Michal Křen, Alexandr Rosen, Michal Štourač, Martin Vavřín, Pavel Vondřička

Paralelní korpus InterCorp po sedmi letech

Abstract
The paper presents the architecture and the current state of the parallel corpus InterCorp, including an outline of its recent development and a comparison with other parallel corpora. This is followed by an overview of the data collection procedure that covers text selection criteria, data format, conversion, alignment, lemmatization and tagging. Among the specific tools, we focus on the on-line alignment editor InterText and the parallel search engine interface Park. Finally, we discuss challenges and prospects of the project.

1. Úvod

Korpus InterCorp je hlavním výstupem stejnojmenného projektu, jehož cílem je vybudovat rozsáhlý paralelní synchronní korpus pokrývající velké množství jazyků (Vavřín a Rosen 2008). Práce na něm byly zahájeny v roce 2005 v rámci řešení výzkumného záměru MSM0021620823 *Český národní korpus a korpusy dalších jazyků* (2005-2011). Na jeho tvorbě se významnou měrou podílejí zejména pedagogové a studenti FF UK v Praze, ale i další spolupracovníci Ústavu Českého národního korpusu FF UK (ÚČNK), například Katedra německého jazyka a literatury PedF MU v Brně, Ústav románských jazyků a literatur FF MU, Slovanský ústav AV ČR aj.

ÚČNK je hlavním koordinátorem zodpovědným za organizaci, financování a standardizaci dat, provozuje centrální datové úložiště a poskytuje jednotlivým účastníkům technickou podporu, konzultace a školení. Každý jazyk má svého koordinátora, který zajišťuje akvizici cizojazyčných textů, jejich zarovnávání, a v neposlední řadě také nábor studentů, kteří vlastní práci provádějí. Koordinátoři jsou také zodpovědní za celkovou kvalitu zpracování textů ve svém jazyce.

InterCorp obsahuje převážně manuálně zarovnané beletristické texty a také výběr automaticky zarovnaných publicistických článků z webových stránek Project Syndicate (http://www.project-syndicate.org/). Současný rozsah zpřístupněných dat je 72 milionů zarovnaných slov ve 22 jazycích (viz Tabulka 1), na české straně je to celkem 41 milionů slov v 652 textech (v počtu textů není zahrnutý Project Syndicate). InterCorp byl zpočátku plánován jako korpus malý a spíše doplňkový, teprve v dalších letech došlo k velkému nárůstu datové základny, jejíž současný rozsah mnohonásobně převyšuje původní plán. Tento nárůst si vynutil nejen zdokonalování stávající databáze pro správu textů, ale také urychlený vývoj nových nástrojů pro zarovnávání (InterText) a pro vyhledávání v korpusech (Park).

jazyk	počet slov (v tisících)	počet textů
angličtina	5 695	Syndicate + 49
bulharština	1 135	15
dánština	190	5
finština	1 247	19
francouzština	3 141	Syndicate + 21
chorvatština	6 735	96
italština	2 817	28

litevština	353	17
lotyština	1 085	33
maďarština	1 123	17
němčina	8 846	Syndicate + 100
nizozemština	3 914	58
norština	2 158	21
polština	4 716	80
portugalština	1 312	18
rumunština	671	5
ruština	2 951	Syndicate + 25
slovenština	6 899	138
slovinština	992	16
srbština	1 724	27
španělština	10 905	Syndicate + 108
švédština	3 673	47
CELKEM	**72 280**	**943**

Tabulka 1: Rozsah zpřístupněné části korpusu InterCorp z února 2011; počty slov jsou uváděny včetně textů Project Syndicate.

2. Architektura korpusu InterCorp

InterCorp není pouhým souborem několika různých paralelních korpusů, jde o jeden ucelený mnohojazyčný paralelní korpus s jednotnou strukturou. Každý text zařazený do InterCorpu musí mít českou verzi a nejméně jednu verzi v dalším jazyce. Česká verze textu je vždy jen jedna, a právě s ní je zarovnaná každá cizojazyčná verze téhož textu; čeština tak plní roli tzv. pivota. Přestože tedy neexistují jiná než česko-cizojazyčná zarovnání, je díky této struktuře možné se z libovolné cizojazyčné verze daného textu dostat ve dvou krocích (přes češtinu) do jiné existující cizojazyčné verze. Výše uvedené platí bez ohledu na to, zda je česká verze originál nebo překlad.

Obrázek 1: Struktura korpusu InterCorp s centrálním postavením češtiny.

Celý korpus je v současné době uložený v XML souborech v kódování UTF-8. Pro každou česko-cizojazyčnou dvojici verzí textu jsou informace o zarovnání obsaženy ve zvláštním zarovnávacím souboru, ze kterého vedou odkazy do vlastního textu. Tento formát, tzv. oddělené zarovnání (stand-off alignment), má tu výhodu, že umožňuje měnit nebo přidávat další verze textů a údaje o zarovnání bez zásahu do české verze. Zarovnávací soubor určuje pro každou dvojici jazykových verzí posloupnost tzv. segmentů, tj. vět s navzájem si odpovídajícími překlady, které obě jazykové verze pokrývají.

Protože každá věta má v celém korpusu jednoznačný identifikátor, může se na něj zarovnávací soubor odkazovat. Následující příklady ukazují zarovnání české věty s identifikátorem "cs:Waltari-EgyptanSinuhet:0:1696:1", tj. první věty odstavce s pořadovým číslem 1696 české verze románu Egypťan Sinuhet od Mika Waltariho:

```
<link type='2-1' xtargets='es:Waltari-EgyptanSinuhet:0:1508:1 es:Waltari-EgyptanSinuhet:0:1508:2;cs:Waltari-EgyptanSinuhet:0:1696:1' status='man'/>

<link type='1-1' xtargets='fi:Waltari-EgyptanSinuhet:0:995:1;cs:Waltari-EgyptanSinuhet:0:1696:1' status='man'/>
```

První řádek je ze španělsko-českého zarovnávacího souboru, druhý z finsko-českého. Zatímco v prvním případě odpovídají této jediné české větě dvě španělské věty (zarovnání typu 2-1), ve druhém jedna věta finská (zarovnání typu 1-1) – toto zarovnání je v korpusu nejčastější. Běžné jsou ale samozřejmě i jiné typy zarovnání, např. 2-2, 2-3, nebo také 1-0. V případě typu 1-0 jde o větu, která ve druhém jazyce nemá ekvivalent.

3. Zpracování a formát dat

Výběr textů je v zásadě v pravomoci jednotlivých koordinátorů. Berou při něm v úvahu mnoho někdy i protichůdných kritérií: současnost textu a překladu (přesné vymezení je pro každý jazyk obecně různé), jeho kvalitu nebo nutnost volby překladů ze třetího jazyka v

případě malých jazyků. Další prioritou je budování mnohojazyčného společného jádra, tedy zpracování vybraných textů v co největším množství jazyků.

Vybraný text je zadán studentům, kteří mají za úkol ho získat v elektronické podobě. Pokud je to možné, snažíme se texty získat přímo od nakladatelů – to probíhá obvykle za asistence koordinátorů jednotlivých jazyků. Jako zdroj českých verzí textů slouží také Český národní korpus (ČNK). Pokud není z různých důvodů možné získat text v elektronické podobě, přichází na řadu skenování s optickým rozpoznáváním znaků (OCR). To provádějí studenti daného jazyka pomocí programu FineReader (http://www.abbyy.com/). Bohužel žádný z nám známých programů není schopen zajistit takovou kvalitu rozpoznání, aby bylo možné texty zahrnout do korpusu bez korektur. Nedílnou součástí skenování a OCR jsou tedy také manuální korektury.

Po korekturách se texty ve formátu *.rtf nebo *.doc konvertují ve dvou krocích do formátu XML. V prvním kroku se text převede pomocí makra ve Visual Basicu do kódování Unicode a některé znaky se nahradí za entity (&, <, >). Odstavce jsou v tomto kroku označeny značkou <p> a pokud je to možné, jsou zachovány i řezy písma. Ve druhém kroku je text tokenizován (rozdělen na slova) a segmentován (vyznačeny hranic vět). Na české straně k tomu slouží program *tokenize* Pavla Květoně, pro cizí jazyky používáme *Punkt* (Kiss a Strunk 2006) v implementaci z http://www.nltk.org/. Výsledky obou programů jsou ještě dolaďovány pomocí našich interních skriptů.

Takto zpracované texty jsou načteny do editoru paralelních textů InterText (viz část 4). V něm jsou texty automaticky zarovnány pomocí programu *hunalign* (Varga et al. 2005; http://mokk.bme.hu/en/resources/hunalign/). V prostředí InterTextu studenti kontrolují a opravují automaticky zarovnané texty, kromě zarovnání je možné opravovat i zbývající překlepy a chyby v segmentaci. Po nich text přebírají a opět kontrolují koordinátoři jednotlivých jazyků a nakonec ještě jednou hlavní koordinátor tak, aby se zajistila pokud možno co nejvyšší kvalita korektur, segmentace a zarovnání výsledných textů.

Po celou dobu je průběh práce na textu zaznamenáván do interní databáze textů, kde jsou současně uloženy veškeré bibliografické informace. V okamžiku, kdy se chystá zveřejnění nové verze korpusu, jsou všechny texty z InterTextu vyexportovány ve formátu XML s kódováním UTF-8 společně s příslušnými zarovnávacími soubory, které párují jednotlivé věty vždy ze dvou zarovnaných jazykových verzí. Ke každému textu jsou navíc z databáze textů připojeny bibliografické informace. U jazyků, pro které máme k dispozici příslušné nástroje, je provedena lemmatizace a/nebo morfologické značkování (viz část 5). Nakonec jsou všechny texty převedeny do sloupcového formátu (každé slovo s případným lemmatem a značkou na nové řádce), který vyžaduje korpusový manažer Manatee (Rychlý 2000), a oindexovány. Zarovnávací soubory jsou zpracovány mimo Manatee speciálním nástrojem, který je připraví pro použití v rozhraní Park.

Kromě textů zpracovávaných touto standardní cestou se daří rozšiřovat obsah korpusu také o texty z hromadných zdrojů, jako jsou například publicistické texty z *Project Syndicate* (http://www.project-syndicate.org/), *Presseurop* (http://www.presseurop.eu/) nebo povídkový soubor *Můj rok 1989* z Goethe Institutu. Zpracování těchto textů jde mimo popsaný hlavní proud (OCR, korektury, databáze textů atd.), produkce výsledného XML je naopak ve všech krocích od stažení až po zarovnání automatická, ovšem s mnoha manuálními intervencemi, jejichž cílem je zejména kontrola struktury textů nutná pro zvýšení kvality výsledného plně automatického zarovnání. Automatické zarovnání těchto textů sice není manuálně revidováno, přesto ale zajišťuje dostatečnou spolehlivost pro zařazení do nabídky textů korpusu InterCorp, který je tak významně obohacen o další zdroje a druhy textů.

4. InterText a jeho hlavní rysy

InterText je editor paralelních textů vytvořený nově pro projekt InterCorp, ale s ohledem na možnost snadného použití i pro jiné podobné projekty. V současné době je k dispozici jen ve formě webového rozhraní spravujícího textové struktury uložené v SQL databázi na centrálním serveru, ale ve vývoji je i samostatná verze aplikace použitelná lokálně na osobních počítačích.

InterText umožňuje jak editaci zarovnání elementů (v našem případě obvykle vět) do paralelních segmentů (tj. skupin paralelních vět), tak korektury samotných textů. Opravovat lze nejen překlepy v textu, ale do jisté míry i strukturu textu: spojováním nebo rozdělováním lze měnit chybné dělení zarovnávaných elementů (vět). Každá změna textu se zaznamenává do protokolu, takže je možné zpětně kontrolovat potenciálně destruktivní činnost příliš horlivých editorů a v budoucnu může být implementována i jednoduchá reverzní funkce pro snadné odstranění takovýchto chybných či jinak nepatřičných "korektur". Možnost zasahování do textu či jeho struktury je navíc možné individuálně omezovat. Pokud dojde ke změně struktury textu, je InterText schopen automaticky přečíslovat identifikátory zarovnávaných elementů. Ošetřeny jsou také rizikové situace, kdy se jeden uživatel může pokusit změnit strukturu textové verze, která je současně zarovnána s jinou verzí – v našem případě hrozí takové nebezpečí pouze u českých pivotních textů: systém neumožní změnit strukturu tam, kde by mohlo dojít k porušení nějakého jiného, už existujícího zarovnání daného textu.

InterText sám o sobě nedokáže zarovnávat texty automaticky, ale byl za tímto účelem propojen s již existujícími osvědčenými nástroji – především jde o *hunalign,* lze ale použít také nástroj *TCA2* (http://gandalf.aksis.uib.no/tca2/; podrobnější popis viz Vondřička 2010). Paralelní texty se tedy mohou nejdřív zarovnat automaticky a uživatel pak výsledné zarovnání už jen kontroluje a opravuje. Editor si též po celou dobu uchovává detailní přehled o tom, které části textu (segmenty) byly zarovnány automaticky a které byly již zkontrolovány či opraveny ručně. Pro případy nekompletních překladů, s nimiž mají nástroje pro automatické zarovnání vážné problémy, nabízí také možnost dodatečného opětovného automatického zarovnání části textu po provedení dílčích úprav (ruční vyznačení chybějící části textu).

Pro usnadnění běžné práce byl InterText rozšířen také o funkce vyhledávání v textu a možnost zakládání záložek na problematická místa. Uživatelé si také mohou texty a jejich zarovnání exportovat do svého počítače, i když importování nových textů do systému je v projektu omezeno pouze na správce, kteří kontrolují jejich vstupní kvalitu a formát. Pro účely snadné správy velkého množství textů je k systému možno přistupovat i pomocí lokálních skriptů na serveru, a tak importovat či exportovat texty hromadně.

V zájmu hladké integrace do projektu InterCorp je InterText na různých úrovních propojen s databází textů a uživatelů, kteří se na práci v projektu InterCorp podílejí. Uživatelé jsou proto rozděleni do tří skupin s různými oprávněními přístupu a manipulace s texty a jejich zarovnáními: správci mají plný přístup k celému systému, koordinátoři pouze k textům a zarovnáním, za které jsou zodpovědní, a editoři jen k textům, které jim koordinátoři přidělili ke zpracování. Koordinátoři mají přitom možnost si pravomoc nad jednotlivými texty vzájemně předávat. Díky integraci s databází textů a spolupracovníků projektu InterCorp je možné automaticky delegovat texty patřičným zodpovědným koordinátorům a provádět automatické vyúčtování práce odvedené jednotlivými editory.

InterText byl vytvořen jako obecný nástroj, a snaží se proto oddělovat vlastnosti specifické pro projekt InterCorp od obecných principů paralelních korpusů. Je schopen pracovat s libovolnými texty ve formátu XML a díky využití univerzálního kódování Unicode ve všech komponentách může pracovat s libovolnými jazyky; tato podpora je v praxi závislá též na použitém internetovém prohlížeči a jeho vlastnostech. Umožňuje libovolně zarovnávat jakýkoli pár paralelních verzí jednoho textu, a není tedy omezen na jednu centrální (pivotní) jazykovou verzi. Zarovnávat je možné libovolné elementy XML struktury navzájem, i když

jen na jedné úrovni; nelze tedy například zarovnávat současně a nezávisle na různých úrovních hierarchické struktury textu, jako jsou kapitoly, odstavce, věty, slova atd.

InterText byl i s podrobnou dokumentací uvolněn ve formě zdrojových kódů (většinou v jazyce PHP) pod licencí GNU General Public License v3 a je k dispozici na stránce http://wanthalf.saga.cz/intertext. O jeho využití už projevilo zájem několik dalších evropských projektů paralelních korpusů.

5. Lemmatizace a morfologické značkování

Uživatelé i aplikace často pracují s texty, které nejsou nijak lingvisticky analyzovány. Vzhledem k tomu, že není možné dosáhnout stoprocentně spolehlivé automatické analýzy, ani její výsledky při obrovských kvantech dat ručně opravit, může být lingvistická anotace korpusu zavádějící. Přesto je pro řadu úkolů výhodné lingvistickou anotaci a výsledky automatických metod využívat, a to i u paralelních korpusů.

V současné době obsahuje korpus InterCorp morfologické údaje (zjednoznačněné podle kontextu) ve 14 jazycích, a to ve všech textech a u každého tvaru; z toho u 11 jazyků je uveden i základní tvar (lemma). U těchto jazyků lze značky a lemmata využít při zadávání dotazu i prohlížení výsledků. Aktuální stav, popis značek a údaje o nástrojích použitých na jednotlivé jazyky lze najít na stránce http://www.korpus.cz/intercorp/?req=page:info, podrobnější popis dále zmíněné problematiky značkování a tokenizace uvádí např. Rosen (2010).

Ve všech případech využíváme již existující nástroje založené na metodách strojového učení, u některých flektivních jazyků v kombinaci s automatickou morfologickou analýzou izolovaných tvarů. Z praktických důvodů se nesnažíme o vlastní, jednotný přístup k taxonomii slovních druhů a morfologických kategorií cestou modifikace existujících nástrojů a vytvářením vlastních trénovacích dat. Důsledkem je stav, kdy každý ze 14 takto zpracovaných jazyků si s sebou nese vlastní způsob kódování morfologických kategorií. Diskrepance čistě notační povahy by nepředstavovaly zvláštní problém, ale rozdíly mezi jednotlivými sadami morfologických značek často odrážejí odlišná teoretická východiska a vzájemně neslučitelné koncepce taxonomie. Tak např. česká sada značek, která vychází z tradiční klasifikace slovních druhů, obsahuje zvláštní třídy pro řadové číslovky, a také pro přivlastňovací, ukazovací a vztažná zájmena. Ve značkových sadách řady jiných jazyků se však všechny nebo aspoň některé z těchto slovních druhů považují za adjektiva. Někdy jsou naopak cizí značky podrobnější než české – rozlišují např. apelativa od proprií nebo adjektivní a substantivní užití ukazovacích zájmen.

Místo nerealistické vize převést všechny značky ze všech dotčených jazyků do jednotné obsahově a formátově konzistentní podoby a odpovídajícím způsobem upravit morfologickou anotaci textů počítáme v budoucnu s vytvořením taxonomie, která umožní popsat vztahy mezi nekompatibilními sadami značek prostřednictvím hierarchie kategorií, uspořádané podle míry obecnosti a podle kritéria klasifikace slovních druhů (morfologického, syntaktického a sémantického). Pak by bylo možné pracovat s jednotnou taxonomií, která by brala v úvahu i jen částečně se překrývající významy některých značek.

Závažnějším problémem než různorodé morfosyntaktické značkování jsou problémy s identifikací hranic slov – tokenizací, která značkování předchází a která musí být se značkováním v souladu. Stav v korpusu InterCorp je tedy opět ovlivněn použitým nástrojem a není stejný pro všechny jazyky. Spřežky typu *nač*, *abychom*, *udělals* a *tys* v češtině; *aux* a *cure-dents* ve francouzštině; *zum*, *deutsch-französisch* a *Jelzin-Ära* v němčině se nedělí a zůstávají jako jeden tvar s jednou značkou i lemmatem, na rozdíl od srovnatelných tvarů *nań*, *żebyśmy*, *zrobiłeś*, *tyś* a *niemiecko-rosyjski* v polštině; *padne-li* a *Tchaj-wan* v češtině; *dit-il* ve francouzštině nebo *can't* (tokenizováno jako *ca n't), I'm* a *John's* v angličtině. Korpusový manažer však nedokáže uchovávat původní a tokenizovaný tvar zároveň, takže nenajde tvary

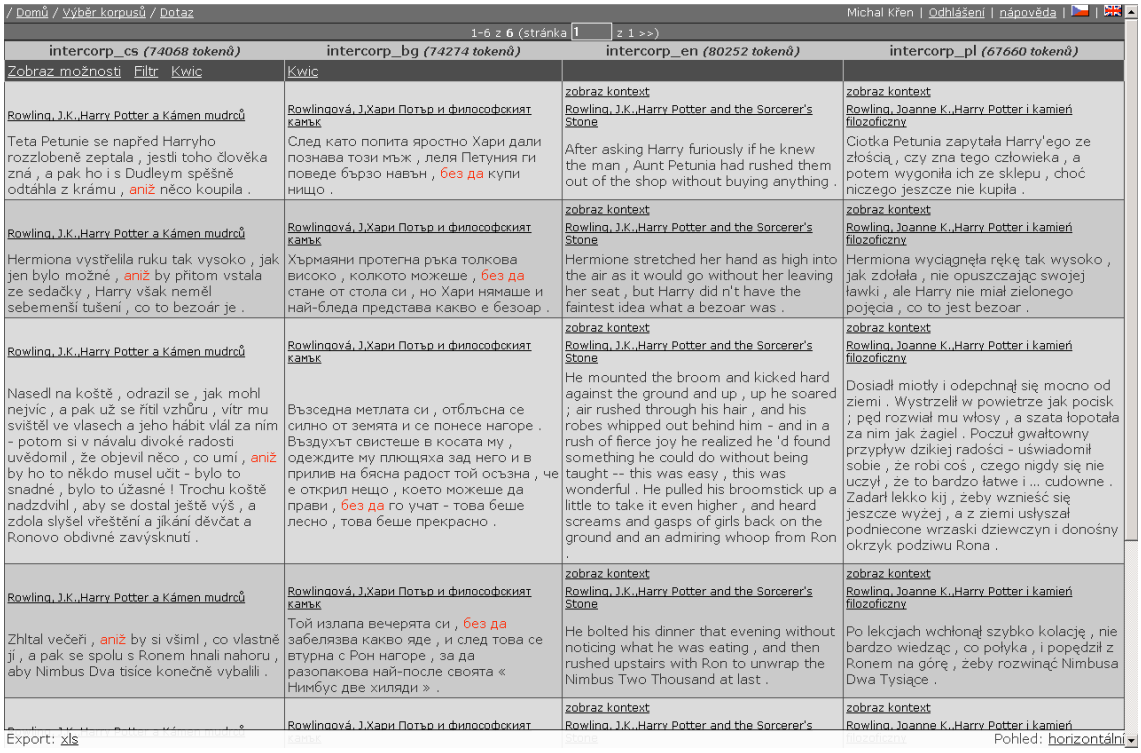
zadané při hledání v původní, nerozdělené podobě, ani je takto nezobrazí. Podobný problém je se španělskými víceslovnými výrazy, které se tokenizují, značkují i lemmatizují naopak jako jeden lexém: *Estados Unidos*, *al mismo tiempo*, *tendrán que*. Způsob, jakým jsou tyto jevy zpracovány a prezentovány uživateli, tedy bohužel zakrývá původní znění textu, které je tak přístupné pouze interně v podobě před lemmatizací a značkováním. Optimální řešení však předpokládá netriviální úpravu korpusového manažeru, kterou v dohledné době nelze očekávat. Uživatelé si proto těchto skutečností musejí být při práci s korpusem vědomi.

6. Park a další možnosti přístupu ke korpusu

Výsledky projektu byly zpočátku přístupné jenom lidem, kteří se na něm podíleli, a to ve formě textů zarovnaných ve dvou jazycích (čeština + cizí jazyk). To bylo způsobeno zejména tím, že pro vyhledávání v paralelních korpusech neexistovalo vhodné rozhraní. Pro interaktivní zarovnávání i vyhledávání v již zarovnaných paralelních textech se používal program ParaConc (Barlow 2002), který byl tehdy přes své nedostatky vyhodnocen jako nejvhodnější. jedním ze zásadních nedostatků, v jehož budoucí odstranění jsme (marně) doufali, je absence podpory Unicode. ParaConc je navíc samostatný program vázaný na platformu MS Windows, který vyžaduje lokální přístup k celým textům a neumožňuje oddělit vyhledávání od editace, což je pro účely projektu nevhodné. Později se také projevila jeho nespolehlivost, například generování ne vždy validního formátu XML při exportu. V současné době se ParaConc v rámci projektu nepoužívá, zarovnávání a veškeré korektury textů před vstupem do korpusu se od roku 2010 provádějí výhradně pomocí nástroje InterText, zatímco pro vyhledávání v hotových textech je všem uživatelům k dispozici rozhraní Park.

Park umožňuje prohledávat paralelní korpusy většího rozsahu (co do počtu slov i jazyků) uložené na centrálním serveru, řídit přístup k nim pomocí uživatelských hesel i omezovat pro jednotlivé uživatele kontext vyhledaných výrazů. V době zahájení prací na projektu však žádný takový nástroj neexistoval. Rozhodli jsme se proto vytvořit vlastní vyhledávací rozhraní jako nadstavbu nad korpusovým manažerem Manatee, používaným v ČNK pro vyhledávání v jednojazyčných korpusech. První verze tohoto nového rozhraní byla na adrese http://www.korpus.cz/Park/ veřejně spuštěna na podzim 2008, kdy byla také napojena na databázi uživatelů ČNK. Tím byl InterCorp de facto zařazen mezi veřejně přístupné korpusy ČNK a zároveň se stal prvním nereferenčním korpusem ČNK, tj. korpusem, který není od okamžiku svého zveřejnění neměnný, naopak se počítá s jeho neustálým rozšiřováním a zdokonalováním.

Park umožňuje uživateli plně využívat možností dotazovacího jazyka CQL při hledání v několika zadaných jazycích a textech současně, měnit způsob zobrazení výsledných konkordancí a také výsledky exportovat pro použití v dalších programech. Jeho vývoj se však od začátku potýkal s neustálými problémy, způsobenými tím, že aplikační rozhraní Manatee nepočítá s paralelními korpusy, ale hlavně absencí dokumentace k Manatee, což vývoj neúměrně zdržovalo. Další zdržení si vyžádalo přepracování Parku potřebné kvůli sjednocení českých verzí a přechodu na oddělené zarovnání (stand-off alignment), takže teprve počátkem roku 2011 přibyla k jeho základním funkcím možnost filtrování výsledků dotazu. K plnohodnotné práci s korpusem stále chybějí funkce pro vytváření náhodného vzorku, třídění výsledných konkordancí a základní statistické funkce, jakými jsou frekvenční distribuce nebo výpočet kolokací.



Obrázek 2: Ukázka práce s korpusem ve vyhledávacím rozhraní Park.

Z těchto důvodů byly na podzim roku 2009 zpřístupněny také všechny jednojazyčné verze korpusu pomocí standardního rozhraní Bonito (Rychlý 2007) na adrese http://www.korpus.cz/corpora/intercorp/. Tento alternativní způsob přístupu k textům InterCorpu sice neumožňuje využívat informace o zarovnání mezi jednotlivými jazyky, na druhé straně je ale možné k jednotlivým jazykovým verzím přistupovat jako k samostatným korpusům a používat při práci s nimi řadu funkcí Bonita, které v Parku chybějí.

Kromě výše zmíněných způsobů vyhledávání v korpusu, vhodných především pro běžné koncové uživatele, roste potřeba využívat stále rozsáhlejší a cennější data vzniklá v rámci projektu také pro jiné účely, zejména pro strojový překlad a počítačové zpracování přirozeného jazyka vůbec. V nejbližší budoucnosti se proto nedílnou součástí projektu stane také poskytování jazykových dat InterCorpu zájemcům z řad domácích i zahraničních vědeckých a výzkumných institucí, s omezeními danými pouze platnou legislativou, zejména autorským zákonem. Typickým příkladem takových dat mohou být česko-cizojazyčné překladové páry vět v promíchaném pořadí.

7. Plánovaná vylepšení a perspektivy do budoucna

Paralelních korpusů je v současné době řada, InterCorp je však ojedinělý tím, že obsahuje poměrně velké množství převážně beletristických textů v mnoha jazycích, navíc je veřejně přístupný přes vyhledávací rozhraní. Počtem jazyků i textů jej sice převyšuje korpus OPUS (http://opus.lingfil.uu.se/), ten je však koncipován jako “otevřený”, takže obsahuje jen některé specifické, volně dostupné typy textů (záznamy debat Evropského parlamentu, softwarové manuály, filmové titulky). Mezi dalšími částečně srovnatelnými projekty s větším počtem jazyků jsou to korpusy obsahující texty jednoho typu, např. JRC-Acquis (http://langtech.jrc.it/JRC-Acquis.html) se zákony Evropské Unie, nebo menší korpus “slovanských a jiných jazyků” ParaSol (http://www-korpus.uni-r.de/ParaSol/), který podobně jako InterCorp klade důraz na beletrii.

Preference beletrie jako výrazově bohatého žánru je u InterCorpu dána požadavky hlavní cílové skupiny uživatelů a zároveň spoluřešitelů projektu – akademických pracovníků a studentů jazykových oborů na filozofických fakultách, a při vhodném výběru textů není hlavní role beletrie u menšího celkového objemu textů na jazyk příliš na závadu. Z absence srovnatelného projektu je však zřejmé, že jde o poměrně náročný způsob výstavby korpusu, který je vhodné doplnit také jinými zdroji. Tato potřeba je zvláště patrná u jazyků, ve kterých bude objem beletristických textů vždy nedostatečný.

Problému nevyváženosti mezi jazyky a typy textů konkurují problémy techničtější povahy. Přetrvává především omezený repertoár funkcí ve vyhledávacím rozhraní a nezanedbatelné nejsou ani problémy související s kvalitou větné segmentace a zarovnáním, a také s lingvistickým značkováním – stále vysoký počet neanotovaných jazyků, tokenizace spřežek a víceslovných výrazů bránící intuitivnímu zadávání dotazů a disparátní značkové sady. Způsoby řešení všech těchto problémů jsou zjevné nebo již byly naznačeny, typografické nedostatky a chyby v segmentaci a zarovnávání plánujeme řešit možností navrhnout opravu chyby přímo ve vyhledávacím rozhraní.

Mezi bližší cíle patří plánované rozšíření korpusu o další žurnalistické a také právnické texty z webových zdrojů (PressEurop, Project Syndicate, Acquis Communautaire), a postupně i o další jazyky (čínština, arabština, albánština, romština, vietnamština). Lingvistickou anotaci chceme dále rozšiřovat na více jazyků. Dříve či později přijde na řadu i anotace syntaktická, v podobě strukturní i funkční, a zarovnávání po slovech, slovních spojeních a větných členech, které umožní kromě jiného i zvýraznění ekvivalentu hledaného výrazu v paralelních konkordancích.

Paralelní korpus je svou podstatou alespoň potenciálně spojen předivem vztahů s dalšími korpusy jednojazyčnými i paralelními, a to bez ohledu na institucionální či národní hranice. Proto je žádoucí vyšší integrace s českou částí ČNK v podobě společného vyhledávacího rozhraní. Konkordance získané na základě jednojazyčného dotazu by pak mohly být doplněny informací, zda nejsou k dispozici v některých dalších jazycích. V úvahu připadá i virtuální integrace více paralelních korpusů z více institucí a zemí tak, aby byly přístupné všechny najednou z jednoho vyhledávacího rozhraní.

Literatura

Barlow M., 2002, ParaConc: Concordance Software for Multilingual Parallel Corpora. In *Proceedings of the LREC 2002*, Las Palmas, 20-24.

Kiss T., J. Strunk, 2006, Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*, 32, č. 4, 485-525.

Rosen A., 2010, Morphological tags in parallel corpora. In *InterCorp: Exploring a Multilingual Corpus*, eds F. Čermák, P. Corness, A. Klégr, NLN, Praha, 205-234.

Rychlý P., 2000, *Korpusové manažery a jejich efektivní implementace*. Disertační práce, FI MU Brno.

Rychlý P., 2007, Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno, 65-70.

Varga D., L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy, 2005, Parallel Corpora for Medium Density Languages. In *Proceedings of the RANLP 2005*, Borovec, 590-596.

Vavřín M., A. Rosen, 2008, Intercorp: A Multilingual Parallel Corpus. In *Труды международной конференции "Корпусная лингвистика - 2008"*, Издательство СПбГУ, Санкт-Петербург, 97-104.

Vondřička P., 2010, TCA2 - nástroj pro zarovnávání paralelních textů. In *Mnohojazyčný korpus InterCorp: Možnosti studia*, eds F. Čermák, J. Kocek, NLN, Praha, 225-231.