]

О

History of changes of the merged-in/appended element:

DELETED with a par-break in the merger [by Vacík, Ivo on 2013-02-19 20:23:33]

д повикот на Караѓорѓе:

"Да го убие секој својот субаша"! се решаваат социјалните прашања на Балканот по пат на национални ослободителни движења и војни.

Figure 3: *InterText server* – editor interface detail showing the history of changes of one element in a Macedonian text.

InterText also cares for parent elements ("containers") of the alignable text elements: it means that when aligning sentences in a standard text, the user can see also where paragraphs start, and the paragraphs can also be split or merged if necessary.

A basic search functionality is available as well. It can be used to search both for text, sentences by their identifier, changed (edited) sentences and for special types of segments in the alignment.[9] Regular expressions are also supported in the text search.

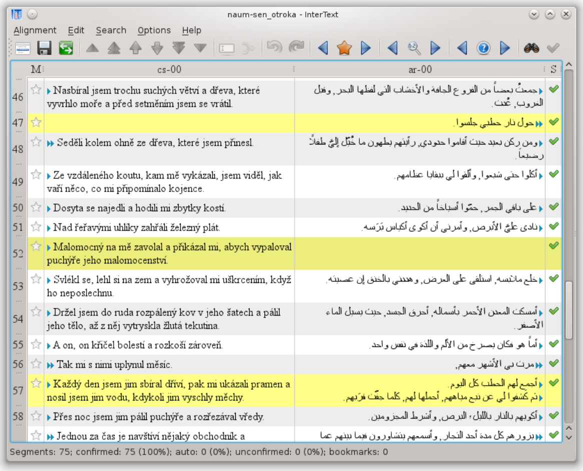
For the purpose of text and alignment management, *InterText* also keeps track of the status of each alignment: it can be "open" for changes or reserved by a "remote editor" (see below), it can be marked as "finished" when ready for the inclusion into the final corpus, or it can be "closed" by the administrator (or "blocked" for some other reason). External scripts (or "triggers") can be automatically run when the status of some alignment is changed by a user and they may also check and decide whether the text or alignment fulfils some particular formal criteria and whether its status really can be changed. The triggers may also initiate financial clearing of the finished job and eventually start further processing of the texts – export, tagging, conversion to the final format or even automatic inclusion into the corpus repository.

An additional set of command-line scripts is available for external control (scripting) and especially for easy batch import and export of texts and their alignments. While the PHP scripts cannot be used as any kind of library in third-party applications, they can be easily run from command-line or from other scripts or applications in order to automatically trigger actions such as import or export of texts or creating initial automatic alignments between imported language versions.

InterText editor is provided with a thorough user-guide (in English and Czech) in the form of an integrated web page.

4. InterText editor

The personal desktop application *InterText editor* has been lately developed both as an external editor for *InterText server*, overcoming the limitations of an on-line, web-based editor, and as an independent, standalone alignment editor for purely personal projects. It implements most of the functionality of *InterText server*, except for the user-level and permission system, the command line scripts and the log of changes. On the other hand, it offers more comfortable and faster operation using both mouse and keyboard, undo and redo functionality and full search and replace capability. It also allows for a more detailed customization of the interface. It uses the Qt framework and it is thus available for most common platforms:[10] *MacOS X, Linux* (and other *POSIX* systems) and *MS Windows*.



	M	cs-00	ar-00	S
46	☆	Nasbíral jsem trochu suchých větví a dřeva, které vyvrhlo moře a před setměním jsem se vrátil.	جمعتُ بعضاً من الفروع الجافة والأخشاب التي لفظها البحر, وقبل الغروب, عُدت.	✓
47	☆		حول نار حطبي جلسوا.	✓
48	☆	Seděli kolem ohně ze dřeva, které jsem přinesl.	ومن ركن بعيد حيث أقاموا حدودي, رأيتهم يطهون ما خُيِّل إليَّ طفلاً رضيعاً.	✓
49	☆	Ze vzdáleného koutu, kam mě vykázali, jsem viděl, jak vaří něco, co mi připomínalo kojence.	أكلوا حتى شبعوا, وألقوا لي ببقايا عظامهم.	✓
50	☆	Dosyta se najedli a hodili mi zbytky kostí.	على باقي الجمر, حمّوا أسياخاً من الحديد.	✓
51	☆	Nad řeřavými uhlíky zahřáli železný plát.	نادى عليَّ الأبرص, وأمرني أن أكوي أكياس بَرَصه.	✓
52	☆	Malomocný na mě zavolal a přikázal mi, abych vypaloval puchýře jeho malomocenství.		✓
53	☆	Svlékl se, lehl si na zem a vyhrožoval mi uškrcením, když ho neposlechnu.	خلع ملابسه, استلقى على الأرض, وهددني بالخنق إن عصيته.	✓
54	☆	Držel jsem do ruda rozpálený kov v jeho šatech a pálil jeho tělo, až z něj vytryskla žlutá tekutina.	أمسكت المعدن الأحمر بأسماله, أحرق الجسد, حيث يسيل الماء الأصفر.	✓
55	☆	A on, on křičel bolestí a rozkoší zároveň.	أمّا هو فكان يصرخ من الألم واللذة في نفس واحد.	✓
56	☆	Tak mi s nimi uplynul měsíc.	مرت بي الأشهر معهم,	✓
57	☆	Každý den jsem jim sbíral dříví, pak mi ukázali pramen a nosil jsem jim vodu, kdykoli jim vyschly měchy.	أجمع لهم الحطب كل اليوم. ثم كشفوا لي عن نبع مياههم, أحملها لهم, كلما جفّت قِرَبهم.	✓
58	☆	Přes noc jsem jim pálil puchýře a rozřezával vředy.	أكويهم بالنار بالليل؛ البرص, وأشرط المجزومين.	✓
	☆	Jednou za čas je navštíví nějaký obchodník a	يزورهم كل مدة أحد التجار, وأسمعهم يتشاورون فيما بينهم عما	✓

Segments: 75; confirmed: 75 (100%); auto: 0 (0%); unconfirmed: 0 (0%); bookmarks: 0

Figure 4: *InterText editor* – editor interface showing parallel texts in Czech and Arabic.

InterText editor is able to download texts and alignments from an *InterText server* (see figure 5).[11] While the alignment itself is locked for changes on the server[12] until explicitly released from the (remote) application again,

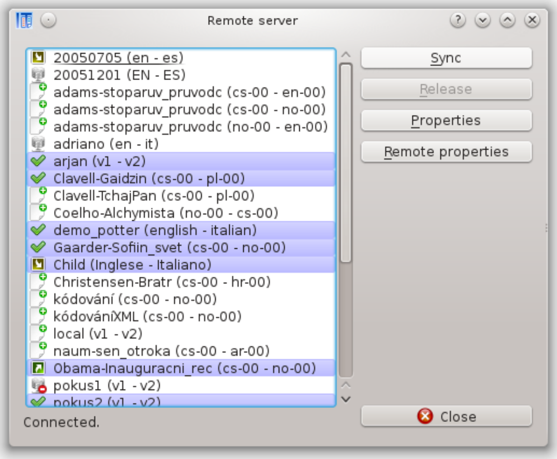
[9] Segments with missing text in one or the other version (n:0 or 0:n) or non-trivial segments (other than 1:1).

[10] The latest versions of Qt open the potential to possibly support mobile platforms (*Android* and *iOS*) as well.

[11] Upload of new alignments to the server is actually also supported, if permitted by the administrator of the *InterText server*.

[12] By changing its status to "remote editor".

changes to the text contents may still be done on the server (e.g. when the same text is part of some other alignment) – such changes are tracked and may be synchronized with the local copy in *InterText editor*. In that way, the remote user can stay informed about possible changes in the text resulting from other people's work, and both the text and the alignment can be worked on when off-line. The result (or its current state) may be submitted back to the server anytime when the remote user gets on-line, until the alignment is finally released from the remote application.



20050705 (en - es)
20051201 (EN - ES)
adams-stoparuv_pruvodc (cs-00 - en-00)
adams-stoparuv_pruvodc (cs-00 - no-00)
adams-stoparuv_pruvodc (no-00 - en-00)
adriano (en - it)
arjan (v1 - v2)
Clavell-Gaidzin (cs-00 - pl-00)
Clavell-TchajPan (cs-00 - pl-00)
Coelho-Alchymista (no-00 - cs-00)
demo_potter (english - italian)
Gaarder-Sofiin_svet (cs-00 - no-00)
Child (Inglese - Italiano)
Christensen-Bratr (cs-00 - hr-00)
kódování (cs-00 - no-00)
kódováníXML (cs-00 - no-00)
local (v1 - v2)
naum-sen_otroka (cs-00 - ar-00)
Obama-Inauguracni_rec (cs-00 - no-00)
pokus1 (v1 - v2)
pokus2 (v1 - v2)

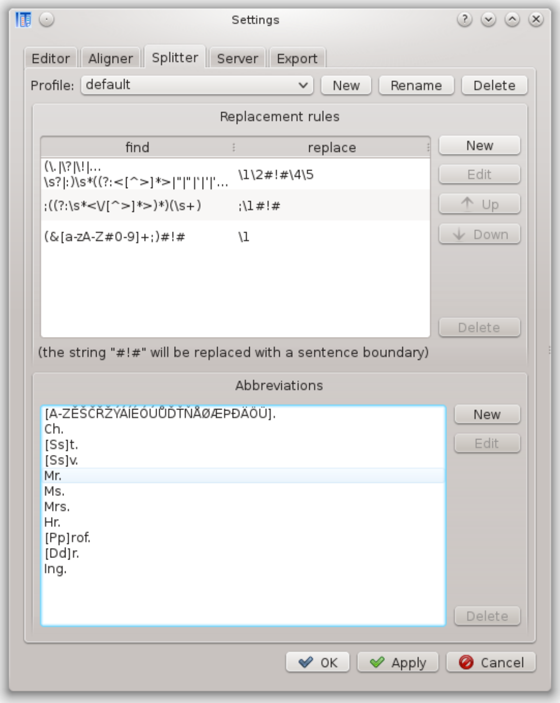
Connected.

Figure 5: *InterText editor* – dialog showing list of alignments available on a remote *InterText server* and their different state of synchronziation with the local repository.

As already mentioned, *InterText editor* is also targeted at individual use by linguists with limited technical skills and without any need to install *InterText server* and to run a large centralized project with multiple users. It can thus import plain text files or unsegmented XML documents and apply a basic rule-based (fully configurable, see figure 6) sentence splitter on them. It can also import newline-aligned plain text files. Export of texts is fully customizable by the user and predefined configurations include export of new-line aligned texts, ParaConc semi-XML files and basic TMX[13] files.

The application can also compute direct alignment between two language versions of one text that are already indirectly aligned via a third common language version.

InterText editor is provided with a comprehensive user-guide (in English) in PDF, explaining also the most important basic principles of text alignment and common caveats for inexperienced editors.



Profile: default

```
find                                        replace
(\.|\?|\!|...
\s?|:)\s*((?:<[^>]*>|"|"|'|'|'...          \1\2#!#\4\5
;((?:\s*<V[^>]*>)*)(\s+)                    ;\1#!#
(&[a-zA-Z#0-9]+;)#!#                        \1
```

(the string "#!#" will be replaced with a sentence boundary)

[A-ZĚŠČŘŽÝÁÍÉÓÚŮĎŤŇÅØÆÞÐÄÖÜ].
Ch.
[Ss]t.
[Ss]v.
Mr.
Ms.
Mrs.
Hr.
[Pp]rof.
[Dd]r.
Ing.

Figure 6: Sentence splitter can have several profiles configured as sets of replacement rules based on regular expressions and exceptions ("abbreviations").

5. Aims of further development

Both versions of *InterText* are being continuously improved and extended within the *InterCorp* project. The goal is to further improve the usability, flexibility, scalability and robustness (especially of the personal application *InterText editor*), in order to make it more and more user-friendly and useful both within the *InterCorp* project and for as many other similar purposes as possible, beyond the demands of the project itself. The tool is meant to generally help linguists and translation scholars (or translators) with various tasks related to alignment of texts, which are otherwise difficult without deeper technical skills. However, no application can offer both full flexibility and simplicity at the same time. Therefore, the user may often be in need to adapt various aspects of the software by means of configuring it for his own particular purpose. Improving configurability and a detailed documentation is thus also a priority.

Of course, *InterText* cannot solve all problems of complex document processing: it cannot replace fully-fledged text editors nor complex XML editors. It cannot solve e.g. the confusion based on the existence of different historical text encodings and formats used in computers for different text documents in different languages. It does not try to reinvent the automatic aligners or other existing sophisticated tools, but it wants to provide a friendlier user interface to these tools for technically unskilled users. It also wants to integrate the different tools related to the creation of parallel texts and corpora.

At the moment, *InterText editor* can directly process plain text files or simple XML-like fragments (e.g. text with HTML markup). Obviously, some simple integrated conversion tool for common document formats (OpenDocu-

[13] TMX stands for *Translation Memory eXchange*, a standardized XML-based format commonly used by translational memories and other CAT (Computer-Aided/Assisted Translation) applications or machine translation tools.

ment,[14] Office Open XML,[15] etc.) might be desirable for personal use as well.

The output from *InterText* can be (in many cases) directly used by the *ParaConc* concordancer tool or by CAT or machine translation tools supporting TMX files (see above). A more straight-forward connection to some other concordance tool or even a corpus search engine might be desirable, but at the moment there are no commonly used systems generic enough for this task, which could be easily deployed by unskilled desktop users. In addition, further processing by taggers and lemmatizers would also be desirable in most cases, which is a highly language and task dependent job and cannot easily be generalized. The question of interoperability with other downstream tools must therefore stay open until such tools – with a comparable level of user-friendliness – are available.

An important feature of both *InterText* applications is their scalability. At the moment, there are already several thousands of texts and alignments currently available in the *InterText server* database of the project *InterCorp*. The texts include also huge novels of several hundreds of pages (in print), having tens of thousands of aligned sentences. Yet the applications must stay reasonably fast and responsible in all cases. Currently they can keep up, but small optimizations are still being worked on.

Because collaborative work and desire for high-quality output are main features of the project *InterCorp*, further support in this field may be desirable, such as a possibility to add notes and comments to the alignment or to the individual texts. While it would be extremely difficult to make *InterText* work as a universal XML editor to access all arbitrary metadata stored in the documents (beside of the text contents) for different projects, limited access to some particular attributes – such as editor's notes or comments – might be sufficient at this stage of work.

The need to cooperate with several dozens of paid external editors – often students – implies also the need for monitoring their activities and time spent on the process of alignment verification of different texts. For this purpose, *InterText server* is already able to log all activities in the system, but at the moment, there is no analytical tool or even visualisation of the data available in *InterText*. The project management is still considering the possibilities to evaluate this data and project the results into more rightful financial rewards for work on texts with different level of difficulty and time requirements.

InterText is also being explored and tested by several individuals as well as university projects around the world for different purposes and we are trying to provide support for their specific needs as much as possible.

6. Availability and license

Both *InterText server* and *InterText editor* are available under the GNU General Public License v3 from the author's web page `http://wanthalf.saga.cz/intertext/` or via the website of the ICNC at `http://ucnk.ff.cuni.cz`.

7. Acknowledgements

The author is grateful to Knut Hofland and Øystein Reigem[16] for introduction into the principles of the TCA(2) aligner, the authors of the automatic aligner *Hunalign* and to The Research Council of Norway for supporting the research stay leading (among others) to the implementation of *InterText* within their mobility programme Yggdrasil.

This paper and software resulted from the implementation of the Czech National Corpus project (LM2011023) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

8. References

Michael Barlow. 1992. Using Concordance Software in Language Teaching and Research. In W. Shinjo et al., editors, *Proceedings of the Second International Conference on Foreign Language Education and Technology*, Kasugai, Japan. LLAJ & IALL.

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

Knut Hofland and Stig Johansson. 1998. The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In Stig Johansson and Signe Oksefjell, editors, *Corpora and Crosslinguistic Research: Theory, Method and Case Studies*. Rodopi, Amsterdam.

Knut Hofland. 1996. A Program for Aligning English and Norwegian Sentences. In G. Perissinotto, editor, *Research in Humanities Computing 5. Selected papers from the ACH/ALLC Conference*, pages 165–178, Oxford. Clarendon Press.

Michal Křen, Alexandr Rosen, Michal Štourač, Martin Vavřín, and Pavel Vondřička. 2011. Paralelní korpus InterCorp po sedmi letech [The parallel corpus InterCorp after seven years]. In František Čermák, editor, *Korpusová lingvistika Praha 2011: 2 - Výzkum a výstavba korpusů*, volume 15 of *Studie z korpusové lingvistiky*, pages 105–115, Praha. Ústav Českého národního korpusu.

Alexandr Rosen and Martin Vavřín. 2012. Building a multilingual parallel corpus for human users. In Nicoletta Calzolari et al., editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2447–2452, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

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.

[14] OASIS open standard format supported natively by OpenOffice.org, Libre Office, Google Docs, IBM Lotus Symphony, etc.

[15] Open standard used by Microsoft Office.

[16] Both of UNI Digital (previously UNIFOB Aksis / HIT-senteret) in Bergen, Norway.