M. Vavřín, A. Rosen

INTERCORP: A MULTILINGUAL PARALLEL CORPUS[1]

1. Introduction

The goal of the project *InterCorp*, under way since 2005 with the Institute of the Czech National Corpus (ICNC) as the project leader, is to build a parallel synchronous corpus for Czech and most languages studied at the Faculty of Philosophy and Arts of Charles University in Prague. The current list includes 22 languages (see Table 1) plus Czech, which serves as the *pivot*: all 'foreign' texts have their Czech counterparts, while a foreign text may have no counterpart in another foreign language[2]. Yet the corpus is not a mere collection of subcorpora including Czech and another language: two or more foreign texts can be queried and the results displayed without the pivot, as long as the texts in the relevant languages are present.

Unlike corpora involving more than a few languages[3], but like some other corpora oriented primarily towards linguists, students and translators as direct users[4], *InterCorp* consists mainly of fiction, a

1 The work reported here is supported by the Czech Ministry of Education, grant no. 0021620823 (The Czech National Corpus and Corpora of Other Languages).

2 Currently, most of our texts are only in Czech and one other language. However, some texts have already reached up to 15 versions (Milan Kundera's novel The Unbearable Lightness of Living).

3 Two most obvious examples could be *Opus – an open source parallel corpus*: **http://urd.let.rug.nl/tiedeman/OPUS/**, and *The JRC-Acquis Multilingual Parallel Corpus*: **http://langtech.jrc.it/JRC-Acquis.html**

4 The Regensburg Parallel Corpus, see *von Waldenfels, R*. Compiling a Parallel Corpus of Slavic Languages. // Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) № 9. 2006. München, P. 123-138. **http://www.uni-regensburg.de/Fakultaeten/phil_Fak_IV/Slavistik/RPC/**

genre best approximating the needs of the project participants (twelve departments, two of them outside Charles University[5], each responsible for at least one language pair). The challenging task has been to integrate fruits of their expertise and effort into a common shared resource. Their substantial involvement made a distributed mode of pre-processing inevitable. The role of ICNC is to provide technical infrastructure, develop methodologies and facilitate coordination of all participants, while making sure that the integrity of texts is not compromised and their alignment is of acceptable quality. Coordinators for specific languages, mostly members of the participating departments, are responsible for the choice of texts, their pre-processing, and semi-automatic alignment, usually employing students to do scanning and proofreading[6].

2. Pre-processing

After selecting a specific text, the coordinator in charge of the language asks a student to provide its electronic version. The student is instructed to search first for an already available file, in the archives of ICNC or elsewhere. Whenever possible, the publishing house is contacted with a request to provide the electronic version of the text. If a file is not available, the text is scanned and OCR'd[7]. Proofreading is a requirement for all our scanned texts, as none of the OCR tools we tested provides results of satisfactory quality.

Proofread texts (as .doc or .rtf files) are exported from MS Word using a Visual Basic macro into a quasi-XML format accepted by

5 Masaryk University, Brno: Department of German Language and Literature, Faculty of Education; Palacký University, Olomouc: Department of Dutch Studies, the Faculty of Arts

6 See **http://www.korpus.cz/intercorp/?lang=en** for more details.

7 See **http://finereader.abbyy.com/** for details about the OCR tool used.

ParaConc[8], a PC-based tool for building and using parallel corpora, used by the project participants mainly for alignment and checking the results. This is performed in two steps.

In the first step, paragraph boundaries and some formatting commands present in the text are translated into XML tags. Special mark-up characters (&, <, >) are rendered as character entities. Due to the limitations of *ParaConc*, the texts are converted into a language-specific Windows encoding, while more character entities are used to represent characters outside the character set.

In the second step, sentence boundaries are identified. For Czech, we use a rule-based splitter[9], for other languages a tool based on an unsupervised learning algorithm[10]. See fig. 1 for a sample tagged text.

Fig. 1. A text sample with structural markup

<p id="697"><s id="697.1">Ten druhý pár se usadil v zaprášeném malém caffè plném mušinců, kde jsme konečně našli volná místa.</s>

<s id="697.2">Byli starší, avšak jejich doprovodná skupina byla stejně veselá a milující.</s>

<s id="697.3">Ženich měl vpravdě oslnivý bílý motýlek a frak a snědá hezká nevěsta, která blýskala bělmy i oslnivými zuby a které šíji zdobily módně nakrátko sestřižené vlasy, si oblékla těsné krátké šaty bez ramínek.</s>

<s id="697.4">Byly z tmavě rudého saténu a doplňovaly je dlouhé rukavice, střevíčky s vysokými podpatky a malý klobouček sedící jí v týle.</s>

---

8 *Barlow, M.* ParaConc: Concordance software for multilingual parallel corpora. // Language Resources for Translation Work and Research, LREC 2002. P. 20–24. See **http://www.athel.com/para.html.**

9 Progam *tokenize* by Pavel Květoň.

10 The Punkt sentence tokenizer, in an implementation from **http://nltk.org/**. See *Kiss, T., Strunk, J.* Unsupervised Multilingual Sentence Boundary Detection // Computational Linguistics. 2006. № 32. P. 485–525. This method has now superceded the formerly used less satisfactory internal algorithm of *ParaConc*.

<s id="697.5">Všechny doplňky barevně ladily se šaty.</s>
<s id="697.6">A ta nevěsta byla tak v sedmém měsíci těhotenství.</s>

Next, the text pairs are aligned, again in *ParaConc*. Paragraphs are aligned manually (the tool alerts the user to a mismatch), before automatic alignment is performed by a built-in implementation of the Gale-Church algorithm[11]. The alignment results are checked by students, coordinators and the main coordinator to ensure the best possible alignment quality of all texts. The above sample, aligned with a corresponding foreign texts, is exported from *ParaConc* with additional tags for aligned segments, see fig. 2.

Fig. 2. A text sample with alignment markup

<s id="697.1"><seg id="2683">Ten druhý pár se usadil v zaprášeném malém caffè plném mušinců, kde jsme konečně našli volná místa.</seg></s>
<s id="697.2"><seg id="2684">Byli starší, avšak jejich doprovodná skupina byla stejně veselá a milující.</seg></s>
<s id="697.3"><seg id="2685">Ženich měl vpravdě oslnivý bílý motýlek a frak a snědá hezká nevěsta, která blýskala bělmy i oslnivými zuby a které šíji zdobily módně nakrátko sestřižené vlasy, si oblékla těsné krátké šaty bez ramínek.</s>
<s id="697.4">Byly z tmavě rudého saténu a doplňovaly je dlouhé rukavice, střevíčky s vysokými podpatky a malý klobouček sedící jí v týle.</seg></s>
<s id="697.5"><seg id="2686">Všechny doplňky barevně ladily se šaty.</s>
<s id="697.6">A ta nevěsta byla tak v sedmém měsíci těhotenství.</seg></s>

Obviously, this style of alignment tagging is specific to a language pair: the same text aligned with its counterpart in yet another language would most likely have different `<seg>` tags. Multiple versions of an aligned Czech text are the price for distributed pre-

11 *Gale, W. A., Church, K. W.* A Program for Aligning Sentences in Bilingual Corpora // Computational Linguistics. 1993. № 19. P. 75–102.

processing using software tools available. However, a stand-alone alignment annotation file, referring to <s> rather than <seg> tags, can always be extracted for a given pair of texts and the <seg> tags can be discarded.

In the next step, the aligned texts are cleaned (*ParaConc* may insert tags in somewhat erratic ways) and transformed into a regular XML format in the UTF-8 encoding, including bibliographical data extracted from a database of texts available within the project. This database is also used for tracking the passage of a text through the pre-processing stages.

Finally, the texts can be morphologically tagged and/or lemmatized. This option depends on the availability and performance of suitable language-specific tools. Czech is the first target, but in addition to Czech and some "high density" languages where such resources are easy to obtain, we intend to have this basic level of linguistic annotation for at least some other richly inflected languages, for the obvious benefit to the user despite the possibly challenging differences in language-specific tagsets.

3. Parallel web interface

At the time of writing, a pilot version of a server-based corpus search tool is available for searching a part of the corpus. The web-based interface is built on top of the corpus manager *Manatee*[12], already used in the monolingual part of the Czech National Corpus. The tool includes a modification of the *Manatee*'s GUI *Bonito*, currently modified and optimized by the ICNC staff to satisfy user's needs, about to be launched for routine search tasks shortly. (More details and samples in the camera-ready version.).

---

12 *Rychlý, P., Smrž, P.* Manatee, Bonito and Word Sketches for Czech. // Proceedings of the Second International Conference on Corpus Linguistics. 2004. Saint-Petersburg. P. 124–132. **http://nlp.fi.muni.cz/projects/bonito/**

4. The current state and the issue of balance

A balanced parallel corpus is much harder to built than a monolingual one: some texts are rarely translated. It may not be a serious problem for some purposes (such as collecting training data for stochastic machine translation), and an approach open to any available type of texts, usually resulting in a massive prevalence of one or a few types, is well justifiable. But priorities in our project were different, and a realistic approach nearest to the goal of a resource including *natural* language was the preference for literary texts. This is not the shortest path to obtain as much data as possible, but one that provides a much richer resource for most of our potential users. An additional bonus is the relatively easy alignment of fiction, as compared with some other genres, a mix of translation sources and targets, and – at least for some titles – a chance to acquire the same texts in multiple languages.

Table 1 below shows large differences in the number of word tokens and texts across language pairs. In some languages there is a long way to go before considerations of balance turn into a topic. But for some languages, where a critical mass of texts has been reached, additional genres are already in focus in order to obtain a more varied setup. Legal texts, technical manuals, or software documentation are all easily available. Non-fiction, poetry, and drama are not substantially more difficult than fiction. However, unlike with monolingual data, there will never be enough translations of newspaper articles, or even multi-lingual editions of complete periodicals. The conclusion that a parallel corpus can never be balanced is unavoidable, but a less ambitious target of some reasonable mix is realistic.

Table 1. List of subcorpora and their size.
(as of January 2008, updated figures in camera-ready version)

Language	No. of Czech word tokens (thousands)	No. of foreign word tokens (thousands)	No. of completed titles	No. of titles not yet completed
Bulgarian	867	868	16	3
Croatian	1 240	1 306	23	24
Danish	73	96	3	1
Dutch	795	919	20	13
English	1 802	2 099	26	44
Finnish	216	190	4	3
French	804	971	20	2
German	1 459	1 657	20	62
Hungarian	1 196	1 184	23	2
Italian	1 815	2 088	21	2
Latvian	307	292	8	2
Lithuanian	0	0	0	15
Macedonian	0	0	0	7
Norwegian	551	562	5	3
Polish	2 123	2 066	38	6
Portuguese	1 517	1 744	20	1
Russian	1 053	1 051	19	2
Serbian	1 049	1 122	12	4
Slovak	0	0	0	4
Slovene	393	430	5	9
Spanish	3 935	4 500	59	8

Swedish	1 729	1 993	31	10
TOTAL	**22 924**	**25 138**	**373**	**227**

4. Perspectives and conclusion

The substantial share of manual effort, especially during the alignment phase, where a tool intended for a slightly different environment of a single user is used, and the related technical issues are obvious targets for optimization. An alternative alignment method is previewed for some texts that should be processed in a way bypassing the standard path, such as legal documents freely available online in multiple languages[13]. With the possibility of using morphological annotation, the synergic effects of more sophisticated, content-based alignment methods, known to perform better with lemmatized texts, can lead to results justifying their integration into the routine pre-processing path.

Another crucial point is the legal status of texts in the corpus and potential restrictions in accessing them due to copyright concerns. We believe we should find a solution to open the web search for concordances to the public, even at the cost of shuffling the sequential order of sentences if necessary, while leaving the context in place for more restricted audience. This concern of availability extends to our potential partners abroad, as we wish to share experience and results with people and institutions of similar interests.

13 For a comparison of potential candidates with references to previous endeavours of a similar kind see *Rosen, A.* In Search of the Best Method for Sentence Alignment in Parallel Texts // Computer Treatment of Slavic and East European Languages: Third International Serminar, Bratislava. 2005. P. 174–185.