Středoškolská technika 2013

Setkání a prezentace prací středoškolských studentů na ČVUT

VYUŽITÍ JAZYKOVÝCH KORPUSŮ NA STŘEDNÍCH ŠKOLÁCH

(S DŮRAZEM NA PARALELNÍ KORPUS INTERCORP)

Alžběta Vítková

Gymnázium Třebíč
Masarykovo náměstí 9/116, Třebíč

Prohlášení

Prohlašuji, že jsem svou práci vypracovala samostatně, použila jsem pouze podklady (literaturu, SW atd.) uvedené v přiloženém seznamu a postup při zpracování a dalším nakládání s prací je v souladu se zákonem č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) v platném znění.

V Třebíči dne 10. března 2013 podpis: ……

Poděkování

Tímto bych chtěla poděkovat PhDr. Olze Nádvorníkové, PhD. z Ústavu románských studií Filozofické fakulty Univerzity Karlovy za uvedení do problematiky korpusové lingvistiky, ale hlavně za cenné rady a trpělivost při vedení a konzultování této práce.

Děkuji také celému pracovnímu kolektivu Ústavu Českého národního korpusu Filozofické fakulty Univerzity Karlovy, zejména pak PhDr. Olze Richterové, za umožnění spolupráce a aktivní participace na projektu InterCorp.

Dále bych chtěla poděkovat Mgr. Monice Peroutkové z Gymnázia Třebíč, Městské knihovně v Třebíči, panu Františku Jůzovi a žákům tříd 6.G a 3.A Gymnázia Třebíč za podporu při realizaci praktické části práce.

ANOTACE

Ve své práci se snažím přiblížit studentům středních škol projekt paralelního korpusu InterCorp, který je vytvářen studenty a pedagogy Filozofické fakulty Univerzity Karlovy v Praze. Tento korpus je oblíbený mezi odborníky, ovšem já se zaměřuji na jeho propagaci na středních školách, protože jsem přesvědčená, že je užitečný i pro studenty a širokou veřejnost. Mým cílem je nejen motivovat žáky k jeho užívání, ale i odhalit případné nedostatky, které mohou studenta od práce s InterCorpem odradit, a pomoci Ústavu Českého národního korpusu FF UK navrhnout vylepšení.

KLÍČOVÁ SLOVA

Korpus – korpusová lingvistika – střední školy – popularizace

ANNOTATION

In my paper I try to present a project of parallel corpus InterCorp to students of high schools. This corpus is created by students and teachers of the Faculty of Arts of Charles University in Prague. This project is popular with experts, however, I focus on its advertising at high schools as I am convinced that it is really useful for students and the general public, too. The aim of my work is not only to motivate learners to use InterCorp but also to discover potential shortcomings which might discourage a student from working with this corpus. Then I would like to help the Institute of the Czech National Corpus to suggest some improvements.

KEY WORDS

Corpus – corpus linguistics – high schools – popularization

Obsah:

Úvod

S korpusovou lingvistikou a paralelním korpusem InterCorp jsem se poprvé setkala na stáži na Filozofické fakultě Univerzity Karlovy. Potkala jsem několik studentů, kteří pomocí InterCorpu zpracovávali seminární nebo diplomové práce, a položila jsem si otázku – může takový projekt být užitečný i mladším žákům, přesněji řečeno středoškolákům? A pokud ano, jak?

Svou práci jsem tedy rozdělila na dvě části, část teoretickou a praktickou. V teoretické části práce se snažím vymezit poznatky, jež byly o korpusové lingvistice a především korpusu InterCorp dosud shromážděny. Pokládám totiž za důležité, aby byl čtenář seznámen jak s obsahem korpusové lingvistiky, její historií a současným stavem výzkumu v této oblasti, tak s jejím praktickým využitím. Vzhledem k zaměření praktické části se soustřeďuji na vytvoření základního povědomí o paralelním korpusu InterCorp.

Praktickou část jsem se rozhodla koncipovat jako šetření. Mým cílem bylo zjistit, jak významnou roli hraje propagace v motivaci žáků při práci s InterCorpem, a na základě zpětné vazby navrhnout vylepšení popularizace tohoto korpusu. Ústav Českého národního korpusu Filozofické fakulty Univerzity Karlovy, který projekt zaštiťuje, totiž podle mého názoru nepropaguje InterCorp natolik, aby mohl být využíván širší veřejností. Ke zjištění odpovědi na své výzkumné otázky jsem vytvořila pracovní list o pěti úkolech pro dvě skupiny studentů francouzštiny. V první skupině jsem se snažila o popularizaci projektu – vedla jsem studenty k pochopení logiky a přínosu úkolů, nechala jsem je aktivně se zapojit do řešení relativně složitých problémů. Studenty z druhé skupiny jsem nechala poradit si s pracovním listem samostatně, bez jakékoli pomoci a vnější motivace. Všichni nakonec vyplnili dotazník, který zkoumal, co konkrétně a jak moc studenty motivuje při práci s paralelním korpusem InterCorp. Po vyhodnocení dotazníků bylo možné identifikovat rezervy a navrhnout způsoby jejich eliminace.

Svou prací bych tak chtěla pomoci popularizovat v České republice nepříliš známou disciplínu, a to korpusovou lingvistiku, konkrétně paralelní korpus InterCorp.

1 Korpus

1.1 Korpus a korpusová lingvistika

Korpusová lingvistika je disciplína lingvistiky zkoumající jazyk pomocí elektronických jazykových korpusů a zabývající se i výstavbou těchto korpusů, jejich zpracováním a příslušnou metodologií.[1] Jde o disciplínu poměrně mladou, protože pro plné využívání jejích možností je třeba mít pevné technické zázemí. Na významu tedy získala až v době rozšiřování prvních osobních počítačů v 90. letech 20. století.

František Čermák, jeden ze zakladatelů korpusové lingvistiky u nás, uvádí následující definici korpusu: „Korpus je rozsáhlý, elektronicky uložený, zpracovávaný a přístupný soubor jazykových dat ve standardizovaném formátu, tj. jednotlivých forem a textových celků a/nebo vzorků psaného a mluveného jazyka, cíleně shromážděný jako referenční zdroj pro vědecké studium jazyka a pro zpracování užitných jazykových nástrojů, příruček a jiných artefaktů.“[2] V praxi to znamená, že korpus je jakási databanka lingvistického materiálu, vhodného jak pro výzkum lexikální a sémantický (význam slov), tak pro gramatický (zkoumání gramatických jevů a tendencí jazyka). Obsáhlejší korpusy, jako je např. i Český národní korpus *(viz kapitola 2)*, obsahují kromě toho ještě i diachronní korpusy, které umožňují zkoumat vývoj jazyka.

Korpus je bezpochyby jedním z nejvhodnějších zdrojů lingvistických dat, který převyšuje jakýkoli jiný dostupný zdroj dat jak kvalitou, tak kvantitou i vytěžitelností.

1.1.1 Bližší vymezení korpusu

Jazykový korpus má mnoho společného s internetovými vyhledávači (Google), ale má navíc různé přednosti. Především jde o spolehlivost údajů – je jasně dané, kolikrát se daný jev v korpusu nachází a s jakým množstvím dat se pracovalo. Citační normy pro korpus jsou také pevně vymezené. Dále jsou data vyvážená, je dbáno o to, aby byl výsledek co nejpřesnější a např. beletrie nebyla upřednostňována před publicistickými texty. Velkou výhodou je i znalost kontextu a možnost vyhledání konkordancí.[3] Pomocí konkordance, což je bezprostřední kontext hledaného výrazu, můžeme třeba zjistit, jaké slovo nejčastěji stojí před slovem *nehty*. Můžeme se tak vyhnout chybám typu *pěstované nehty* (namísto *pěstěné*).

Ze základních pojmů korpusové lingvistiky je třeba zmínit *token* = výskyt slovního tvaru v korpusu, *typ* = slovní tvar jako takový (např. *rukama*, *pohybuji*), *lemma* = základní slovní tvar (např. ruce, pohybovat), *konkordance* = výskyt slovních tvarů s kontextem, *tag* = značka.

[1] KOCEK, J. – KOPŘIVOVÁ, M. – KUČERA, K. (eds): *Český národní korpus – úvod a příručka uživatele.* ÚČNK FF UK, Praha 2000, s. 4.
[2] ČERMÁK, F. – BLATNÁ, R. (eds): *Manuál lexikografie*. Nakladatelství a vydavatelství H & H, Praha 1995, s. 52.
[3] URL: <http://ucnk.ff.cuni.cz/co_je_korpus.php> [Cit. 14.12.2012].

1.1.2 Historie korpusu a korpusové lingvistiky

Už na začátku 17. století se začaly objevovat první pokusy o vytvoření rozsáhlého lexikografického díla, např. ve Velké Británii vytvořil Samuel Johnson svůj Dictionary of English Language. Ačkoli byl vytvářen pouze ručně, obsahoval přes 1 milion slov. V podobném duchu pak pokračovaly práce na území anglofonních zemí.[4] Právě tyto země, zejména Velká Británie, by se daly označit za průkopníky korpusové lingvistiky a dodnes určují trendy v této oblasti.

První počítačově zpracované korpusy pro lexikografy se začaly objevovat na počátku 60. let 20. století, respektive už roku 1959, kdy byl vytvořen anglický The Survey of English Usage.[5] Za první korpus jako takový ale můžeme považovat až americký, počítačově zpracovaný The Brown Corpus, poprvé zveřejněný roku 1961.[6] Čítal okolo milionu slov.

Tvorba korpusů ovšem stále ještě zůstávala nepříliš oblíbenou disciplínou. V 60. letech totiž Noam Chomsky, jeden ze zakladatelů generativní lingvistiky, naprosto odmítl korpusy a vůbec celou korpusovou lingvistiku.[7] Počítačové technologie byly na takové úrovni, že se nedal předpokládat výraznější vývoj, proto o korpusy jevil zájem málokdo.[8]

V 70. letech se postupně objevovaly různé korpusy podobné Brown Corpus, např. Lancaster/Oslo-Bergen (LOB) Corpus, vytvořený Geoffrey Leechem. Z dalších, navazujících na Brown Corpus, stojí za zmínku Kolhapur (korpus indické angličtiny), Wellington (korpus novozélandské angličtiny), Australian Corpus of English (korpus australské angličtiny), The Frown Corpus (korpus americké angličtiny počátku 90. let) a The FLOB Corpus (korpus britské angličtiny 90. let).[9] Od té doby na poli korpusové lingvistiky převládají anglicky mluvící země, korpusy angličtiny se dodnes řadí mezi nejpropracovanější a nejpestřejší. Angličtina se také používá jako komunikační jazyk nejen mezi korpusovými lingvisty, ale i ostatními uživateli korpusů.[10]

90. léta 20. století znamenala všeobecně velký rozvoj korpusů.[11] Začaly být využívány nové počítačové technologie, softwary a hlavně se masově rozšířil internet. Kromě korpusů

---

[4] Poslední velký slovník angličtiny, sestavený bez elektronické databáze, vyšel roku 1961.

[5] URL: <http://www.ucl.ac.uk/english-usage/about/history.htm> [Cit. 15.12.2012].

[6] URL: <http://icame.uib.no/brown/bcm.html> [Cit. 15.12.2012].
Na jeho tvorbě se mimo jiné podílel i lingvista českého původu Henry (Jindřich) Kučera, který jej ve spolupráci s W. Nelsonem Francisem dokončil roku 1964.

[7] V jednom rozhovoru z roku 2004 Noam Chomsky na otázku, co si myslí o současné tendenci považovat korpusy za významnou složku výzkumu, odpověděl doslova: *„Korpusová lingvistika neexistuje. To je to samé jako tvrdit, že fyzikové a chemici si místo toho, aby se spolehli na pokusy, raději vezmou videokazety. Na těch videokazetách bude nahráno, co se děje v každém okamžiku ve světě. Pak sesbírají ty videokazety a v nich možná teoreticky najdou nějakou generalizaci. Víte, jenže věda takhle nefunguje.“* (přeloženo z anglického originálu, dostupného z URL: <http://www.corpus4u.org/forum/upload/forum/2005052811133696.pdf> [Cit. 15.12.2012].)

[8] NÁDVORNÍKOVÁ, O.: *Analýza predikačního potenciálu francouzských tvarů na –ANT.* Diplomová práce, Filozofická fakulta Univerzity Karlovy v Praze, ved. J. Tláskal, Praha 2003, s. 3.

[9] URL: <http://en.wikipedia.org/wiki/Corpus_linguistics> [Cit. 15.12.2012].

[10] O tom svědčí např. i terminologie – i pro Český národní korpus jsou klíčová právě anglická slova, jako např. *tag*, *word*, *token* atd.

[11] Například první počítačový korpus mluvených textů byl ale vytvořen už roku 1971 v rámci The Montreal French Project. Obsahoval milion slov, což inspirovalo lingvistku Shanu Poplackovou k vytvoření mnohem většího korpusu mluvené francouzštiny v okolí Ottawy.

anglického jazyka tak vznikají korpusy jiných národních jazyků, diachronní korpusy i korpusy speciální: korpusy soukromé korespondence, telefonních hovorů, tiskových konferencí v Bílém domě nebo korpus právních dokumentů.

1.1.3 Popis a využití korpusů

Data v korpusu jsou získávána trojím způsobem, a to konverzí elektronicky uložených textů, skenováním nebo manuálním přepisem tištěných textů, poznámek či rukopisů.

Poté je korpusy třeba *tagovat* (z angl. slova *tag* = značka) neboli *značkovat*. K jednotlivým tvarům nebo *lemmatům* (základním tvarům) jsou přidávány mluvnické (morfologické) značky, pomocí nichž pak může uživatel vyhledávat například pouze podstatná jména rodu mužského neživotného apod. Morfologické značky neboli *tagy* jsou tedy velmi důležité zejména pro morfologickou analýzu.

Další podstatnou vlastností korpusu je to, jestli je nebo není *lemmatizovaný*. Lemmatizace neboli proces, kdy je slovo převedeno do svého základního tvaru,[12] je ale užitečná zejména u flektivních jazyků, jako jsou jazyky slovanské (mezi nimi i čeština) nebo latina. V případě jazyků analytických (např. angličtina) je lemmatizace méně používaná. Protože angličtina nezná pády, např. slovo *table* bude mít stále stejný tvar nehledě na kontext (*on the table, above the table*…). Není to ale pravidlem, protože další z analytických jazyků, francouzština, lemma využije při časování (např. sloveso *écrire* – *j'écris, il écrivait, ils écriraient, écrivant* …).

Důležité je i vymezit pojem konkordance, což je uspořádaný soubor výskytů slovních tvarů s jejich textovým okolím, často s označením zdroje. Standardním formátem konkordancí je KWIC (Key Word in Context), který umožňuje i volbu délky kontextu po obou stranách, uspořádání alfabeticky vzestupně nebo sestupně nebo frekvenční uspořádání (vzestupně i sestupně). Dále je ovšem možné hledat i podle souvýskytu dalšího slova.[13]

Na tvorbě korpusů se ovšem podílejí nejen odborníci, ale i studenti (v České republice převážně filozofických fakult, matematicko-fyzikálních fakult nebo fakult informatiky). Studenti se tedy naučí nejen pracovat s korpusem, ale i vidět problematiku z druhé strany, kdy se některé operace musí zadávat ručně, což je velmi pracné. Podstatné je i takzvané *alignování* neboli zarovnávání, které umožňuje zvláště v případě korpusu paralelního vyhledávat konkordance.[14]

Jak uvádí Čermák s Blatnou, je ovšem také nutné vědět, že i tagování má svá úskalí. Vždy odráží jen určitou jazykovou teorii svých tvůrců, tím pádem je relativní, protože jiní lidé mohou mít na pojetí a označení dat jiný názor. Data se svou interpretací tedy zjednodušují a deformují. Tagování nikdy nebude úplné a zcela bez chyb a nikdy nevytlačí

[12] Základním tvarem je myšlen tvar slovníkový, tj. infinitiv u verb, nominativ u substantiv apod. Pomocí lemmatu jsou pak nalezeny všechny tvary tohoto slova (např. u slovesa *chtít – chtěl by, chci, budou chtít, chtěl jsi*, …)

[13] ČERMÁK, F. – BLATNÁ, R. (eds): *Manuál lexikografie*. Nakladatelství a vydavatelství H & H, Praha 1995, s. 55.

[14] Zarovnává se většinou podle delších úseků (od několika slov po několik vět).

prosté neoznačkované korpusy, které jediné uchovávají relativně autentický text bez vnášené interpretace.[15]

Přístupy výzkumu se pak dají z obecného hlediska rozdělit na *corpus-based* (tedy výzkum na korpusu založený, resp. korpusem ověřovaný) a *computer-driven* (tedy výzkum korpusem řízený či inspirovaný). Computer-driven je způsob v českém prostředí poměrně nový a nepříliš užívaný. Například se využívá pro výzkum synonymie (rozdíl ve využívání dvojic synonymních *slov medicína-lékařství, abonent-předplatitel*), dále také pro rozdělování slov do slovních druhů nebo pro zkoumání terminologie.

Korpusy jsou přínosné ještě z několika dalších hledisek.

Nejprve se nabízí přínos pro práci lingvistů a hlavně lexikografů. Zaprvé je díky korpusu je možné ověřit, zdali dané slovo (či daný tvar slova) existuje. To je užitečné například při skloňování některých problematických výrazů (*školné*: 7. pád = *školném* či *školným*?). Zadruhé lze z korpusu vyčíst, jak je daná varianta vhodná v dané komunikační situaci. Dá se tak vyhnout například použití slova *zřídit* (ve smyslu *zranit*) ve formálním projevu. Místo toho použijeme právě slovo *zranit*. Zatřetí se dá posuzovat i synchronie a diachronie daného výrazu – zdali jde o archaismus, historismus (pak se dá zkoumat i vývoj onoho slova) nebo jde-li o slovo běžně užívané i v dnešní době. Začtvrté je možné zjistit formální informace o prvku (např. s jakým pádem se váže předložka *vedle*), povahu jeho okolí (např. jestli se dané slovo vyskytuje na začátku, nebo konci věty). Dále je možno kontrolovat a analyzovat metajazyk. Metajazykem je myšlena jakási „řeč o řeči“, např. *NaCl = chlorid sodný.*

Samozřejmě se dají zkoumat i kolokace, frazeologie nebo konkordance. Korpusy jsou často využívány i při tvorbě slovníků speciálních, např. frekvenčních, retrográdních, homonymních, synonymních a opozitních, slovníků morfémů nebo tezauru. Na základě korpusů se vytvářejí i aplikované slovníky a lexikální příručky.

Obecně vzato existuje jen málo oborů, které se dají studovat primárně jinak než skrze jazyk. Korpus je tedy užitečný i pro nelingvisty (obsahoví specialisté, jako např. historikové, mohou využít korpus při sledování terminologie). Ze zcela laického hlediska je možné v korpusu vyhledávat informace podobné jako u lingvistů, ovšem z daleko pragmatičtějších důvodů, např. jestli opravdu existuje citoslovce *kiš, kiš*. V tomto ohledu ovšem není korpus zcela spolehlivý (vzhledem k povaze dat, která nemohou obsáhnout všechny oblasti). Např. slovo *zeugma*, prokazatelně existující, korpus SYN2010 nenajde, jelikož nebylo užito v jeho databázi.

Pro cizince může korpus jazyka, který není jeho rodným, znamenat ucelenou databanku informací (například zdali se užívá častěji tvar *policisti* nebo *policisté*, v jakých případech je vhodné je použít a v jakých ne).

[15] ČERMÁK, F. (ed.): *Korpusová lingvistika Praha 2011 – 2 Výzkum a výstavba korpusů*. Nakladatelství Lidové noviny, Praha 2011, s. 15.

I domácí student ovšem může ocenit korpus, například když píše písemnou práci. Korpus sice nedokáže nahradit Pravidla českého pravopisu nebo Slovník spisovné češtiny, ale může pomoci uživateli s výběrem jazykových prostředků a s jejich vhodným zvolením v souvislosti s komunikační situací.

1.2 Typologie korpusů

Korpusy mohou být velmi rozličné a jejich rozdělení nemusí být zcela jednoznačné.[16] Čermák a Blatná uvádějí, že se nabízí rozdělení dle hledisek reprezentativnosti, stáří textů, původu získaných dat a počtu jazyků:

a) z hlediska reprezentativnosti a účelu na obecné a specializované korpusy
b) z hlediska stáří textů na synchronní a diachronní korpusy
c) z hlediska původu získaných dat na korpusy psaných a mluvených textů
d) z hlediska počtu jazyků na jednojazyčné a paralelní korpusy

Existuje však ještě mnoho dalších dělení korpusů, například na referenční a nereferenční. Referenční korpus zůstává neměnný po celou dobu existence, kdežto korpus nereferenční je v průběhu let postupně doplňován a upravován.[17]

1.2.1 Obecné a speciální korpusy

Reprezentativnost je klíčovým pojmem typologie korpusů, je ovšem nutné ji chápat relativně. Obecné korpusy bývají nazývány nejreprezentativnějšími právě z důvodu, že jsou velmi obsáhlé (v řádech desítek milionů slov), tudíž vzorek bude spolehlivý. Na obecných korpusech se nejčastěji zkoumá stav současného jazyka, jsou tak velmi používané.

Mezi obecné korpusy v Českém národním korpusu (ČNK) patří např. SYN, který vznikl spojením dílčích korpusů (SYN2000, SYN2005, SYN2010 a další, podrobně v kapitole 2.2). Patří k obsáhlejším korpusům, velikostí 1,3 miliardy slovních tvarů je největší v České republice.[18]

Druhou částí národních korpusů pak bývá korpus speciální. Na rozdíl od obecného monitoruje pouze určitou specifickou část jazyka (například romány pouze jednoho autora, studentský slang, dialekty nebo korespondenci). Díky speciálním korpusům tak můžeme analyzovat mimo jiné vyjadřovací schopnosti určitého autora (analyzovat styl psaní Bohumila Hrabala apod.). V ČNK se jedná např. o korpus ORWELL.[19]

1.2.2 Synchronní a diachronní korpusy

Synchronní korpus je pravděpodobně nejběžnějším korpusem. Mapuje současný jazyk, proto zahrnuje jen několik posledních desetiletí, jelikož jazyk se stále vyvíjí. Na základě

[16] ČERMÁK, F. – BLATNÁ, R. (eds): *Manuál lexikografie*. Nakladatelství a vydavatelství H & H, Praha 1995, s. 50-53.
[17] URL: <http://ucnk.ff.cuni.cz/n_neref.html> [Cit. 17.12.2012].
[18] URL: <http://ucnk.ff.cuni.cz/struktura.php> [Cit. 17.12.2012].
[19] Korpus ORWELL je ručně označkovaný korpus románu George Orwella „*1984*“.

synchronních korpusů lze zkoumat tendence jazyka současného, např. při zpracování frazeologie (ustálená slovní spojení, jako např. *mít namále*).

Diachronní korpus se naproti tomu věnuje několika (popř. všem) stádiím jazyka. Oproti synchronnímu chápe jinak pojem reprezentativnosti. V diachronním korpusu ji totiž lze vztáhnout pouze na dané časové období, v němž byly texty sepsány. Žánrová nevyváženost je dalším typickým znakem diachronního korpusu, obvykle se zde vyskytují texty náboženské, veršované a jiné speciální.[20] Na diachronních korpusech lze zkoumat např. vývoj daného slova (změny jeho pravopisu – ortografický/morfologický vývoj, významu – lexikální/etymologický vývoj, postavení ve větě – syntaktický vývoj apod.)

Naprostá většina korpusů v ČNK je synchronní, mezi diachronní patří např. DIAKORP nebo DOTKO.

1.2.3 Korpusy psaných a mluvených textů

Korpusy psaných textů (nebo také psané korpusy) jsou klasickým a nejběžnějším zástupcem korpusů.

Korpusy mluvených textů (mluvené korpusy) bývají synchronní. Dříve totiž nebylo možné sbírat materiál k výzkumu, protože technika nebyla na dostatečné úrovni, a tak žádné diachronní korpusy mluvené neexistují. V případě mluvených korpusů je velmi složité získat autentické vzorky mluveného jazyka i v dnešní době, tudíž vznikají jen omezenější menší korpusy. Hlasové záznamy, z nichž je mluvený korpus složený, zohledňují různé sociolingvistické faktory: pohlaví, věk a vzdělání mluvčích a typ (formální – odpovědi na dané otázky; a neformální – spontánní mluvní projevy). S mluvenými korpusy souvisí i tvorba korpusů nářečních (dialektických), které se dají získat právě jen díky mluveným projevům mluvčích. Dají se tak využít v oblasti lexikologie a morfologie.

ČNK zahrnuje jak korpusy psaných textů (již zmiňovaný SYN), tak korpusy textů mluvených, např. ORAL2008.

1.2.4 Jednojazyčné a paralelní korpusy

Jednojazyčný korpus zahrnuje jazyková data pouze jediného (zpravidla národního) jazyka, kdežto korpus paralelní se věnuje dvěma a více jazykům. Tyto korpusy jsou vytvářené z překladů a jejich originálů.

Korpusy paralelní jsou poměrně novým odvětvím korpusové lingvistiky. U nás v roce 2005 vznikl InterCorp *(viz kapitola 3)*, ve světě dále Europarl Parallel Corpus, OPUS a jiné.[21]

Další kapitola se bude zabývat Českým národním korpusem, jeho popisem a využitím.

[20] KOCEK, J. – KOPŘIVOVÁ, M. – KUČERA, K. (eds): *Český národní korpus – úvod a příručka uživatele.* ÚČNK FF UK, Praha 2000, s. 8-9.

[21] Jedním z evropských paralelních korpusů devadesátých let dvacátého století byl paralelní korpus románu George Orwella 1984, který vznikl v rámci projektu Multext-East.

2 Český národní korpus

Český národní korpus (ČNK) je kontinuální projekt, jehož produkty (jednotlivé konkrétní korpusy) mapují a monitorují různé podoby českého jazyka s cílem zpřístupnit uživatelům co nejbohatší zdroj jazykových dat a příslušné nástroje k jejich využívání.[22]

2.1 Historie Českého národního korpusu

Prvním krokem směrem ke korpusům u nás byly počítačově čitelné slovníky (např. Retrográdní slovník současné češtiny, vydaný nakladatelstvím Academia roku 1986). Daleko přínosnější pro budoucí projekt ČNK ovšem byly zkušenosti z práce na několika slovníkových databázích, vznikající již v 80. a na počátku 90. let.

Posléze bylo vytvořeno několik databází, čítajících mezi sto a dvěma sty tisíci slovy. Významný podíl tvořily databáze Filozofické fakulty Masarykovy univerzity nebo např. práce J. Hajiče a J. Drózda.

Databáze se postupně rozrůstaly, ovšem izolovaně a nekoordinovaně. Roku 1988 tedy vznikla pod záštitou Kybernetické společnosti „Iniciativní skupina pro přípravu počítačových korpusů textů a slovníků“, která si kladla za cíl sjednotit metody lexikografické práce se zahraničím a zlepšit koordinaci jednotlivých projektů.

První podoba toho, čemu dnes říkáme ČNK, vznikla roku 1991, kdy se skupina lingvistů a matematiků rozhodla sdružit a vytvořit „Počítačový fond češtiny“. ÚČNK jako takový byl založen roku 1994 na základě iniciativy řady jednotlivců z různých pracovišť, kteří začali už před lety pociťovat naléhavou potřebu vybudovat velký korpus představující dostatečnou materiálovou (datovou) základnu umožňující tvorbu nových, kvalitativně lepších slovníků češtiny, gramatik a dalších jazykových příruček. Od té doby je Ústav Českého národního korpusu veden prof. Františkem Čermákem.[23]

2.2 Popis Českého národního korpusu

V ČNK se nacházejí korpusy synchronní (psaného i mluveného jazyka), diachronní a paralelní. Paralelnímu korpusu InterCorp se budu podrobně věnovat v kapitole 3.

Rozdíl mezi Českým národním korpusem a národními korpusy ostatních zemí je mimo jiné také cenový. British National Corpus (Britský národní korpus) nebo Frantext (francouzský národní korpus) jsou zpoplatněné, kdežto část (přibližně pětina) ČNK je volně přístupná pro kohokoliv, zbylá část po registraci.

Přístup do celého ČNK je tedy zdarma, je ovšem nutné získat přihlašovací jméno a heslo. Stačí vyplnit elektronický formulář na http://ucnk.ff.cuni.cz/prohlaseni.php, který umožní plný přístup do korpusu. Stejné přihlašovací jméno a heslo slouží jak pro ČNK, tak pro InterCorp.

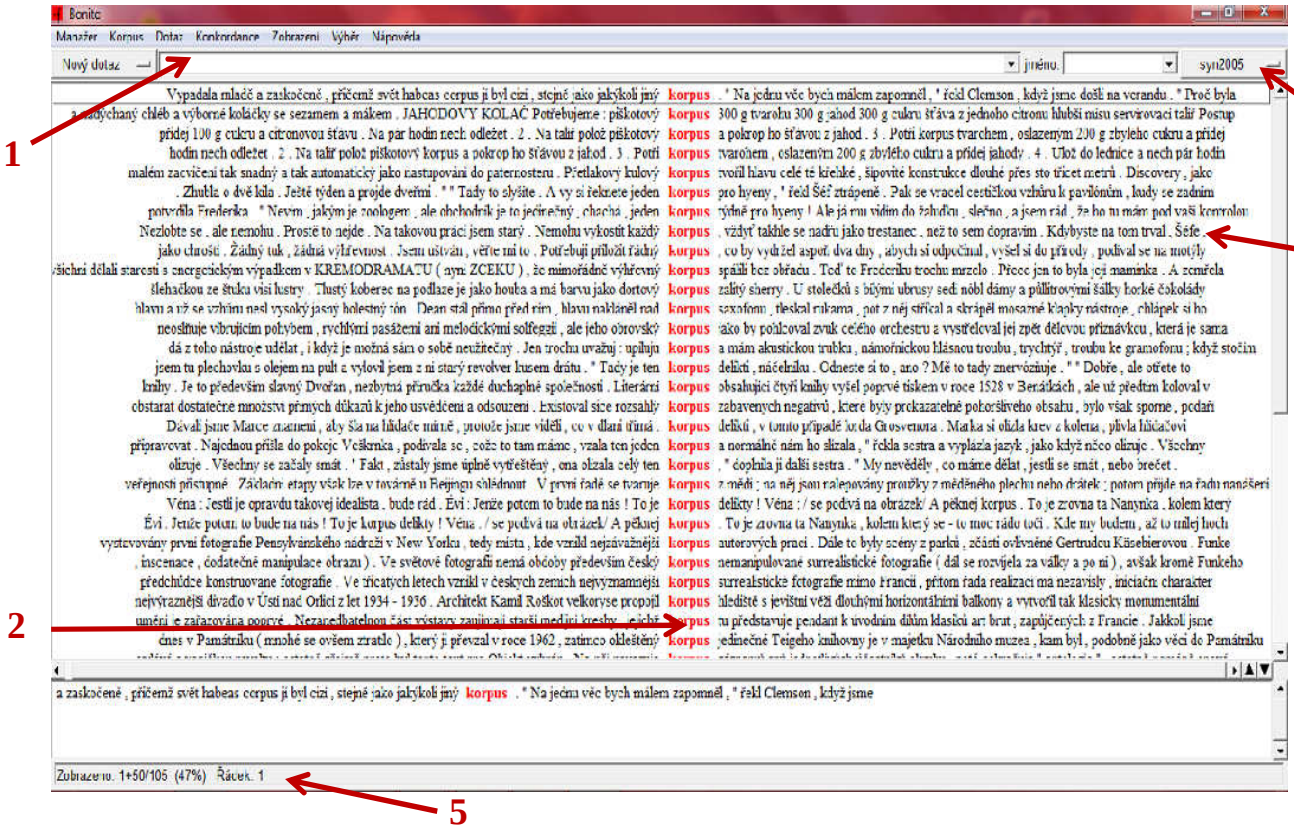
---

[22] KOCEK, J. – KOPŘIVOVÁ, M. – KUČERA, K. (eds): *Český národní korpus – úvod a příručka uživatele.* ÚČNK FF UK, Praha 2000, s. 10.

[23] ŠULC, M. *Korpusová lingvistika: první vstup.* Nakladatelství Karolinum, Praha 1999, s. 47.

Pro plné využití možností ČNK jsou dvě možnosti. Zaprvé lze použít přístup přes webové stránky http://www.korpus.cz/hledat_v_cnk.php.[24] Zadruhé je možné stáhnout si zdarma korpusový manažer Bonito ze stránek http://ucnk.ff.cuni.cz/bonito/instalace.php. Bonito pak umí například zobrazit konkordance (kontext), vyhledávat podle lemmat, vytvářet statistiky, subkorpusy a podobně.

Obr. 1: Ukázka vzhledu programu Bonito (vyhledávání slova *korpus*) ve formátu KWIC



Vysvětlivky čísel: 1 – dotazový řádek; 2 – vyhledaný výraz (KWIC – Key Word in Context); 3 – výběr korpusu; 4 – konkordanční řádek; 5 – stavový řádek.

Na obrázku je možné vidět vyhledávání slovního tvaru *korpus* v korpusu SYN2005. V horní liště jsou na výběr možnosti dalšího rozšíření informací o vyhledávání (např. statistiky, vytvoření subkorpusu, export výsledků, zdroje textů aj.). Výsledky jsou zarovnány podle hledaného výrazu (*korpus*). Z obrázku lze i vyčíst, že v korpusu SYN2005 je slovo *korpus* obsaženo 105krát.

[24] Webové vyhledávání funguje pomocí rozhraní NoSketch Engine, což je projekt, který kombinuje korpusové manažery Manatee a Bonito.

Texty pro Český národní korpus jsou získávány pěti různými způsoby:[25]

I. prostřednictvím smluv s nakladateli a vydavateli
II. využíváním textů dostupných na internetu
III. skenováním
IV. manuálním přepisem
V. darem od autorů

2.2.1 Korpusy psaného jazyka (synchronní)

V českém jazyce neexistují žádná striktní kritéria pro stanovení časové hranice mezi diachronními a synchronními korpusy. Podle Kocka, Kopřivové a Kučery byl v rámci ČNK přijat jako hranice rok 1990 jak v beletrii, tak v publicistických i odborných textech. Do synchronního korpusu se ovšem zařazují i v současnosti čtení starší autoři, kteří se narodili roku 1880 a později. Dále jsou do korpusu zařazeny také knihy publikované od konce 2. světové války (tyto texty jsou ovšem v korpusu zastoupeny řidčeji).[26]

Dále je zaznamenána snaha o vyváženost korpusu, takže např. beletrie je v korpusu pouhých 15%. V ČNK tedy není obsažena veškerá literatura daného období, ale jen určitý reprezentativní vzorek.

Co se týče složení, Český národní korpus obsahuje 15% krásné literatury *(viz výše)*, 25% naučné literatury (informativních nenovinových textů) a 60% publicistických textů (novinových textů denního tisku).[27]

Mezi synchronní korpusy psaného jazyka v Českém národním korpusu patří v první řadě korpus SYN, který se dále člení na korpusy SYN2000, SYN2005, SYN2006PUB, SYN2009PUB a SYN2010. Dohromady čítá 1,3 miliardy slov. Korpus není referenční, to znamená, že se v průběhu let stále vyvíjí. Všechny subkorpusy jsou lemmatizované i morfologicky označkované.[28]

Mezi další synchronní korpusy psaného jazyka v ČNK patří FSC2000, stomilionový korpus, který byl použit jako referenční zdroj Frekvenčního slovníku češtiny.[29] Dalším zástupcem synchronních psaných korpusů je CZESL-PLAIN, dvoumilionový žákovský korpus češtiny z roku 2012. Dále je k dispozici LINK, téměř dvoumilionový korpus sestavený z odborných lingvistických textů. Byl vytvořen roku 2008. Druhým nejméně obsáhlým korpusem Českého národního korpusu je osmisettisícový korpus KSK-DOPISY z roku 2006, který se věnuje korespondenci přelomu 20. a 21. století. Posledním synchronním psaným

[25] KOCEK, J. – KOPŘIVOVÁ, M. – KUČERA, K. (eds): *Český národní korpus – úvod a příručka uživatele.* ÚČNK FF UK, Praha 2000, s. 19.
[26] KOCEK, J. – KOPŘIVOVÁ, M. – KUČERA, K. (eds): *Český národní korpus – úvod a příručka uživatele.* ÚČNK FF UK, Praha 2000, s. 13.
[27] Beletrie je v korpusu zastoupena z více než dvou třetin beletrie prózou, zatímco poezie tvoří jednu dvacetinu. V odborných textech jsou téměř rovnoměrně rozloženy vědy o umění, sociální vědy, přírodní vědy, technika, ekonomie a řízení a životní styl. Malé procento tvoří i administrativa, víra a náboženství a právo a bezpečnost.
[28] URL: <http://ucnk.ff.cuni.cz/struktura.php> [Cit. 26.12.2012].
[29] ČERMÁK, F. – KŘEN, M. (eds): *Frekvenční slovník češtiny*. Nakladatelství Lidové noviny, Praha 2004.

korpusem je osmdesátitisícový ORWELL, korpus románu George Orwella „1984“. Korpus byl zveřejněn roku 2003.

2.2.2 Korpusy mluveného jazyka (synchronní)

Žádný korpus mluveného jazyka není označkovaný ani lemmatizovaný. Prvním z nich je ORAL2006, milionový korpus neformální mluvené češtiny. Navazuje na něj ORAL2008, milionový sociolingvisticky vyvážený korpus neformální mluvené češtiny. Dalším korpusem mluveného jazyka je SCHOLA2010, korpus vyučovacích hodin z roku 2010. Má asi 790 tisíc slov. Dále jsou k dispozici PMK a BMK – Pražský mluvený korpus a Brněnský mluvený korpus. První jmenovaný byl zveřejněn roku 2001, druhý roku 2002. PMK čítá přibližně 675 tisíc slov, BMK o něco méně – 490 tisíc slov.

Díky těmto korpusům je možné zkoumat hlavně lexikologii a morfologii. Pražský a Brněnský mluvený korpus umožňují vyhledávat dialektologické rozdíly, např. tendence v Čechách prodlužovat samohlásky a na Moravě je zkracovat. SCHOLA2010 zaznamenává mluvu ve školách, dá se na ní studovat např. formálnost nebo neformálnost výpovědí jak žáků, tak učitelů.

2.2.3 Diachronní korpusy

Diachronní korpusy ČNK, zvláště pak DIAKORP, jsou budovány s cílem vytvořit elektronickou materiálovou základnu pro výzkum vývoje češtiny od prvních dochovaných souvislejších záznamů (2. polovina 13. století) zhruba do poloviny 20. století. Do diachronního korpusu jsou zařazovány pouze dobově a útvarově autentické texty, tj. takové, u nichž se dá předpokládat, že do nich nebyly vneseny později žádné prvky pozdějšího jazykového stavu.[30]

V Českém národním korpusu jsou obsaženy dva diachronní korpusy, a to DIAKORP a DOTKO. DIAKORP je dvoumilionový korpus diachronní složky ČNK, zveřejněný roku 2005. DOTKO je pak korpus dolní lužické srbštiny, v němž převažují texty z let 1848–1933. Obsahuje přibližně 2 miliony slov a byl zveřejněn roku 2010.

2.3 Využití Českého národního korpusu

Využití ČNK se samo o sobě moc neliší od obecného využití korpusů *(viz kapitola 1.1.3).* Není možné říct, jaké aplikace lze očekávat od korpusu, výstupů totiž může být neomezeně mnoho. Základní využití je kromě lingvistického především lexikografické, v oblasti jazykového překladu, jazykové pedagogiky a další, např. sociolingvistické nebo psycholingvistické.

Podle Čermáka k hlavním problémům řešeným pomocí korpusů, možnostem a otevřeným otázkám současného dění v oblasti výstavby a studia korpusu patří zaprvé možnost zkoumat frazémy a idiomy. Při srovnání například s internetovými vyhledávači lze na korpusu zvolit vyhledávání pomocí slovního spojení, což zaručí větší přesnost výsledků.

[30] KOCEK, J. – KOPŘIVOVÁ, M. – KUČERA, K. (eds): *Český národní korpus – úvod a příručka uživatele.* ÚČNK FF UK, Praha 2000, s. 14.

Stejným způsobem lze spolehlivě identifikovat specifické typy kolokací, to znamená odhalovat tendence jazyka spojovat některé výrazy (např. lze zjistit, jak často se v češtině vyskytuje substantivum mezi dvěma adjektivy). Dají se také identifikovat větší textové struktury.[31]

Dalším přínosem korpusu je bezpochyby umožnění a zpřístupnění označkovaných víceslovných jednotek. Velká výhoda korpusu tkví právě v označkování, v praxi to znamená rozlišování gramatických kategorií daných slov. Uživatelům vychází vstříc i reverzibilní hledání, což znamená, že své kroky mohou vzít zpět a hledat znovu. Internet také nabízí velmi řídce propojení orální a grafické verze slov, tj. mluvený korpus.

Výhoda Českého národního korpusu spočívá i ve spolehlivosti dat a jejich kritérií. Korpusy jsou reprezentativní, výsledky tedy poté budou vyvážené. Lze na nich také vytvářet subkorpusy, které mohou sloužit jako základ dalších výzkumů.

Díky Českému národnímu korpusu a korpusové lingvistice obecně se dají promýšlet i konsekvence zapojení víceslovných jednotek do celého popisu jazyka. Dá se tedy zjistit, že existují monokolokabilní jednotky (tj. taková slova, která se vyskytují pouze v jednom frazému), např. slovo *eminentní* se v 80% případů vyskytuje před slovem *zájem*.[32] Stejně jako všechny ostatní nereferenční korpusy se ale ČNK stále vyvíjí, proto nelze říci, že by všechny druhy využití výše jmenované byly definitivní.

V této kapitole jsem tedy shrnula informace o nejobsáhlejším českém korpusu, Českém národním korpusu. Jak již bylo výše uvedeno, jeho nedílnou součástí je kromě korpusů češtiny i paralelní korpus InterCorp, kterému se v následující kapitole budu věnovat podrobněji.

[31] ČERMÁK, F. (ed.): *Korpusová lingvistika Praha 2011 – 2 Výzkum a výstavba korpusů*. Nakladatelství Lidové noviny, Praha 2011, s. 24.
[32] CVRČEK, V. – KOVÁŘÍKOVÁ, D.: *Možnosti a meze korpusové lingvistiky.* Naše řeč 94, 2011, s. 127.

3 InterCorp

Korpus InterCorp je hlavním výstupem stejnojmenného projektu, jehož cílem je vybudovat rozsáhlý paralelní synchronní korpus pokrývající co největší počet jazyků. Na jeho tvorbě se významnou měrou podílejí nejen pracovníci Ústavu Českého národního korpusu, ale také pedagogové a studenti Filozofické fakulty Univerzity Karlovy v Praze a další spolupracovníci ÚČNK.[33]

InterCorp se skládá ze dvou částí, a to *jádra* a *kolekce*. Jádrem korpusu InterCorp jsou ručně zarovnané, převážně beletristické texty (z 2. poloviny 20. století). V současnosti je rozsah jádra 91,5 milionů slovních tvarů. Kromě toho obsahuje korpus takzvané kolekce, což jsou automaticky zpracované texty. Převážně se jedná o publicistické články z webových stránek Project Syndicate a Presseurop, dále je to balíček právních textů Acquis Communautaire. V kolekci je možné najít přes 451 milionů slovních tvarů.

3.1 Historie InterCorpu

Projekt InterCorp byl spuštěn roku 2005, kdy získal grant Ministerstva školství, mládeže a tělovýchovy na roky 2005–2011. Cílem bylo vybudovat paralelní synchronní korpusy přibližně 25 jazyků s češtinou ve středu.

Roku 2008, po třech letech příprav, byla databáze, čítající 25 milionů slov, zpřístupněna veřejnosti pomocí webového rozhraní Park.[34] Park byl prvním softwarem, který umožnil vyhledávat výsledky pro více jazyků najednou. Korpus obsahoval 19 jazyků, slova ovšem ještě nebyla ani lemmatizovaná, ani označkovaná.

V dalších letech přibývaly další jazyky (nejvíce mezi lety 2011 a 2012). Největší vzestup počtu slov byl zaznamenán opět mezi lety 2011 a 2012. Nyní je tak k dispozici pátá verze korpusu, obsahující 27 jazyků, z nichž je 17 označkovaných a 14 lemmatizovaných. V jádře korpusu je 91,5 milionů slov, zatímco v kolekci přes 451 milionů.[35]

V současnosti se také plánuje rozšíření korpusu o další jazyky a v prvním čtvrtletí roku 2013 by měla být rozšířena i programová podpora InterCorpu (*více v kapitole 3.2*).

3.2 Popis InterCorpu

Cílem projektu InterCorp je vytvoření paralelních, tj. o překlady opřených korpusů češtiny a všech velkých jazyků evropských i většiny tzv. malých, a to na základě dat Českého národního korpusu a ve spojení s ním.[36] Každý cizojazyčný text má v korpusu InterCorp svou českou verzi, čeština je tedy takzvaný pivot. Česká verze textu (jak originál, tak překlad) je zarovnaná s jednou nebo více verzemi cizojazyčnými. InterCorp se řídí zásadou, že každý

[33] URL: <http://www.korpus.cz/intercorp/?req=page:info> [Cit. 15.12.2012].
[34] Původním záměrem bylo, aby InterCorp v roce 2010 dosáhl minimálně 2,5 milionů slov. Jak je vidět, už v roce 2008 měl desetkrát větší rozsah, než bylo původně zamýšleno.
[35] URL: <http://www.korpus.cz/intercorp/?req=page:releaseNotes> [Cit. 15.12.2012].
Na stránkách projektu InterCorp je zveřejněna detailní historie verzí, včetně popisu vývoje manuálu Park.
[36] URL: <http://www.korpus.cz/intercorp/dokumenty/VZpara.pdf> [Cit. 15.12.2012].

jazyk v něm obsažený má minimálně jednoho garanta z Filozofické fakulty Univerzity Karlovy.

V současné době zahrnuje InterCorp kromě češtiny také jiné *slovanské jazyky* (běloruštinu, bulharštinu, chorvatštinu, makedonštinu, polštinu, ruštinu, slovenštinu, slovinštinu, srbštinu a ukrajinštinu). V korpusu jsou obsaženy i *jazyky germánské* (angličtina, dánština, němčina, nizozemština, norština a švédština). Dále jsou zahrnuty *románské jazyky* (francouzština, italština, katalánština, portugalština, rumunština a španělština). Z dalších jazyků (baltských, ugrofinských apod.) jsou v korpusu obsaženy i arabština, finština, litevština, lotyština a maďarština. Zatím pouze v kolekcích je možné nalézt také řečtinu. Dále se plánuje rozšíření korpusu o romštinu, hindštinu a vietnamštinu. InterCorp tedy předčil očekávání, protože v něm bude možné zkoumat nejen jazyky tradičních evropských větví, ale i jiné indoevropské jazyky, kterými se mluví např. v Asii.

Korpus InterCorp je dostupný z internetu přes webové vyhledávací rozhraní Park, dostupné z adresy http://korpus.cz/Park/login, je nutné ovšem znát uživatelské jméno a heslo. Používá se stejné heslo jak pro Český národní korpus, tak pro InterCorp.

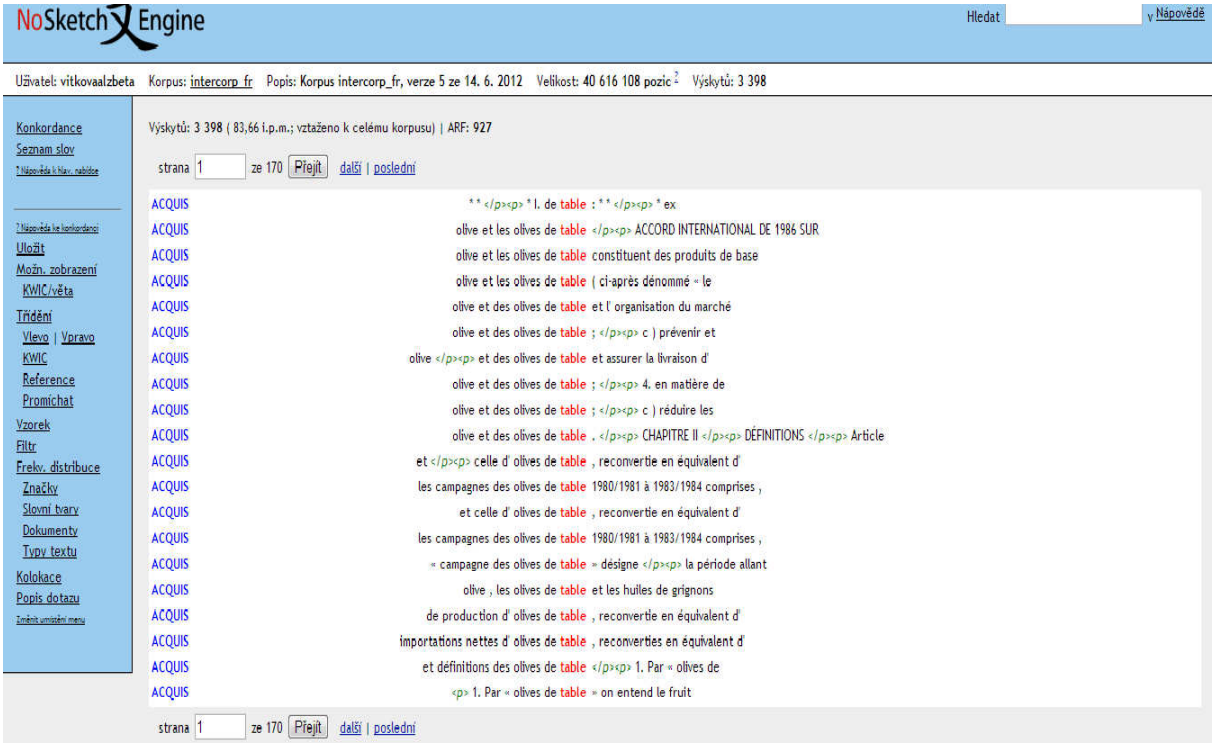
Obr. 2: Ukázka vzhledu rozhraní Park (vyhledávání lemmatu slova *stůl*) ve formátu KWIC

Vysvětlivky čísel: 1 – počet výskytů; 2 – název korpusu; 3 – vyhledaný výraz (KWIC – Key Word in Context); 4 – možnosti exportu do excelové tabulky; 5 – výběr horizontálního nebo vertikálního pohledu; 6 – zobrazení širšího kontextu

Na obrázku můžeme vidět srovnání čtyř subkorpusů (čtyř různých jazyků). Jde o češtinu, angličtinu, francouzštinu a italštinu. Dané výsledky jsou pak zarovnány podle vět.

Na přelomu března a dubna 2013 by mělo být spuštěno nové, modernější webové rozhraní NoSketch Engine, pomocí něhož lze v současné době vyhledávat pouze v jednojazyčných částech paralelního korpusu. Po jeho spuštění bude uživatelům umožněno například vytvářet i subkorpusy, ukládat je a dále s nimi pracovat. Otevře se také možnost využívat frekvenční distribuci apod., tedy pokročilé, přesto velmi potřebné funkce.

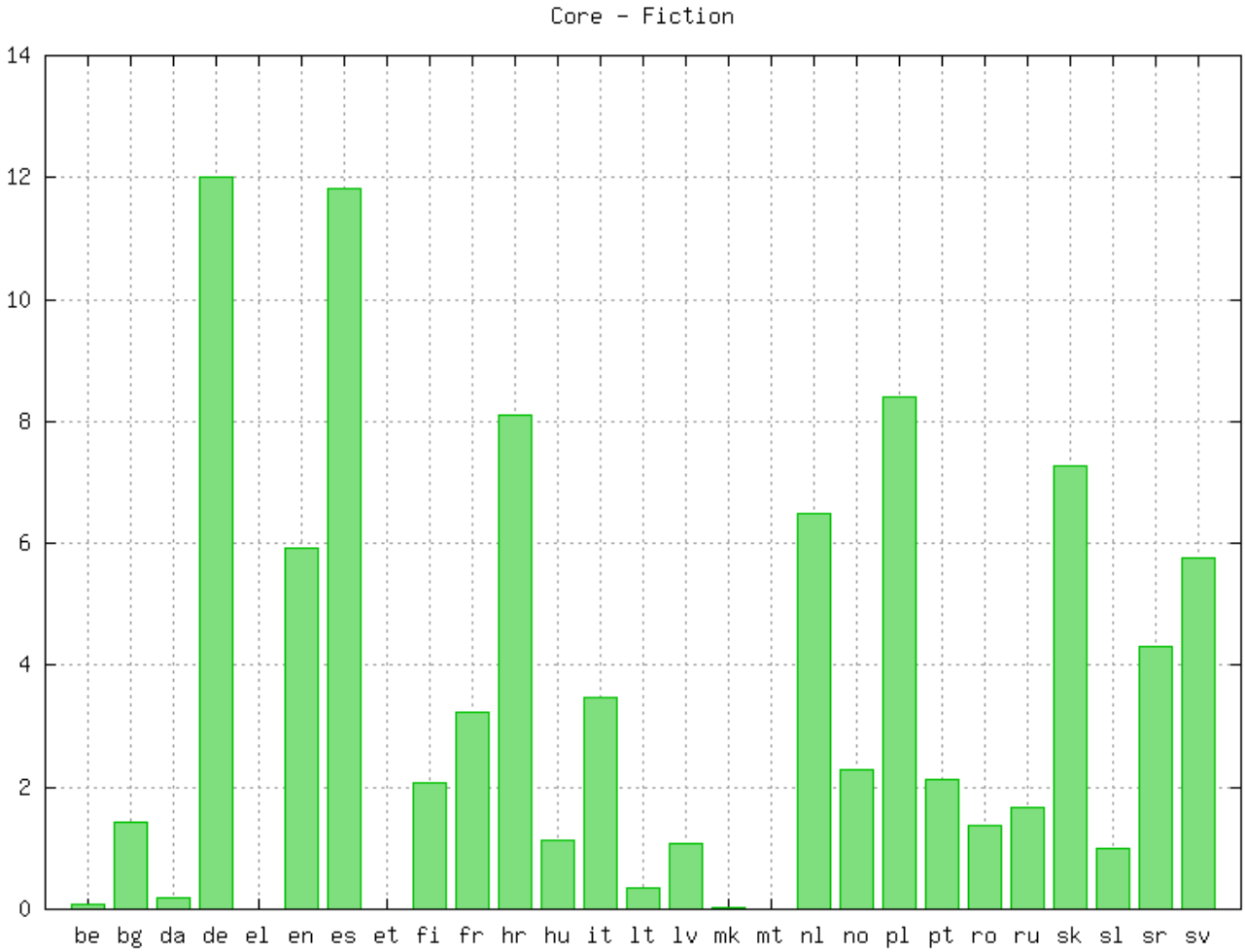
Obr. 3: Ukázka vzhledu rozhraní NoSketch Engine (vyhledávání lemmatu slova *table*) ve formátu KWIC



Na obrázku je velmi pěkně vidět srovnání se starším rozhraním Park. NoSketch Engine je vizuálně přívětivější (je tedy jasné, že bude pro uživatele příjemnější s ním pracovat), mnohem důležitějším pokrokem je ale menu v levé liště, kde lze zvolit naprosto přesná kritéria vyhledávání. Zjednodušení představuje i rychlé zobrazení zdroje (zde ACQUIS). Při kliknutí se totiž zobrazí všechny dostupné informace o daném zdroji, které je v Parku potřeba složitě dohledávat. V horní liště uživatel vidí i podrobnější popis korpusu, např. jak aktuální verzi právě prohledává.

Kromě InterCorpu však v Evropě existují ještě další paralelní korpusy, např. OPUS, který sice zahrnuje více jazyků a textů než InterCorp, na druhou stranu obsahuje jen některé specifické, volně dostupné typy textů. Tím pádem ztrácí na reprezentativnosti. JRC-Acquis, další z evropských paralelních korpusů, zase obsahuje jen zákony Evropské unie. ParaSol, korpus slovanských a jiných jazyků, je mnohem méně obsáhlý (přestože stejně jako InterCorp klade důraz na beletrii).

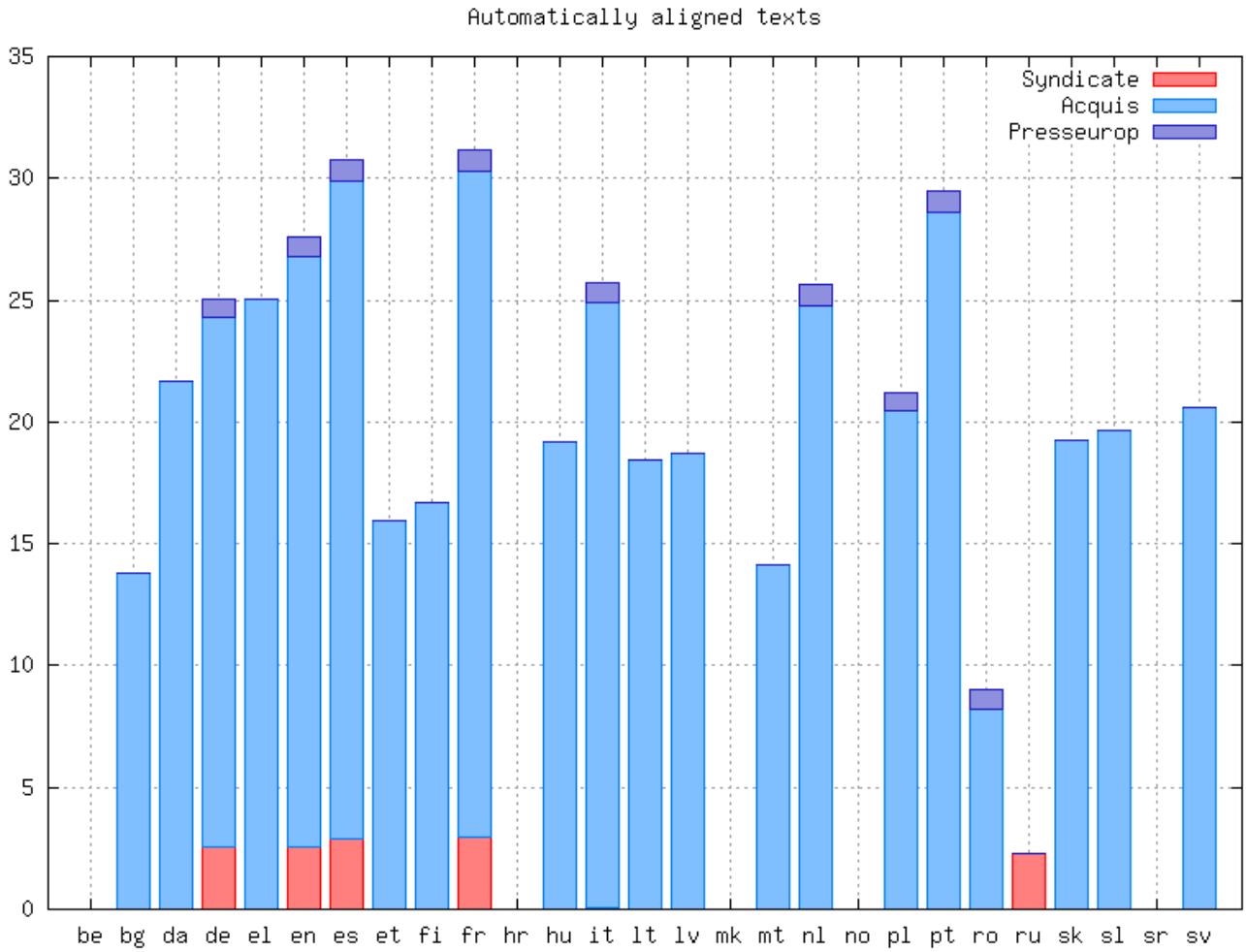
Obr. 4: Složení korpusů jednotlivých jazyků – jádro[37]



Jak vidíme v grafu, v jádře korpusu (převážně beletristické části korpusu) se nachází nejvíce textů v němčině a španělštině. Němčina se španělštinou tvoří dohromady jednu čtvrtinu celého jádra. Naopak například řečtina nebo estonština není v jádře zastoupena vůbec.[38]

[37] URL: <http://www.korpus.cz/intercorp/?req=page:info> [Cit. 27.12.2012].
[38] Tento graf pochází z roku 2011, jsou na něm tedy zaznamenána data z verze 4.

Obr. 5: Složení jednotlivých korpusů – kolekce[39]



Z grafu lze vyčíst, že jazyky jsou v kolekci (publicistické a právní texty) zastoupeny mnohem rovnoměrněji. V kolekci je nejvíce zastoupena francouzština s 31 miliony slov, těsně za ní následuje španělština s více než 30 miliony slov a portugalština s 29 miliony slov. Naopak chorvatština nebo běloruština nejsou zastoupeny vůbec. Jak je vidět, naprostou většinu textů kolekce tvoří balíček právních textů Acquis, v používanějších jazycích Evropské unie je zastoupen i Presseurop. Němčina, angličtina, španělština, francouzština a ruština jsou obsaženy i v Syndicate.[40]

3.3 Využití InterCorpu

InterCorp nachází využití v mnoha oblastech. Předně je důležitý pro lexikografy, překladatele, učitele a studenty cizích jazyků, translatology, literární vědce nebo komparatisty. Slouží i k vyhledávání informací ve více jazycích (cross-language information retrieval). Důležitý je i pro zjednoznačnění textu v jednom jazyce.

[39] URL: <http://www.korpus.cz/intercorp/?req=page:info> [Cit. 27.12.2012].
[40] I tento graf pochází z roku 2011, jsou na něm tedy zaznamenána data z verze 4.

Lexikografové oceňují vyhledávání pomocí paralelních konkordancí, identifikaci kolokací a jejich ekvivalentů nebo třeba extrakci ekvivalentů.

Překladatelé používají překladovou paměť (Translation Memory), překlad podle příkladů (Example-Based Machine Translation) nebo statistický překlad. Díky InterCorpu si svůj překlad mohou i kontrolovat.

Dále se díky InterCorpu dají srovnávat např. lexikální ekvivalence. V Kunderově Nesmrtelnosti a Rowlingové Harry Potter and the Philosopher's Stone je jeden výchozí text v češtině a druhý v angličtině, vychází se tedy ze stejných měřítek, to znamená, že v obou případech jde o originální jazyk (Nesmrtelnost v češtině a Harry Potter v angličtině).[41]

Pomocí InterCorpu se také dá vyhledat, jak lze do češtiny přeložit např. francouzské slovo *fête*. Slovník[42] uvádí pouze tři ekvivalenty: *svátek, jmeniny, slavnost*.

Tab. 1: Ukázka výsledků, vyhledaných v korpusu InterCorp (slovní tvar *fête*)

Radostně ho pozdravily a hned ho zavedly k sobě do pokoje.	Ils lui firent **fête** et le menèrent aussitôt à leur chambre.
Navrátilci z táborů nebyli k **oslavám** přizváni.	Les déportés de retour des camps n' étaient pas à la **fête**.
Vrhli se na sebe, porazili při tom vzpěry věšáků a padli na matraci z šatů , zatímco na druhém konci bytu právě vrcholila **zábava**.	En se jetant l'un sur l'autre, ils avaient renversé les montants et s'étaient vautrés sur le matelas de vêtements alors que de l'autre côté de l'appartement la **fête** battait son plein.
Já taky nenávidím Halloween: předtím jsme měli Dušičky a svátek Všech svatých, nechápu, proč bylo třeba hledat nějaký **svátek** za velkou louží."	Moi aussi je déteste Halloween: on avait la Toussaint avant, je ne vois pas pourquoi il a fallu aller chercher une **fête** outre-Atlantique .
Děkuji za vaši pozornost a ať **slavnost** pokračuje!	Merci de votre attention et que la **fête** continue!
Langouvé byl sice trochu podivín, ale tak potrhlej zase nebyl… nebo mu to nakukali v Cissenu? … jiný dřevorubecký "úderníci"? pořád dokola mu vyhrávali na granátomet, až mu udělali "**mejdan**" v hlavě… a nasadili mu tam nějaký představy apotheos …	Langouvé était un peu braque, mais pas si tellement… ou alors, c' était à Cissen? … les autres « bûcherons de choc » ? ils avaient pas fait que lui sonner le tromblon, ils y avaient mis la « **fête** » dedans … qu' on était en Apothéose …
Jaká byla belgická **hostina**?	C´était bien la petite **fête** belge ?
Přednosta našeho oddělení slavil **jmeniny** a pozval nás do jedné vinárny, pak se šlo do druhé, do třetí, do čtvrté, do páté…	C'était la **fête** de notre chef de bureau et il nous avait donné rendez-vous chez un marchand de vin. De là, on est allé chez un autre bistrot, puis chez un troisième, un quatrième, un cinquième…

InterCorp našel 231 výskytů slova *fête*. Namátkou se dá vybrat například 8 různých překladů, které mohou pomoci jak překladateli začínajícímu, tak zkušenému. Je tedy jasné, že oproti slovníku má InterCorp tu výhodu, že klade daleko větší důraz na variabilitu synonym.[43]

[41] ČERMÁK, F.: *InterCorp: jeho povaha a možnosti*. Ústav Českého národního korpusu, Karlova Univerzita v Praze 2011.
[42] *Francouzsko-český slovník*. FIN Publishing, Olomouc 1998.

InterCorp má však i jisté nevýhody. Zaprvé – texty nejsou autentické, většinou jde o překlady. Je tak nutné spoléhat na důslednou práci překladatele. Texty také nejsou reprezentativní, paralelně lze totiž získat jen některé typy textů. Předpokladem správného fungování korpusu je také spolehlivé párování alespoň po větách – automatické párování je třeba ručně opravovat. Ne vždy totiž sedí překlady přesně věta od věty.[44] Nakonec je také obtížné získat nástroje, které mají požadované funkce, ale nevyžadují speciální znalosti.[45]

Podala jsem tedy základní informace o korpusové lingvistice, Českém národním korpusu a paralelním korpusu InterCorp. Všechny tyto poznatky ovšem byly více méně teoretické, proto se v další části budu zabývat tím, jak lze InterCorp využít v praxi a jak významná je role propagace při práci s ním.

[43] Dalo by se namítnout, že existují i slovníky synonym, tudíž lze zvolit variantu: Nejprve si vyhledat překlad v překladovém slovníku, poté vyhledat synonyma ve slovníku synonym. Tento postup je sice možný, ale zdlouhavý. InterCorp stejné výsledky vyhledá během chvíle.

[44] Například při vyhledávání výsledků pro tabulku výše (slovo *fête*) se několikrát také experimentálně potvrdilo, že ruční zarovnávání nemusí být vždy úplně přesné.

[45] URL: <http://utkl.ff.cuni.cz/~rosen/public/pc_short_2008.pdf> [Cit. 15.12.2012].

4 Praktická část

4.1 Cíle, metodika a náplň praktické části

4.1.1 Cíle a metodika praktické části

V teoretické části jsem shrnula základní poznatky o korpusové lingvistice a korpusu InterCorp. Snažila jsem se o zdůraznění využitelnosti tohoto paralelního korpusu. Vše ale probíhalo pouze v teoretické rovině, proto jsem se rozhodla prakticky vyzkoušet, jakou souvislost má míra propagace s motivací a úspěšností žáků při práci s korpusem.

Ústav Českého národního korpusu, který zaštiťuje celý projekt, uvádí, že paralelní korpus slouží jako zdroj dat pro teoretické studie, lexikografii, studentské práce, výuku, zejména výuku cizích jazyků, počítačové aplikace, překladatele i veřejnost.[46] Poslední slovo, tedy veřejnost, mě zaujalo, a tak jsem si položila otázku: jak může být paralelní korpus InterCorp užitečný právě veřejnosti, přesněji řečeno mým spolužákům na střední škole? Jakou roli při jejich práci s korpusem hraje vnější motivace, tedy jak moc je ovlivňuje člověk, který jim s prací pomáhá? Je vůbec možné, aby si s takto náročným nástrojem poradili sami? Má na studenty nějaký vliv propagace? Pokud ano, jak velký? Odpověď na tyto otázky jsem zkusila zjistit experimentálně.

Mým záměrem bylo srovnat práci dvou skupin. S jednou jsem pracovala já: nejdříve jsem se snažila studenty namotivovat, pak jsme společně po krůčcích vypracovali úkoly. Druhá skupina byla odkázána sama na sebe: každý student pracoval samostatně bez jakékoli ústní pomoci. Měli k dispozici pouze tištěný návod. Na konci časového limitu, vyhrazeného pro vypracovávání úkolů, dostali dotazníky, které monitorovaly jejich spokojenost a další motivaci k práci s korpusem.

Vytvořila jsem pracovní list, který obsahoval pět úkolů (o logice jejich vytváření více v *kapitole 4.1.2*). Jelikož je InterCorp korpus paralelní, zvolila jsem práci s korpusem česko-francouzským. O pomoc jsem požádala svou učitelku francouzštiny, Mgr. Moniku Peroutkovou, protože jsem potřebovala uvolnit pro svou práci dvě vyučovací hodiny ve dvou třídách. Paní Peroutková mi vyšla ve všem vstříc, a tak mohla praktická hodina proběhnout ve 3.A (třetí ročník čtyřletého studia) a v 6.G (šestý ročník osmiletého studia) třebíčského gymnázia, tedy ve dvou třídách, které už prošly více než třemi lety výuky francouzštiny.

[46] URL: <http://www.korpus.cz/intercorp/> [Cit. 10.2.2013].

4.1.2 Pracovní list

ÚKOLY – PRÁCE S KORPUSEM INTERCORP

Kromě řešení každého úkolu popiš ještě stručně svůj postup.

1. LEXIKOLOGIE

Najdi frekvenci užití slova *gros* v komiksech o Asterixovi. Může být jak adjektivum v mužském i ženském rodě (*gros* i *grosse*), tak substantivum (*gros*).

2. *MORFOLOGIE*

Srovnej dvě možnosti překladu slova bosý: *pieds nus, nu-pieds*. Který překlad je používanější? Napiš, kolikrát se obě spojení v celém korpusu vyskytují.

3. HLÁSKOSLOVÍ/FONETIKA

Vyhledej frekvenci užití dvou homonym – *sain* a *saint*. Slova mohou být v plurálu (*sains, saints*), poněvadž i v tom případě jsou homonymní, ne však v ženském rodě (*saine, sainte*). Které je používanější? (tj. když je uslyšíme v omezeném kontextu, která varianta bude pravděpodobnější?)

4. TVOŘENÍ SLOV

Zjisti, kolikrát se v komiksech o Asterixovi vyskytují slova končící na koncovku *-ette*. Dvě omezení: **ne**smí končit na koncovku *-ettes* (tj. plurál), **ne**započítávej ani zájmeno *cette*.

5. SYNTAX

Najdi, jak často se sloveso *faire* používá v kauzativní konstrukci, tj. v češtině nechat/donutit někoho dělat něco. V praxi to znamená, že se pojí s jiným slovesem v infinitivu a předmětem – tj. pojí se s jiným slovesem (např. *faire découvrir*).

Každý úkol v pracovním listu byl tvořen na základě teoretických poznatků o paralelním korpusu InterCorp. Jak již bylo řečeno, úkoly se týkaly francouzského jazyka, který je velmi gramaticky rozmanitý a dá se na něm zkoumat mnoho zajímavých jevů. Úlohy byly řazeny podle obtížnosti vzestupně.

První úkol se týkal lexikologie, tedy nauky o slovní zásobě. Záměrně jsem zvolila práci pouze s korpusem textů od René Goscinnyho o Asterixovi a Obelixovi. Obelix má totiž jednu velmi výraznou vlastnost, a to velkou váhu, stručně řečeno je tlustý. Proto jsem se rozhodla vyzkoumat, kolikrát se v korpusu vyskytuje slovo *gros* neboli *tlustý*. Chtěla jsem po studentech, aby nepočítali jen s maskuliny, ale i femininy, neboť francouzština zná (stejně jako čeština) mluvnickou shodu. Žáci nakonec zjistí, že ačkoli je tloušťka Obelixova charakteristická vlastnost, v korpusu se konkrétně slovo *tlustý* vyskytuje pouze třicetkrát. Goscinny tedy nepoužívá explicitní vyjádření příliš často.

Druhý úkol, morfologický, se zaměřil na translatologický problém, a to překlad slova *bosý*. V doslovném překladu znamenají oba výrazy – jak *pieds nus*, *tak nu-pieds* – doslova *bosé nohy*. Ve francouzštině přívlastek následuje většinou až za větným členem, na nějž se váže, proto sousloví *pieds nus* není nijak stylisticky neobratné. Tendence používat inverzi (*nu-pieds*) je však méně typická. Žáci tedy měli zjistit, že v korpusu se vyskytuje překlad *pieds nus* desetkrát více než *nu-pieds*. Prakticky tak bylo dokázáno, že francouzština opravdu častěji vytvoří nové slovo spojením dvou jiných v sousloví.

Třetí úkol se týkal hláskosloví, přesněji řečeno homonymie. Slova *sain* (*zdravý*) a *saint* (*svatý*) se totiž vyslovují v singuláru a plurálu maskulin stejně – [sɛ̃]. Cílem studentů bylo zjistit, jaká varianta výkladu je pravděpodobnější, když uslyší např. jen spojení slov [ʒak e sɛ̃], tedy *Jacques est sain/t*, v překladu *Jakub je zdravý/svatý*. Studenti nakonec zjistili, že je čtyřikrát pravděpodobnější, že *je Jakub svatý*.

Čtvrtý úkol se věnoval tvoření slov. Francouzština tvoří deminutiva neboli zdrobněliny nejčastěji přidáním adjektiva *petit* (*malý*), druhým oblíbeným způsobem je derivace sufixem *-ette*. V tomto úkolu studenti zkoumali, kolik takto tvořených slov se vyskytuje v komiksech o Asterixovi, to znamená textu určeném primárně dětem a dospívajícím čtenářům. Je ale důležité si uvědomit, že zájmeno *cette* (*tato*) je odvozeno koncovkou *-tte* od zájmena *ce* (*tento*), proto se do zdrobnělin řadit nebude. Aby si žáci vyzkoušeli i práci s jinými příkazy, nesměly být do výsledků započítány plurálové tvary. Při správném postupu tedy studenti přišli na to, že se v textu vyskytuje pouze minimum deminutiv, tvořených pomocí sufixu.

Pátý úkol, týkající se syntaxe, byl zamýšlen jako nejsložitější. Čeština sice kauzativní konstrukce, což je spojení slovesa nechat/dát s infinitivem, nepoužívá příliš často, zato ve francouzštině (stejně jako např. v angličtině) se vyskytují mnohokrát. To dokazují výsledky, protože v celém korpusu bylo nalezeno přes tři a půl tisíce výskytů této konstrukce.

4.1.3 Dotazník a jeho role

Prvotním cílem dotazníku bylo zjistit, nakolik ovlivnila propagace paralelního korpusu InterCorp postoj studentů k tomuto projektu. Konkrétně jsem se snažila o analýzu různých aspektů, které na středoškoláky působí. Zaprvé mě zajímalo, jak se žákům práce líbila. Druhou otázkou bylo zhodnocení obtížnosti úkolů, třetí zase spokojenosti studentů s jejich výsledky. Čtvrtá, pátá a šestá otázka se zabývaly motivací žáků. Do této otázky byly dotazníky pro obě skupiny stejné. Poslední otázkou se ale lišily. Sedmá otázka u skupiny, s níž jsem pracovala, byla jakousi zpětnou vazbou, nakolik studentům přišla vhod má přítomnost a pomoc. Druhá skupina měla v této otázce rozhodnout, zdali by pomoc nějakého zkušenějšího uživatele InterCorpu ocenili.

Studentům jsem nabídla možnost připsat jakoukoli poznámku či připomínku do dotazníku. Toho využil pouze jediný účastník ze skupiny Bez pomoci, který mi připsal dvě poznámky: k otázce č. 3 (spokojenost s výsledky) a č. 7 (případná pomoc).

Otázka č. 3: *Jsi spokojen se svými výsledky?*

3 – nejsem ani spokojená, ani nespokojená: „Přijde mi, že jsem se to snažila řešit co nejprimitivněji, takže jsem spíš nespokojená, že jsem to neřešila složitěji a odborněji.“

Otázka č. 7*: Myslíš, že by Tě práce s InterCorpem bavila víc, kdybys nemusel/a hledat cestu sám/a a někdo Ti vysvětlil, jak na to?*

ANO – „Kdybych tu byla sama, byla bych velmi nešťastná a vystresovaná z časového limitu a složitosti. Nejprve jsem ani nevěděla, jak to otevřít, co to je a co se vlastně po mně chce. Kdybych tu byla sama, spíš by se mi to nelíbilo. Ale teď se mi to docela zamlouvá. :)“[47]

Dle mého názoru poznámky této studentky vystihovaly nejen pocity její, ale i ostatních žáků. To, že možnost připsat dojmy vlastními slovy využije jen minimum účastníků, jsem předpokládala už při tvorbě dotazníku. Snažila jsem se tedy o to, aby byly otázky formulovány přesně a aby obsáhly všechny potenciální nápady a připomínky studentů, aniž by je museli vytvářet sami.

[47] Studentka jako odpověď na otázku č. 1 uvedla, že práce s InterCorpem se jí 3 – trochu líbila.
Formulací *Kdybych tu byla sama, …* chtěla naznačit, že ačkoli pracovala sama, psychicky jí pomohla přítomnost ostatních účastníků praktické hodiny.

DOTAZNÍK – PRÁCE S KORPUSEM INTERCORP

U každé otázky podtrhni jeden výrok, který Tě nejvíce vystihuje.

1. Líbila se Ti práce s InterCorpem?
   1 – velmi se mi líbila
   2 – docela se mi líbila
   3 – trochu se mi líbila
   4 – moc se mi nelíbila
   5 – vůbec se mi nelíbila

2. Jak obtížné se Ti zdálo používání InterCorpu?
   1 – velmi snadné
   2 – docela snadné
   3 – tak akorát
   4 – poměrně obtížné
   5 – velmi obtížné

3. Jsi spokojen se svými výsledky?
   1 – velmi spokojen/á
   2 – docela spokojen/á
   3 – nejsem ani spokojen/á, ani nespokojen/á
   4 – poměrně nespokojen/á
   5 – naprosto nespokojen/á

4. Máš chuť InterCorp užívat i nadále?
   1 – ano, velkou
   2 – spíše ano
   3 – trochu ano
   4 – spíše ne
   5 – absolutně ne, vůbec žádnou

5. Odrazuje Tě něco od dalšího užívání InterCorpu? *ANO – NE*

6. Pokud ano, co to je? (možno více odpovědí)
   a) nelíbí se mi design (málo barev, stále stejné písmo, grafická nepřehlednost)
   b) nerozumím tomu, jak by mi mohl být nadále užitečný
   c) jeho používání rozumím, ale je komplikované (regulární výrazy apod.)
   d) nerozumím tomu, jak jej používat (nejasný návod apod.)
   e) projekt InterCorp mě vůbec nezaujal
   f) jiné (uveď): _____

7. a) Myslíš, že by Tě práce s InterCorpem bavila stejně, i kdyby Ti s ní nikdo nepomáhal? *ANO – NE*
   b) Myslíš, že by Tě práce s InterCorpem bavila víc, kdybys nemusel/a hledat cestu sám/a a někdo Ti vysvětlil, jak na to? *ANO – NE*

4.2 Průběh praktické hodiny

4.2.1 Praktická hodina ve třídě 3.A Gymnázia Třebíč (skupina S pomocí)[48]

Praktická hodina ve třídě 3.A probíhala v pondělí 4. února 2013 v internetové studovně Městské knihovny v Třebíči. K dispozici bylo 12 počítačů, na nichž pracovalo 12 studentů, a jeden notebook napojený na dataprojektor, na němž jsem já ukazovala postup.

Samotná práce na úkolech byla naplánována na 70 minut, vyplňování dotazníku na 10 minut. Studentům jsem nejprve během 5 minut krátce prezentovala projekt paralelního korpusu InterCorp (stručně jsem představila podstatu, historii a zázemí InterCorpu a objasnila jsem jeho využití). Poté jsem jim rozdala dvoustránkový návod, který jsem na základě různých zdrojů a vlastních znalostí vytvořila speciálně pro tuto skupinu *(viz kapitola 7.2).* Na dalším papíře dostali účastníci experimentu morfologické značky pro francouzskou sekci InterCorpu, které jim pomohly zejména v pátém úkolu.

Poté, co se všichni přihlásili do InterCorpu pomocí speciálně vytvořeného účtu Jiřina Zelená[49], jsem žákům ve zkratce vysvětlila, jak InterCorp funguje – např. kam mají zadávat dotaz, k čemu je dobré lemma, tag apod., jak filtrovat texty a jak vybírat jinojazyčné sekce. Aby ovšem studenti věděli, k čemu jim je InterCorp dobrý i prakticky, vysvětlila jsem jim motivaci prvního úkolu *(viz kapitola 4.1.2).* Po překonání počátečního ostychu se studenti pustili do řešení a navrhovali možné postupy. První úkol tak byl vyřešen za 9 minut.

Po nastínění logiky druhého úkolu *(viz kapitola 4.1.2)* se jej žáci pokusili vyřešit samostatně. Když několikrát neuspěli, zkonzultovali své návrhy se mnou a společně jsme došli k řešení za 6 minut.

Další využití InterCorpu jsem žákům ukázala na třetím úkolu *(viz kapitola 4.1.2).* Bohužel po patnácti minutách užívání přestalo webové rozhraní Park fungovat, proto se u dvou počítačů vytvořily dvě pracovní skupiny, které se snažily přijít na řešení tohoto úkolu. Během 14 minut se jedné ze skupin podařilo uspět v zadávání dotazu, a tak studenti postup prodiskutovali hromadně.

Následovala desetiminutová přestávka, během níž jsem odpovídala na dotazy ohledně InterCorpu. Studenty velmi zajímalo, jak jsem se k němu dostala já osobně jakožto studentka střední školy nebo také kdo ho používá. V této práci svou osobní motivaci uvádím v závěru *(viz kapitola 5).*

Nad čtvrtým úkolem *(viz kapitola 4.1.2)*, u něhož dokonce účastníci dokázali sami odhadnout, jaký význam pro ně má, strávili studenti 18 minut. Největší problém pochopitelně dělalo užití konkordance, o níž jsem se sice předtím zmínila, ale nezdůrazňovala jsem ji, aby studenti zkusili pochopit i složitější části návodu.

[48] Jak již bylo výše uvedeno, skupinou S pomocí označuji skupinu, jíž jsem s vypracováváním úkolů pomáhala já.

[49] Pro účet Jiřina Zelená platilo přihlašovací jméno jirinazelena, heslo: intercorp. Tyto údaje zůstaly všem studentům k dispozici i po skončení projektu. Dva z nich dokonce z vlastní iniciativy ještě týž den studovali informace ze stránek ÚČNK a zkoušeli vyhledávat další výrazy.

Jak jsem předpokládala, nejsložitějším úkolem byl poslední, syntaktický *(viz kapitola 4.1.2)*. Motivaci chápali poměrně dobře (z angličtiny si pamatovali vazbu *make sb do sth*), dokázali také odhadnout, že mají použít CQL, ovšem větší problém dělalo zadávání samotného dotazu. Vypracovávání úkolu zabralo 20 minut, nakonec se ale studenti s mou pomocí dobrali správného výsledku.

Po vyplnění dotazníků jsme se studenty ještě ústně prodiskutovali jejich spokojenost a postřehy či připomínky k průběhu hodiny. Všichni se shodli na tom, že kladně hodnotí spolupráci s dlouhodobějším uživatelem InterCorpu, který jim mohl poskytnout nejen rady z oficiálních materiálů, ale i osobní zkušenosti. Většina také oceňovala, že se dozvěděli o novém zdroji informací, práce je vcelku bavila, bohužel však nespatřovali přínos projektu pro ně jako studenty středních škol, kteří se nechtějí zabývat lingvistikou.

4.2.2 Praktická hodina ve třídě 6.G Gymnázia Třebíč (skupina Bez pomoci)[50]

Praktická hodina ve třídě 6.G probíhala v úterý 5. února 2013 v počítačové učebně Gymnázia Třebíč. K dispozici bylo 16 počítačů, které využilo 7 studentů (z původního počtu 10 sextánů 3 odmítli účastnit se projektu). Důležité je poznamenat, že většina z účastníků se zabývá programováním a baví je matematika, daly se tedy předpokládat zajímavé a inovativní postupy při řešení úkolů.

Studentům jsem e-mailem rozeslala čtyřstránkový návod *(viz kapitola 7.1)*, sestavený z informací dostupných z webových stránek ÚČNK, s nímž měli dále pracovat. Dále dostali k dispozici morfologické značky pro francouzskou sekci InterCorpu. Původně jsem je chtěla nechat pracovat pouze s těmito materiály, ovšem po pěti minutách, během nichž se nikomu nepodařilo přihlásit se do rozhraní, jsem se rozhodla jim ukázat alespoň to, jak se dostat do korpusu. Po přihlášení se jim tedy začal odpočítávat limit 70 minut.

Po 30 minutách jsem žákům rozdala mnou vytvořený laičtější dvoustránkový návod *(viz kapitola 7.2)*. Do této doby měli pouze 2 ze 7 účastníků vyřešený první úkol. Během dalších 5 minut vyřešili první úkol i všichni ostatní.

Po 45 minutách jsem se studentů zeptala na ústní zhodnocení, jak se jim pracuje s novým, stručnějším návodem. Odezva byla velmi kladná a o větší účinnosti svědčí i porovnání množství vyřešených úkolů s pomocí návodu původního a nového.

Po 60 minutách měl jeden z účastníků vyřešené čtyři úlohy, čtyři další účastníci čtvrtou právě začínali. Dva zbývající měli hotové dva úkoly.

Po 70 minutách, na konci časového limitu, dokončil jeden student všech pět úkolů, jeden se pokusil o neúplné vyřešení pátého úkolu. Tři studenti zvládli čtyři úkoly, zbylí dva vyřešili tři úkoly.

Po celou dobu jsem procházela mezi žáky a zkoumala, jaké postupy volí při vypracovávání úloh. Ve většině případů jsem si všimla výrazné tendence nadužívat CQL

[50] Jak již bylo výše uvedeno, skupinou Bez pomoci označuji skupinu, která si s úkoly musela poradit sama.

a morfologické značky. To naznačuje, že nad úkoly studenti velmi přemýšleli, někdy ale až moc komplikovaně. Na druhou stranu si více než polovina účastníků chtěla usnadnit práci tím, že se nesnažili využívat všechny funkce, ale naučili se pracovat s jednou a vše ostatní dělali ručně. Tato tendence se ukázala například při řešení úkolu č. 3 – nejdříve spočítali výskyty slova *sain*, pak slova *sains* a oba výsledky sečetli. Několik studentů také příliš často užívalo hranaté závorky tam, kde jich nebylo třeba.

4.3 Vyhodnocení praktické hodiny

Vyhodnocení jsem rozdělila na několik částí. V první části jsem vyhodnotila úspěšnost studentů při vypracovávání úkolů z pracovního listu. Druhá část je věnovaná podrobnému rozboru dotazníků.

4.3.1 Vyhodnocení úkolů v pracovním listu

V pracovním listu dostali studenti 5 úkolů. Všechny, bez ohledu na obtížnost, jsem ohodnotila 5 body. Dohromady tak mohli studenti dosáhnout až 25 bodů. U každého úkolu (kromě pátého, u nějž je explicitně uveden bodovací klíč) jsem se řídila stejnými hodnotícími kritérii – 2 body za správný výsledek, za správný postup 3 body. Pokud byl postup neúplný nebo ne zcela funkční, strhávala jsem 1, popř. 2 body. Bylo tedy důležitější zvolit správný postup než získat správný výsledek.

Skupina, jíž jsem s prací pomáhala, všechny úkoly s dopomocí vyřešila mnou navrhovaným postupem. Jelikož vypracovávání úloh probíhalo formou diskuse, některé studenty napadaly i jiné, více či méně legitimní postupy. Nakonec jsme se ovšem shodli na tom, že raději zkusí přijít na očekávaný postup, který velmi často označovali za nejjednodušší a nejefektivnější. Každý student tedy získal 25 bodů.

Skupina, v níž studenti museli úlohy řešit samostatně, byla, co se týče vymýšlení postupů, daleko nápaditější. Žáci pečlivě prozkoumali materiály, které jsem jim poskytla, a téměř u každého úkolu si našli vlastní cestu. U této skupiny jsem se rozhodla netrvat na mnou navrhovaných postupech a výsledcích. Pokud byl výsledek řádně odůvodněn jiným možným postupem, dostali studenti plný počet bodů. Všechny výsledky, zvláště pak ty, ke kterým se došlo jiným než původně zamýšleným příkazem, jsem zkontrolovala pomocí InterCorpu a snažila se přijít na to, kde daný student udělal chybu, popř. kde jsou rezervy v návodech a omezení daného dotazu. Pod tabulkou každého úkolu je v případě rozdílnosti odpovědí vysvětleno, proč byl nebo nebyl udělen plný počet bodů. Po rozebrání úkolů jsem zhodnotila celkovou úspěšnost této skupiny.

V průběhu hodiny došlo k několika výpadkům rozhraní Park, proto jsem povolila žákům seskupit se do dvojic. Možnosti řešitelé č. 2 a 3; č. 6 a 7. Uvědomila jsem si i riziko neobjektivity výsledků (někteří účastníci experimentu pracovali sami, někteří své nápady sdíleli s někým jiným), po chvíli ale bylo jasné, že výsledky budou rovnocenné. Studenti, kteří pracovali ve dvojicích, nakonec nedosáhli vyššího bodového zisku. Jediné možné ovlivnění experimentu spočívalo v tom, že tito žáci se podporovali ve vypracovávání úkolu a navzájem si pomáhali, mohli tedy mít větší motivaci.

Skupina Bez pomoci

Úkol č. 1

Správná odpověď: 30

Očekávaný postup: Ruční výběr textů – René Goscinny; Lemma – *gros*

Průměrný počet bodů: 4,7

Tab. 2: Zhodnocení úkolu č. 1

řešitel	řešení	Postup	body
Č. 1	30	Ruční výběr textů – René Goscinny; Slovní tvar – gros.* ; ručně sečteno	**5/5**
Č. 2	30	Ruční výběr textů – René Goscinny; Lemma – gros	**5/5**
Č. 3	30	Ruční výběr textů – René Goscinny; Lemma – gros	**5/5**
Č. 4	31	Ruční výběr textů – René Goscinny; Lemma – gros.{0,3}	**3/5**
Č. 5	30	Ruční výběr textů – René Goscinny; Slovní tvar – gros; Slovní tvar – grosse; Slovní tvar – grosses; ručně sečteno	**5/5**
Č. 6	30	Ruční výběr textů – René Goscinny; Lemma – gros; Lemma – grosse; ručně sečteno	**5/5**
Č. 7	30	Ruční výběr textů – René Goscinny; Lemma – gros; Lemma – grosse; ručně sečteno	**5/5**

Poznámka k č. 1: Student uvedl legitimní postup, přestože InterCorp umožňuje ještě jednodušší příkaz.

Poznámka k č. 4: Student uvedl více méně legitimní postup, ovšem správná odpověď mu nevyšla, a to z toho důvodu, že InterCorp nalezl i tvar sloves (*Idéfix a grossi = Idefix ztloustnul*). Toto sloveso sice vyhovovalo příkazu, ale ne zadání úkolu.

Poznámka k č. 5: Student zvolil nejjednodušší, i když nejméně efektivní postup, a to ruční sečtení obou tvarů. Student se vyvaroval i chyby s množným číslem ženského rodu.

Poznámka k č. 6 a 7: Studenti zvolili téměř stejný postup jako student č. 5. Výsledky by byly správné, i kdyby se v korpusu vyskytovala feminina v plurálu, tudíž tento postup je zcela správný.

Výsledky tohoto úkolu prokázaly, že je studentům nejdříve potřeba vysvětlit význam základních nástrojů, jako je automatické sčítání výskytů nebo správné používání příkazů lemma, slovní tvar apod. Ruční sčítání totiž může způsobovat chyby, které ovlivní vyznění celého úkolu.

Úkol č. 2

Správná odpověď: 51x, 5x

Očekávaný postup: Ruční výběr textů – Vybrat vše; Slovní spojení – *pieds nus*; Slovní tvar – *nu-pieds*

Průměrný počet bodů: 4,7

Tab. 3: Zhodnocení úkolu č. 2

Řešitel	řešení	postup	body
Č. 1	49x, 5x	Ruční výběr textů – Vybrat vše; Slovní spojení – pieds nus; Slovní tvar – nu-pieds; ručně sečteno	**3/5**
Č. 2	51x, 5x	Ruční výběr textů – Vybrat vše; Slovní spojení – pieds nus; Slovní tvar – nu-pieds	**5/5**
Č. 3	51x, 5x	Ruční výběr textů – Vybrat vše; Slovní spojení – pieds nus; Slovní tvar – nu-pieds	**5/5**
Č. 4	51x, 5x	Ruční výběr textů – Vybrat vše; Slovní spojení – pieds nus; Slovní tvar – nu-pieds	**5/5**
Č. 5	51x, 5x	Ruční výběr textů – Vybrat vše; Slovní spojení – pieds nus; Slovní tvar – nu-pieds	**5/5**
Č. 6	51x, 5x	Ruční výběr textů – Vybrat vše; Slovní spojení – pieds nus; Slovní tvar – nu-pieds	**5/5**
Č. 7	51x, 5x	Ruční výběr textů – Vybrat vše; Slovní spojení – pieds nus; Slovní tvar – nu-pieds	**5/5**

Poznámka k č. 1: Postup studenta byl správný, ovšem nebylo nutné vše ručně sčítat. Odchylka od řešení tedy pravděpodobně vznikla lidskou chybou při ručním sčítání.

Soudě podle výsledků byl tento úkol pro studenty nejjednodušší. Všichni zvolili očekávaný postup, což svědčí o tom, že dobře porozuměli návodu. Domnívám se, že chyba studenta č. 1 vyplynula z jeho rychlosti vypracovávání úkolů, protože právě tento student si poradil s prvními dvěma úlohami ještě předtím, než jsem všem rozdala zjednodušený návod. Je tedy možné, že už si svůj postup nekontroloval podle laičtějšího návodu, a proto zjistil až u úkolu č. 3, jak se dá jednoduše zjistit počet výskytů daného výrazu (do té doby vše sčítal ručně).

Úkol č. 3

Správná odpověď: 52x, 202x

Očekávaný postup: Slovní tvar – *sains?* ; Slovní tvar – *saints?*

Průměrný počet bodů: 3,7

Tab. 4: Zhodnocení úkolu č. 3

řešitel	řešení	postup	body
Č. 1	52x, 206x	Slovní tvar – sain; Slovní tvar – sains (poté součet); Slovní tvar – saint; Slovní tvar – saints (poté součet)	**4/5**
Č. 2	45x, 97x	Slovní tvar – sain; Slovní tvar – sains (poté součet); Slovní tvar – saint; Slovní tvar – saints (poté součet)	**2/5**
Č. 3	45x, 97x	Slovní tvar – sain; Slovní tvar – sains (poté součet); Slovní tvar – saint; Slovní tvar – saints (poté součet)	**2/5**
Č. 4	52x, 202x	Slovní tvar – sains? ; Slovní tvar – saints?	**5/5**
Č. 5	52x, 202x	Slovní tvar – sains? ; Slovní tvar – saints?	**5/5**
Č. 6	52x, 206x	Slovní tvar – sain; Slovní tvar – sains (poté součet); Slovní tvar – saint; Slovní tvar – saints (poté součet)	**4/5**
Č. 7	52x, 206x	Slovní tvar – sain; Slovní tvar – sains (poté součet); Slovní tvar – saint; Slovní tvar – saints (poté součet)	**4/5**

Poznámka k č. 1, 6 a 7: Studenti uvedli legitimní, i když velmi zdlouhavý a neefektivní způsob řešení úkolu. Nevyvarovali se ale chyby při sčítání. Na druhou stranu je však zvláštní, že přestože spolupracovali pouze studenti č. 6 a 7 (č. 1 pracoval samostatně), sečetli výsledky se stejnou chybou.

Poznámka k č. 2 a 3: Studenti uvedli stejný postup jako student č. 1. Opět ale udělali chybu při ručním sčítání, což by se jim nestalo, kdyby použili regulární výraz.

Tento úkol už žákům dělal větší problémy, což dokazuje i průměrný počet dosažených bodů. Pouze dva účastníci přišli na nejjednodušší postup, který už ovšem vyžadoval pochopení regulárních výrazů. Zbytek studentů se snažil o ulehčení práce, v důsledku se jim však zdánlivě nekomplikovaný postup nevyplatil, protože pomocí těchto příkazů nenašel InterCorp správný počet výskytů.

Úkol č. 4

Správná odpověď: 10

Očekávaný postup: Ruční výběr textů – René Goscinny; Slovní tvar – .*ette; Konkordance – Negativní filtr: Slovní tvar – *cette*

Průměrný počet bodů: 3

Tab. 5: Zhodnocení úkolu č. 4

řešitel	řešení	postup	body
Č. 1	15	Ruční výběr textů – René Goscinny; Lemma – .*ette; ručně sečteno (bez cette a -ettes)	**4/5**
Č. 2	10	Ruční výběr textů – René Goscinny; Slovní tvar – .*ette; Konkordance – Negativní filtr: Slovní tvar – cette	**5/5**
Č. 3	10	Ruční výběr textů – René Goscinny; Slovní tvar – .*ette; Konkordance – Negativní filtr: Slovní tvar – cette	**5/5**
Č. 4	16	Ruční výběr textů – René Goscinny; Slovní tvar – .{2,}ette	**3/5**
Č. 5	13	Ruční výběr textů – René Goscinny; Slovní tvar – [^c].+ette	**4/5**
Č. 6	---------		**0/5**
Č. 7	---------		**0/5**

Poznámka k č. 1: Student použil komplikovaný a ne zcela správný postup, přesto mu InterCorp poskytl výsledky podobné správné odpovědi. Zde se například ukázala i role jemných nuancí při zadávání dotazů (tomuto studentovi nevyhledal InterCorp žádná slovesa, ačkoli v zadání úkolu nebylo řečeno, že je mají studenti vyloučit z vyhledávání).

Poznámka k č. 4: Student použil jiný postup, než byl původně zamýšlen, ovšem ne úplně scestný. Chybou je ovšem to, že tímto příkazem vyloučil z výsledků slova, která mají před koncovkou -ette jen jedno písmeno, i když nejde o *cette* (např. *jette*). (Přesný výklad příkazu: jakýkoli libovolný znak, opakovaný dvakrát nebo více, následovaný hláskovou skupinou -*ette*.)

Poznámka k č. 5: Student použil zajímavý postup, opět úplně odlišný od předchozích. InterCorp našel více výskytů, než bylo zamýšleno, proto byl strhnut jeden bod za nepřesný výsledek. (Přesný výklad příkazu: jakýkoli libovolný znak – s výjimkou „c“, opakovaný jednou nebo více, následovaný hláskovou skupinou -*ette*.)

V tomto úkolu se projevila velká invence většiny studentů. Nebáli se používat složitější regulární výrazy, proto dostali poměrně vysoké bodové ohodnocení, i když nevyhledali správný počet výskytů. Velmi mě potěšilo, že dva studenti odhalili funkci konkordancí, pomocí níž mohli korigovat své výsledky a správně vyřešit úlohu.

Úkol č. 5

Správná odpověď: 3616

Očekávaný postup: Ruční výběr textů – Vybrat vše; CQL – [lemma="faire"] [tag="VER:infi"]

Průměrný počet bodů: 0,9

Bodování úkolu: V tomto případě jsem bodovala velmi benevolentně – za správný postup plný počet bodů, nehledě na výsledek. Bod jsem udělila za snahu najít řešení, které bylo sice příliš jednoduché, ale bod si zasloužilo.

Tab. 6: Zhodnocení úkolu č. 5

řešitel	řešení	Postup	body
Č. 1	2557	Ruční výběr textů – Vybrat vše; CQL – [lemma="faire"] [tag="VER:infi"]	**5/5**
Č. 2	----------		**0/5**
Č. 3	----------		**0/5**
Č. 4	174	Slovní spojení – faire .*	**1/5**
Č. 5	----------		**0/5**
Č. 6	----------		**0/5**
Č. 7	----------		**0/5**

Poznámka k č. 1: Student zvolil očekávaný a pravděpodobně jediný možný postup. Nepřesnost výsledku přisuzuji časovému limitu a technickým potížím (student si pravděpodobně nevšiml upozornění InterCorpu, že *stále počítá*).

Poznámka k č. 4: Student zvolil ne zcela správný postup. První chybou bylo to, že podle příkazu se sloveso *faire* mohlo vyskytovat pouze v infinitivu, což není předmětem úkolu. Druhá chyba tkvěla v tom, že za jakýkoli znak se mohla dosazovat i interpunkční znaménka či jakákoli jiná slova, tedy nejen slovesa v infinitivu, jak bylo zadáno.

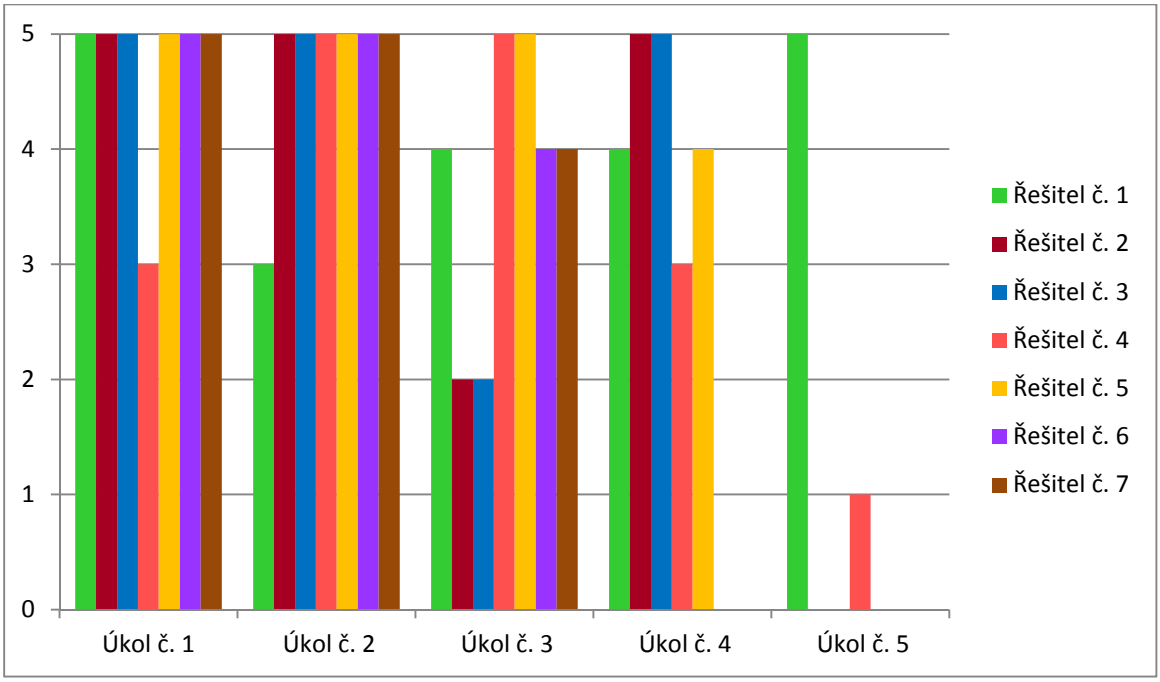
Při zadávání tohoto úkolu jsem předpokládala, že studenti nebudou mít příliš nápadů, jak jej řešit, protože je velmi složitý. Student č. 1, který jako jediný tuto úlohu zdárně vyřešil, měl na její vypracovávání nejvíce času, jelikož si s ostatními úkoly poradil nejrychleji. Ukázalo se tedy, že čas hraje velkou roli. Je možné, že během dalších 15 minut by se podařilo vyřešit úkol více studentům. Obě skupiny (S pomocí i Bez pomoci) však musely mít stejné podmínky, proto jsem časový limit neprodlužovala.

Tab. 7: Zhodnocení řešení všech úkolů skupiny Bez pomoci

řešitel	Úkol č. 1	Úkol č. 2	Úkol č. 3	Úkol č. 4	Úkol č. 5	celkem	
Č. 1	5	3	4	4	5	**21**	
Č. 2	5	5	2	5	0	**17**	
Č. 3	5	5	2	5	0	**17**	
Č. 4	3	5	5	3	1	**17**	
Č. 5	5	5	5	4	0	**19**	
Č. 6	5	5	4	0	0	**14**	
Č. 7	5	5	4	0	0	**14**	
průměrně	*4,7*	*4,7*	*3,7*	*3*	*0,9*	**3,4**	**17**

Průměrně tedy studenti skupiny Bez pomoci dosáhli 17 bodů, což je 3,4 bodu na úkol. Tento výsledek mě velmi potěšil, protože se ukázalo, že i nezkušený uživatel může využívat paralelní korpus InterCorp jen po krátkém zasvěcení.

Graf 1: Zhodnocení řešení všech úkolů skupiny ***Bez pomoci***



Z tabulky a grafu vyplývá, že nejlépe studenti vyřešili úkoly č. 1 a 2. Toto zjištění je vcelku logické, protože obtížnost úkolů byla vzrůstající. Žáci dosáhli lepších výsledků, než jsem očekávala, což přičítám jejich zálibě v matematice a programování. Průměrně dosáhli v každém úkolu 3,4 bodu z 5 – s výjimkou poslední úlohy vyřešila většina účastníků všechny úkoly.

Nejvíce bodů (21 z 25) získal řešitel č. 1, naopak nejméně bodů (14 z 25) získali řešitelé č. 6 a 7. Jak už bylo výše uvedeno, č. 2 a 3 spolupracovali, stejně jako č. 6 a 7. Obě tyto dvojice dostali nejhorší bodové ohodnocení (14 a 17 bodů), z toho tedy vyplývá, že výsledky tato spolupráce nezkreslila.

4.3.2 Vyhodnocení dotazníků

Výsledky jsem zpracovala nejprve do dvou tabulek hodnot z dotazníků. Jedna se týkala skupiny, jíž jsem s prací pomáhala (skupina S pomocí, popř. první skupina). Druhá zachycovala názory skupiny, která byla odkázána sama na sebe (skupina Bez pomoci, popř. druhá skupina).

Druhou část tvoří porovnání grafů, vyplývajících z údajů z dotazníků obou skupin. Každou otázku z dotazníku jsem porovnávala zvlášť, okomentovala jsem oba grafy a nakonec srovnala výsledky.

Tab. 8: Zhodnocení odpovědí z dotazníku skupiny ***S pomocí*** (třída 3.A)

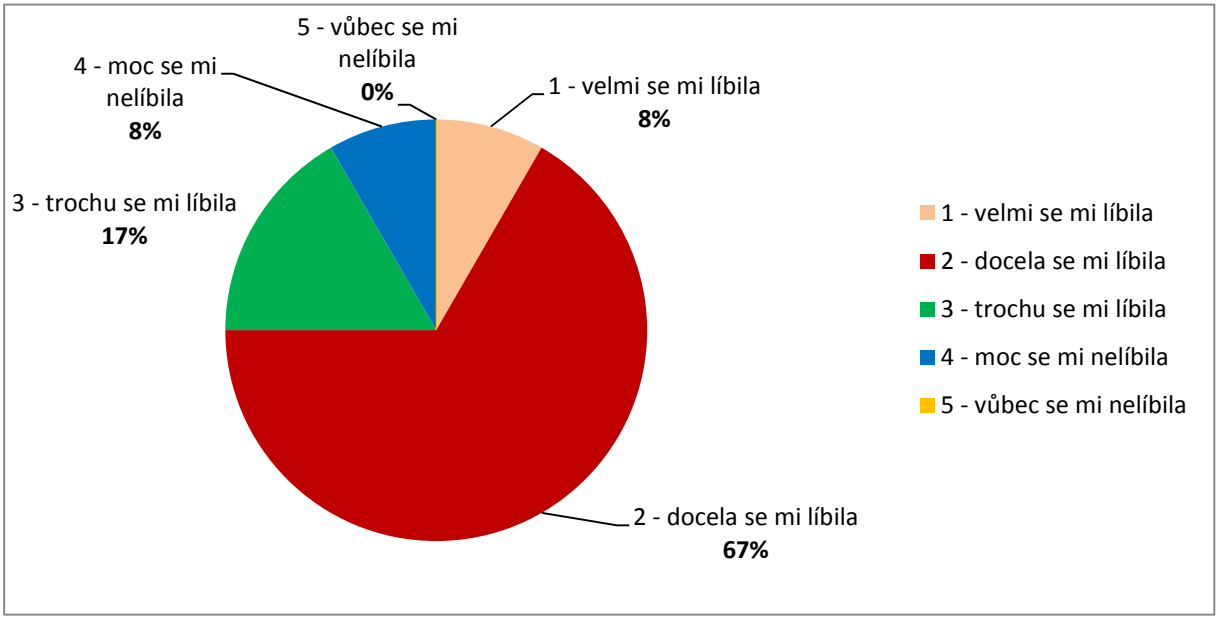
Odpověď	Otázka 1	Otázka 2	Otázka 3	Otázka 4	Otázka 5	Otázka 6	Otázka 7
1	1	0	3	1	----------	----------	----------
2	8	0	5	3	----------	----------	----------
3	2	3	3	5	----------	----------	----------
4	1	8	1	3	----------	----------	----------
5	0	1	0	0	----------	----------	----------
ANO	----------	----------	----------	----------	6	d,c,d,d,ac,ab	1
NE	----------	----------	----------	----------	6	----------	11

Tab. 9: Zhodnocení odpovědí z dotazníku skupiny ***Bez pomoci*** (třída 6.G)

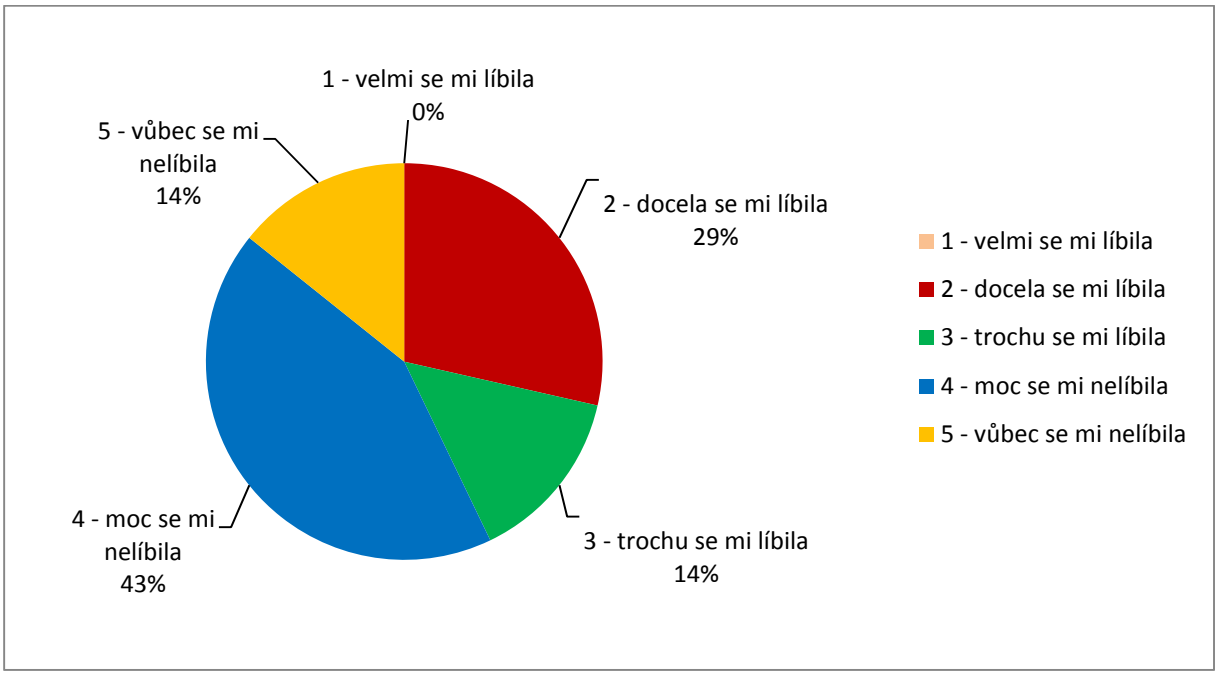
Odpověď	Otázka 1	Otázka 2	Otázka 3	Otázka 4	Otázka 5	Otázka 6	Otázka 7
1	0	0	0	0	----------	----------	----------
2	2	0	2	0	----------	----------	----------
3	1	1	1	1	----------	----------	----------
4	3	2	4	3	----------	----------	----------
5	1	4	0	3	----------	----------	----------
ANO	----------	----------	----------	----------	5	d,bce,bc,bd,bc	6
NE	----------	----------	----------	----------	2	----------	1

Otázka 1 – *Líbila se Ti práce s InterCorpem?*

Graf 2: Vyhodnocení 1. otázky dotazníku skupiny ***S pomocí*** (3.A)



Graf 3: Vyhodnocení 1. otázky dotazníku skupiny ***Bez pomoci*** (6.G)



Naprosté většině, tj. dvěma třetinám studentů, jimž jsem s prací pomáhala, se práce s korpusem InterCorp docela líbila (ohodnotili ji podle školní stupnice na známku 2). Jednomu studentovi se dokonce velmi líbila (známka 1), dvěma se líbila trochu (známka 3) a jednomu se moc nelíbila (známka 4). Průměrná známka činila 2,25.

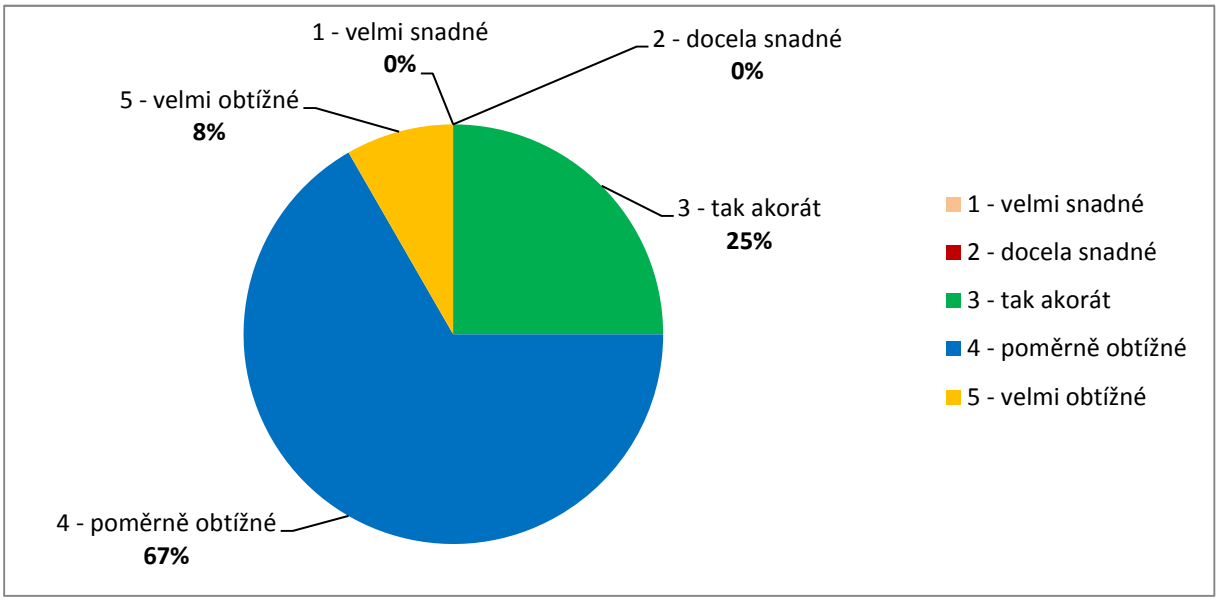
Naproti tomu téměř polovině skupiny, která pracovala samostatně, se práce moc nelíbila (ohodnotili ji podle školní stupnice na známku 4). Jednomu studentovi se vůbec nelíbila (známka 5), jednomu se líbila trochu (známka 3) a zbylým dvěma se docela líbila

(známka 2). Průměrná známka byla více než o 1 stupeň horší než v případě první skupiny, přesněji 3,4.

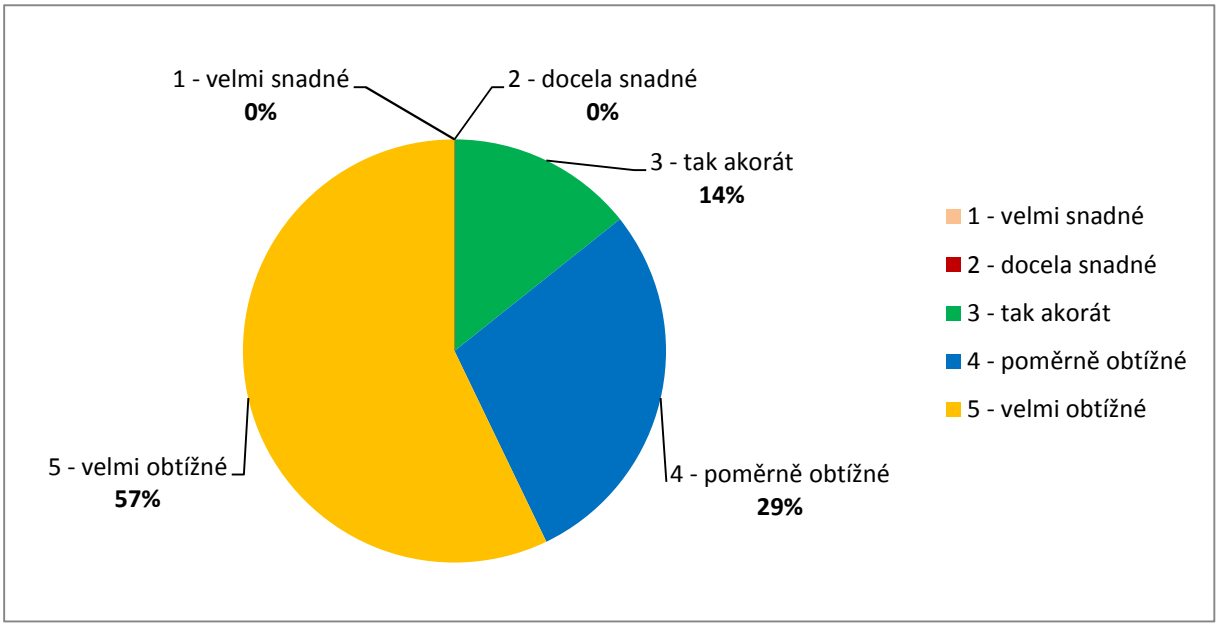
Výsledky svědčí o tom, že studenti, jimž jsem s vypracováváním úkolů pomáhala, hodnotili práci o poznání kladněji než studenti, kteří si museli poradit sami. Navíc se nikomu ze skupiny Bez pomoci práce s InterCorpem nezamlouvala natolik, aby ji ohodnotil známkou 1 (na rozdíl od skupiny S pomocí, kde tuto možnost jeden student zvolil). Jeden účastník z druhé skupiny také uvedl, že se mu práce vysloveně nelíbila (známka 5). Tuto odpověď žádný ze studentů první skupiny nezvolil. V průměru studenti volili známku 3, tedy že se jim práce trochu líbila.

Otázka 2 – *Jak obtížné se Ti zdálo používání InterCorpu?*

Graf 4: Vyhodnocení 2. otázky dotazníku skupiny ***S pomocí*** (3.A)



Graf 5: Vyhodnocení 2. otázky dotazníku skupiny ***Bez pomoci*** (6.G)



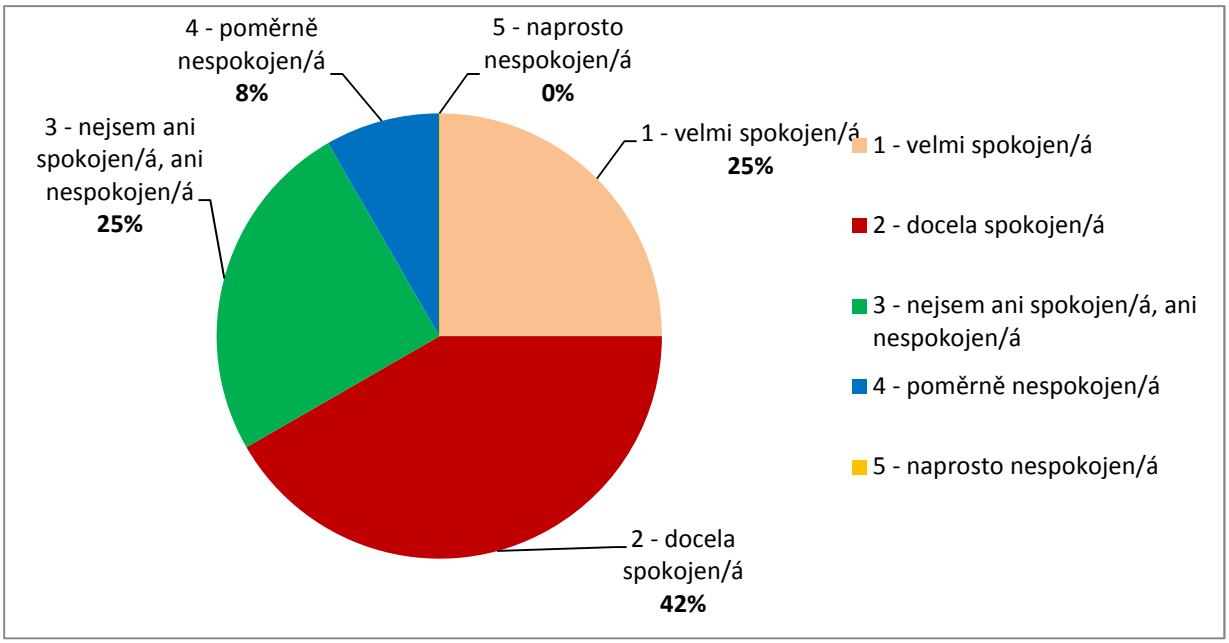
Dvěma třetinám studentů, s nimiž jsem pracovala já (skupina S pomocí), se úkoly zdály poměrně obtížné (známka 4). Čtvrtina z nich zhodnotila obtížnost úkolů známkou 3, tj. obtížné tak akorát. Jednomu studentovi se obtížnost zdála velmi vysoká. Průměrná známka byla 3,8.

U studentů, kteří se s prací potýkali sami (skupina Bez pomoci), převládal názor, že úkoly byly velmi obtížné (známka 5). Úkoly přišly poměrně obtížné (známka 4) dvěma z nich, jednomu se obtížnost zdála adekvátní – tak akorát (známka 3). Průměrně žáci zhodnotili obtížnost na známku 4,3.

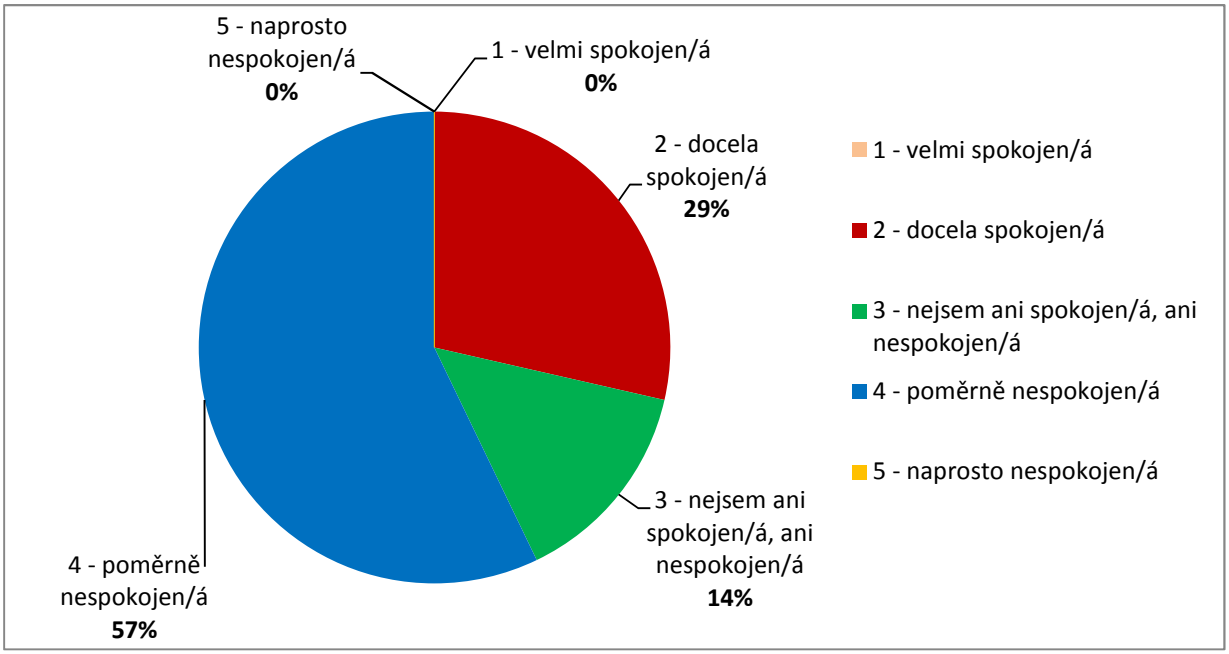
Soudě dle výsledků se žákům, jimž jsem pomáhala, zdála práce nepatrně jednodušší – o 0,5 známky. Ve skupině Bez pomoci ale více než polovina účastníků uvedla, že používání InterCorpu shledali velmi obtížným, zatímco tuto extrémní odpověď zvolil jen jeden student ze skupiny S pomocí. Je také zajímavé, že se ani v jedné skupině nenašel žádný student, který by ohodnotil používání InterCorpu jako velmi snadné nebo docela snadné (známka 1 nebo 2). V globálu se studentům zdálo používání korpusu docela obtížné (známka 4).

Otázka 3 – *Jsi spokojen se svými výsledky?*

Graf 6: Vyhodnocení 3. otázky dotazníku skupiny ***S pomocí*** (3.A)



Graf 7: Vyhodnocení 3. otázky dotazníku skupiny ***Bez pomoci*** (6.G)



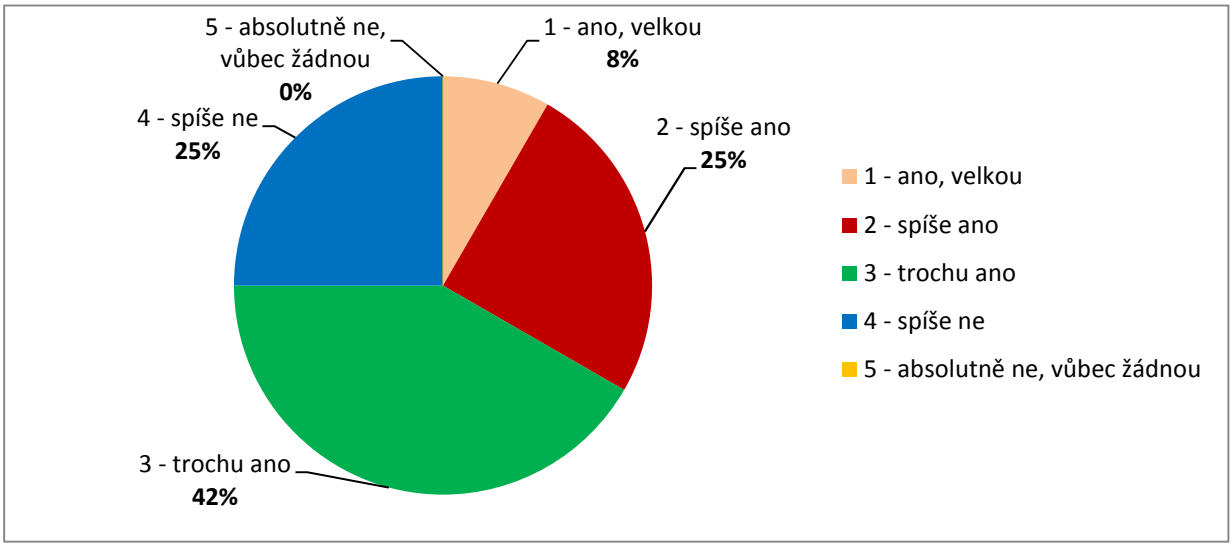
Ze skupiny s pomocníkem byla celá čtvrtina velmi spokojená se svými výsledky (známka 1). Necelá polovina byla docela spokojená (známka 2). Čtvrtina necítila ani spokojenost (známka 3), ani nespokojenost a zbylý jeden člověk byl poměrně nespokojen (známka 4). Průměrná spokojenost se pohybovala těsně pod známkou 2,2.

Skoro třetina studentů, kteří byli odkázáni sami na sebe, byla se svými výsledky docela spokojená (známka 2), ovšem více než polovina se cítila naopak poměrně nespokojená (známka 4). Střední cestu zakroužkoval jeden student (známka 3). Tato skupina byla daleko méně spokojená než skupina Bez pomoci, o celý jeden stupeň – průměrná známka byla 3,3.

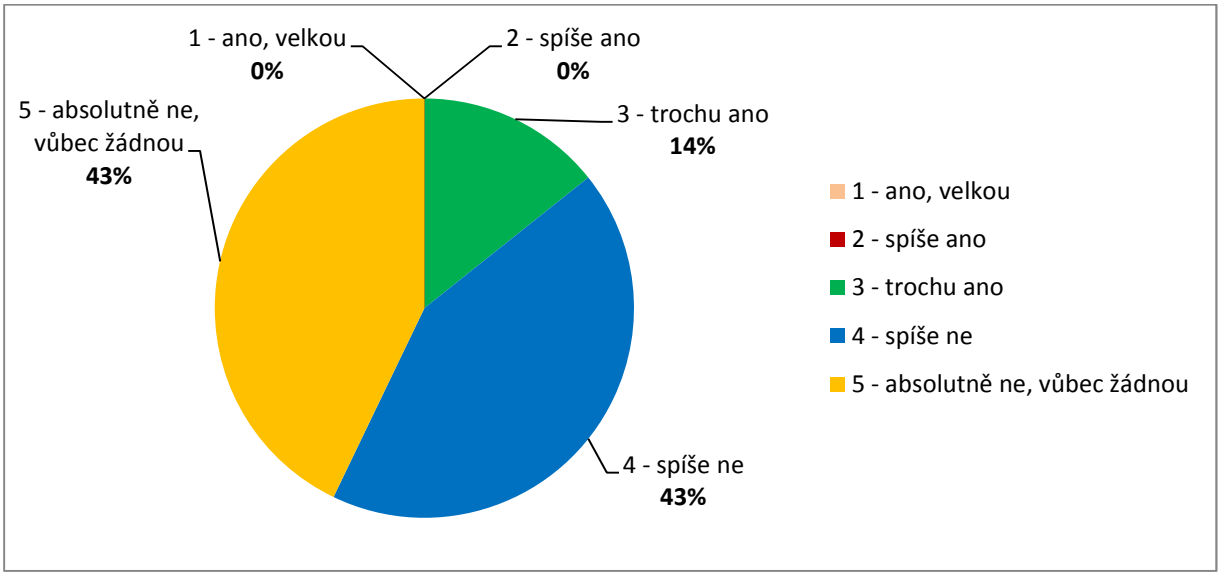
Je tedy jasné, že skupina S pomocí byla znatelně spokojenější se svými výsledky oproti skupině Bez pomoci, která byla vůči svým výsledkům kritická. Absolutní nespokojenost nezvolil nikdo ze zúčastněných, zato tři studenti z první skupiny zhodnotili svou spokojenost na známku 1 (v druhé skupině nebyl nikdo takto spokojený). Průměrně ale byli studenti se svými výsledky spíše spokojení (rozmezí mezi známkou 2 a 3).

Otázka 4 – *Máš chuť InterCorp užívat i nadále?*

Graf 8: Vyhodnocení 4. otázky dotazníku skupiny ***S pomocí*** (3.A)



Graf 9: Vyhodnocení 4. otázky dotazníku skupiny ***Bez pomoci*** (6.G)



Ze studentů, jež jsem motivovala při práci, uvedl záměr využívat korpus InterCorp i nadále jeden žák (známka 1). Čtvrtina zvolila možnost, že chuť do dalšího užívání InterCorpu spíše mají (známka 2), další čtvrtina spíše ne (známka 4). Necelá polovina vyjádřila názor, že jistou chuť k užívání korpusu mají (známka 3). Průměrně studenti odpovídali známkou 2,8.

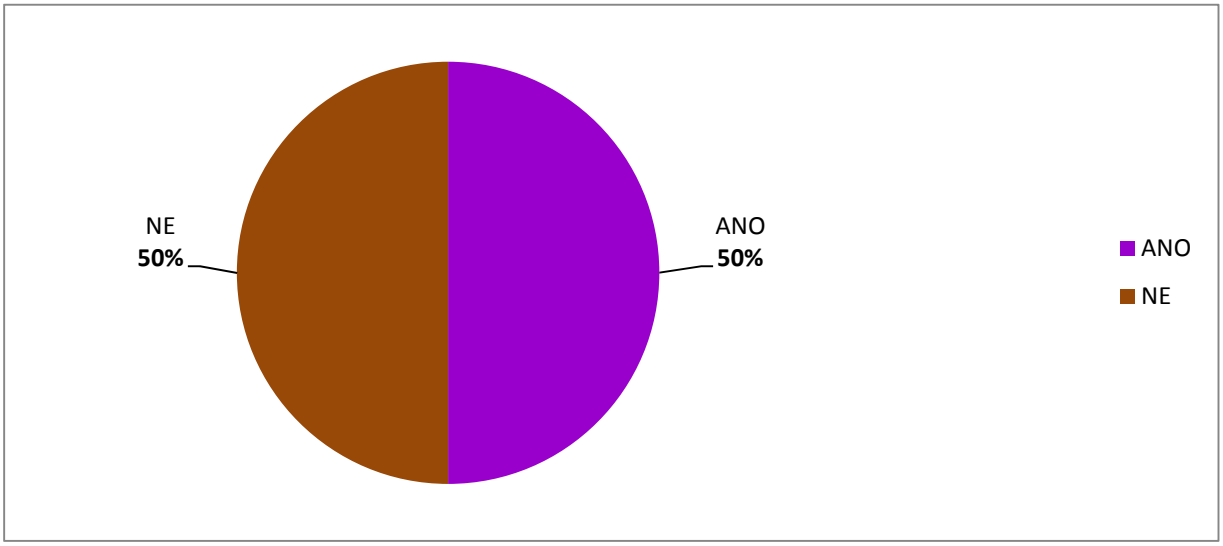
Naopak dvě pětiny dotázaných, kteří pracovali na vlastní pěst, nemají k dalšímu užívání InterCorpu vůbec žádnou chuť (známka 5), stejný podíl (42%) z nich uvedlo, že InterCorp spíše užívat nechtějí (známka 4). Nejlepší známku, a to známku 3, uvedl jeden student, který jistou chuť užívat InterCorp má. Průměrná známka dosahovala hodnoty 4,3.

V této otázce se ukázal nejvyšší rozdíl mezi oběma skupinami, jelikož měl hodnotu jednoho a půl stupně z pěti. Skupina, s níž jsem nepracovala, byla výrazně méně motivovaná

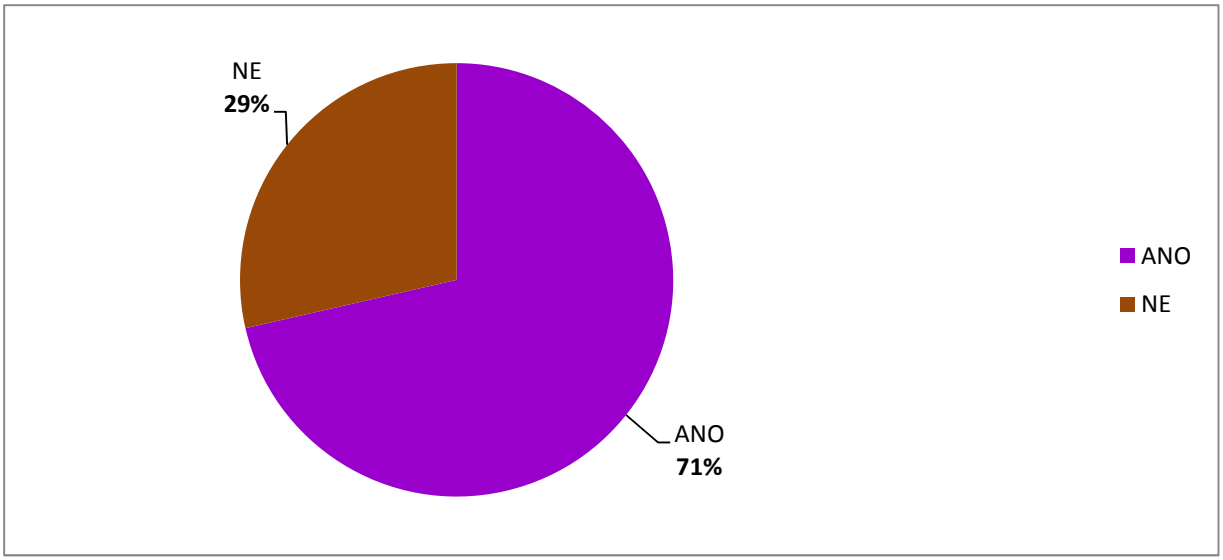
pro další práci s korpusem, přesněji řečeno téměř vůbec. Potvrdila se tak má domněnka, že propagace má na studenty velký vliv.

Otázka 5 – *Odrazuje Tě něco dalšího od dalšího užívání InterCorpu?*

Graf 10: Vyhodnocení 5. otázky dotazníku skupiny ***S pomocí*** (3.A)



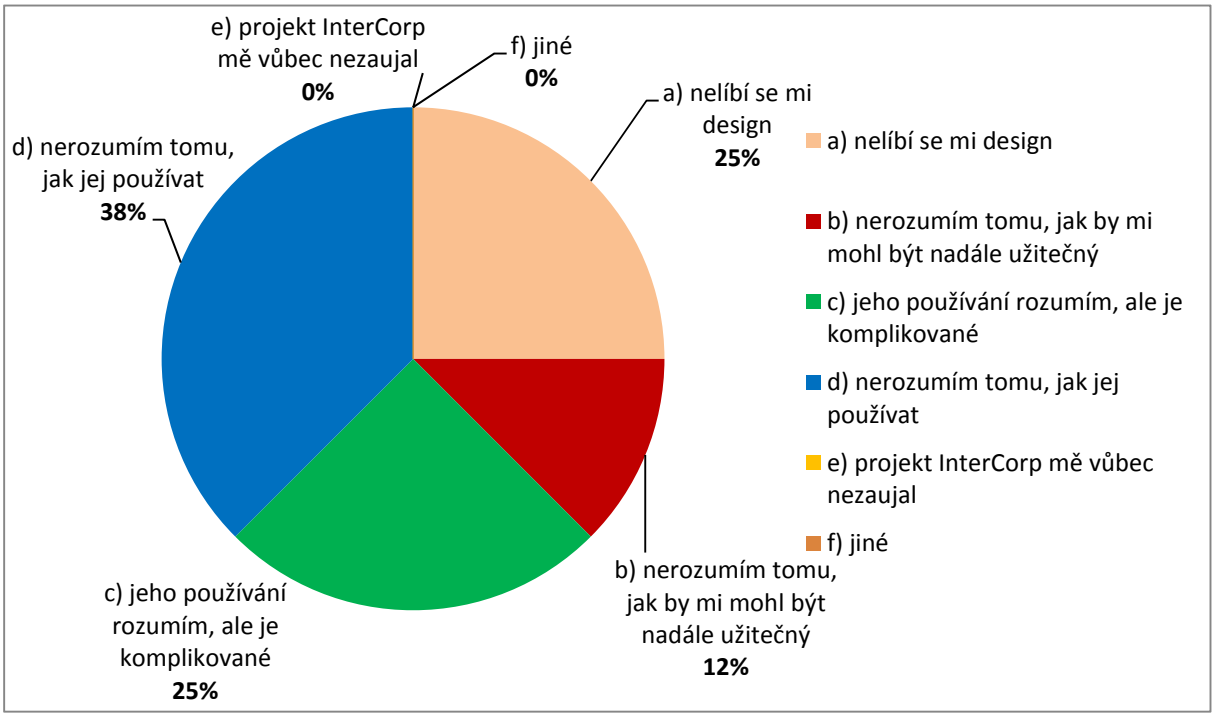
Graf 11: Vyhodnocení 5. otázky dotazníku skupiny ***Bez pomoci*** (6.G)



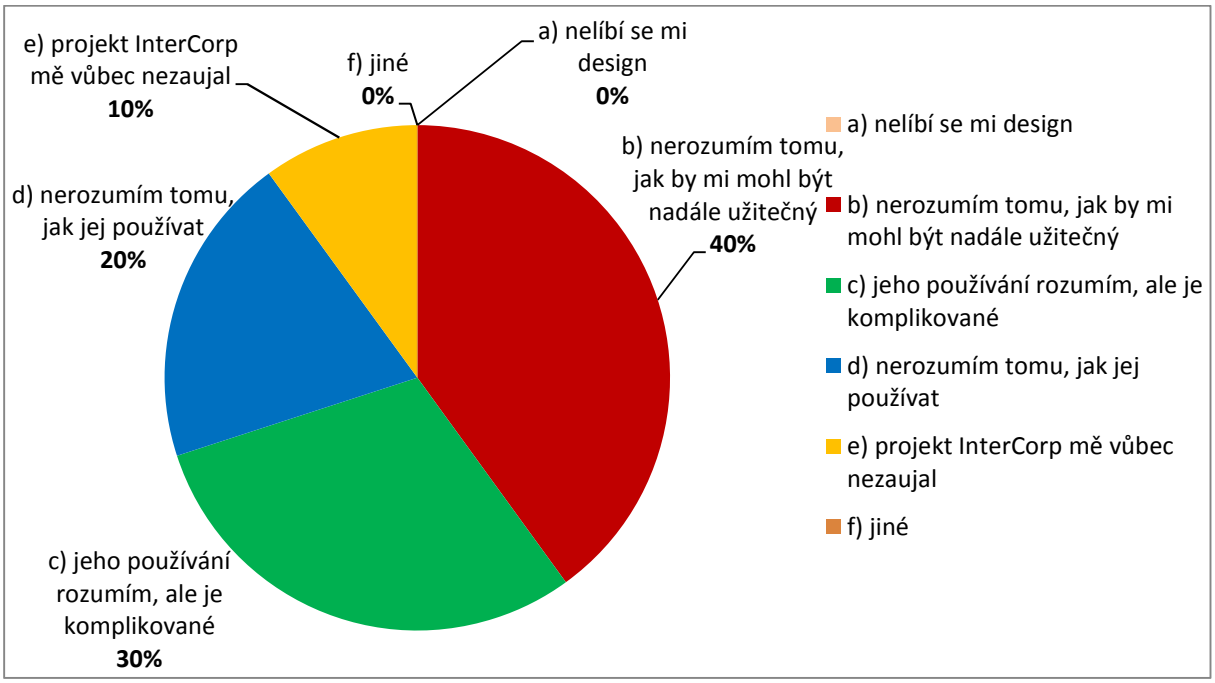
Skupina S pomocí se v odpovědi na otázku, zdali je něco odrazuje od užívání InterCorpu, rozdělila na přesné poloviny. Více než dvě třetiny skupiny Bez pomoci sdílely názor, že existuje nějaká věc, která je od užívání InterCorpu odrazuje. Je tedy zřejmé, že propagace má bezpochyby velký vliv na přístup studentů ke korpusu – rozdíl mezi počtem studentů z 1. skupiny, které něco od používání InterCorpu odrazuje, a počtem studentů z 2. skupiny, kteří sdílejí stejný názor, je více než 20%.

Otázka 6 – *Pokud ano, co to je?*

Graf 12: Vyhodnocení 6. otázky dotazníku skupiny ***S pomocí*** (3.A)



Graf 13: Vyhodnocení 6. otázky dotazníku skupiny ***Bez pomoci*** (6.G)



V této otázce měli žáci možnost zvolit více odpovědí, abych dostala co nejpodrobnější výsledky. Čtvrtinu studentů ze skupiny s pomocníkem, kteří odpověděli kladně na předchozí otázku, odrazuje od dalšího užívání InterCorpu design (nelíbí se jim barvy apod.) – tj. možnost a). Jeden člověk zvolil možnost, že nechápe, jak by mu korpus mohl být nadále užitečný – tj. možnost b). Další čtvrtinu odrazuje jeho používání, kterému rozumí, ale zdá

se jim komplikované – tj. možnost c). Třetina také zvolila možnost, že jim není jasné, jak jej mají používat – tj. možnost d). To je zvláštní, protože tuto odpověď jsem očekávala spíše ve skupině Bez pomoci. Dle mého názoru je důvodem tohoto rozporu fakt, že žádný ze studentů 3.A (tj. skupiny S pomocí) se nezabývá programováním nebo podobnými aktivitami, které podporují analytické myšlení. Z toho vyplývá, že kromě propagace hraje velkou roli v práci s korpusem i zázemí (záliba v jazycích, matematice, logice apod.).

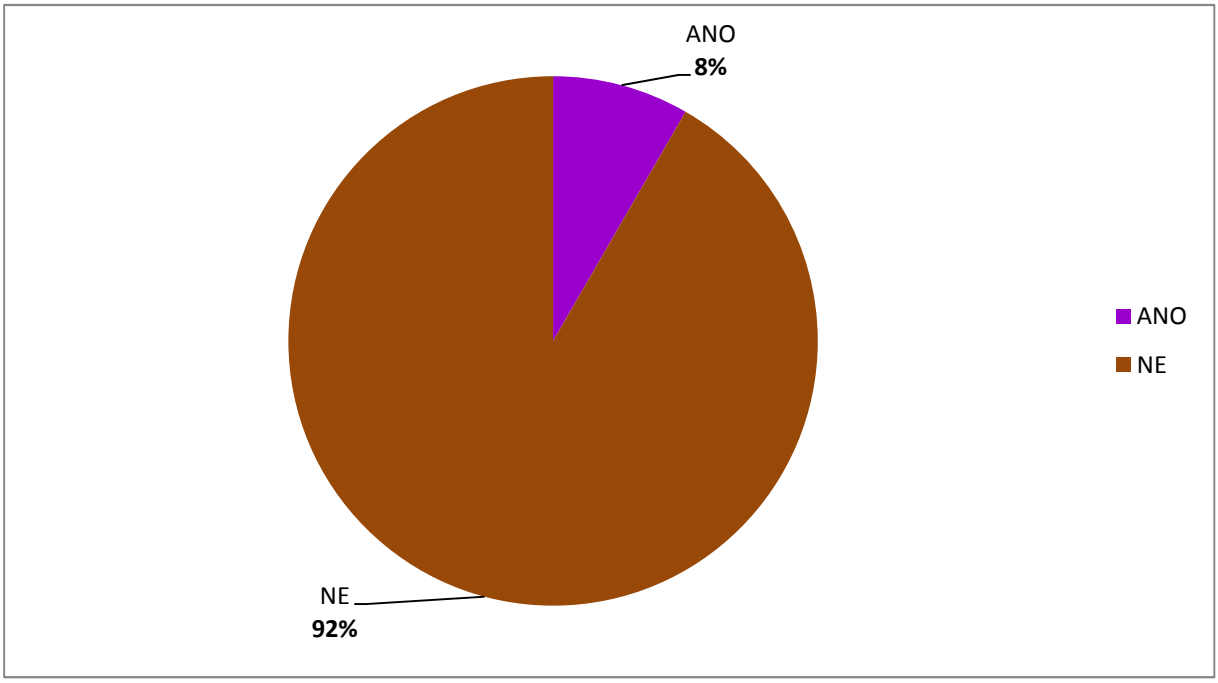
Dvě pětiny respondentů ze skupiny, s níž jsem nepracovala, nerozuměly, jak by jim InterCorp mohl být nadále užitečný – tj. možnost b). Tři studenti rozumí jeho používání, ale zdá se jim komplikované – c). Další dva studenti uvedli, že nerozumí jeho používání – d), jednoho studenta korpus InterCorp vůbec nezaujal – e).

Výsledky ukazují, že skupina, které jsem pomáhala, si daleko negativněji všímala designu, což je logické, protože druhá skupina se soustředila spíše na provozní věci (např. jak InterCorp používat), protože jim používání nikdo nevysvětlil. Zajímavé ale je, že právě tato skupina, které jsem nevysvětlovala používání, uvádí častěji než první skupina, že jeho používání rozumí, ale zdá se jim komplikované. Téměř polovina druhé skupiny také nerozumí tomu, jak by právě jim mohl být InterCorp užitečný, to znamená, že jim chybí motivace. V první skupině tuto možnost zvolil pouze jediný člověk. Pouze ve druhé skupině se jeden respondent rozhodl pro možnost e), totiž že jej projekt InterCorp vůbec nezaujal.

Sedmá otázka byla různá pro skupinu S pomocí a Bez pomoci – obě ale směřovaly k tomu, jak je pro studenty důležitá motivace a pomoc zkušenějšího uživatele.

Otázka 7a – *Myslíš, že by Tě práce s InterCorpem bavila stejně, i kdyby Ti s ní nikdo nepomáhal?*

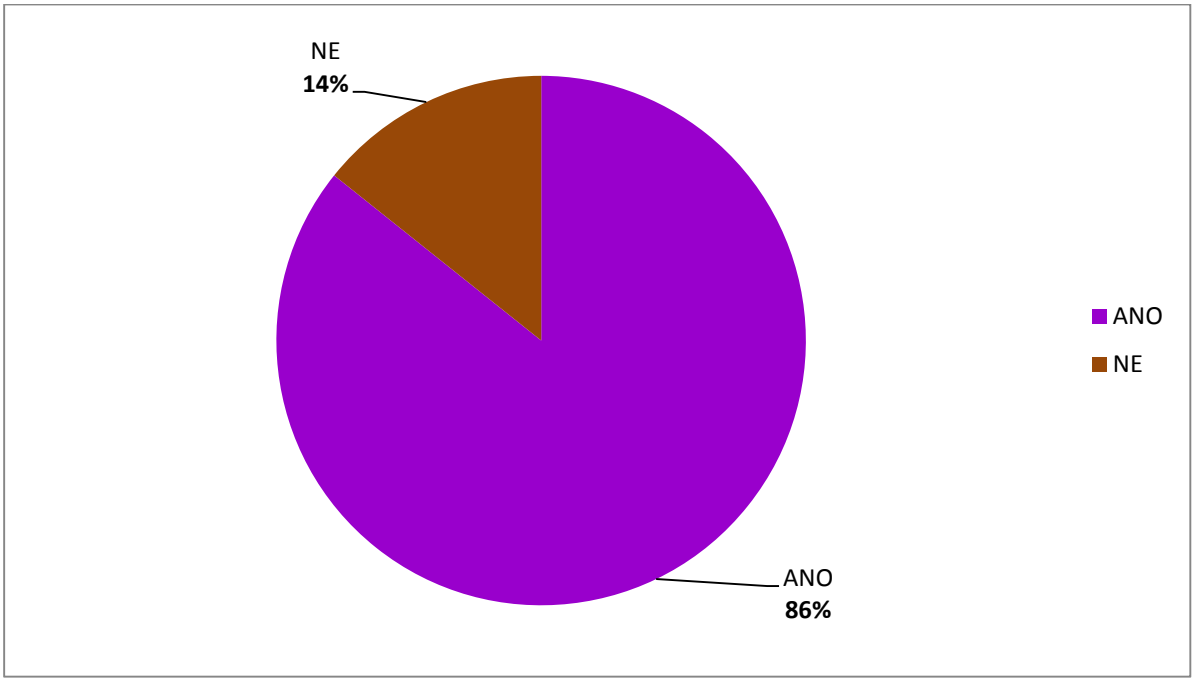
Graf 14: Vyhodnocení 7. otázky dotazníku skupiny ***S pomocí*** (3.A)



Z grafu vyplývá, že naprostá většina respondentů si nemyslí, že by je práce bavila stejně, i kdyby jim s ní nikdo nepomáhal. To dokazuje důležitost jakéhosi prostředníka mezi projektem jako takovým a jeho uživateli. Studenti cítí potřebu mít u sebe někoho, kdo by je motivoval, seznamoval je s jeho využitím a radil jim, jak s korpusem InterCorp zacházet. Na rozdíl od psaného návodu totiž ústní návod umožňuje kladení otázek, což studenti oceňovali. Jelikož úkoly podobného charakteru nikdy neřešili, osobní přístup, který jsem jim poskytovala, studentům pomáhal s orientací v problematice. Samozřejmě i ve zjednodušeném návodu se vyskytovalo mnoho pojmů, kterým žáci rozuměli buď jen zčásti, nebo vůbec, proto jsem jim u každé nejasnosti laicky vysvětlila význam. Účastníci tak nabyli jistoty, že dané úlohy zvládnou. Objektivně však zhodnotili, že bez ústní pomoci by si s úkoly neporadili.

Otázka 7b – *Myslíš, že by Tě práce s InterCorpem bavila víc, kdybys nemusel/a hledat cestu sám/a a někdo Ti vysvětlil, jak na to?*

Graf 15: Vyhodnocení 7. otázky dotazníku skupiny ***Bez pomoci*** (6.G)



Naprostá většina respondentů uvedla, že by je práce s InterCorpem bavila více, kdyby nemuseli cestu hledat sami. To potvrzuje výsledky 7. otázky dotazníku skupiny S pomocí, v níž téměř všichni účastníci uvedli, že by je práce s korpusem nebavila stejně, kdyby jim s ní nikdo nepomáhal. Graf této otázky vypovídá o tom, že studenti ze skupiny Bez pomoci stojí o individuální přístup, ústní pomoc a rady zkušenějšího uživatele. Tento poznatek považuji za velmi důležitý, jelikož také na základě vyhodnocení tohoto bodu dotazníku se pokusím o návrh vylepšení.

5 Závěr

Prvotním podnětem pro tuto práci byla studentská stáž 24 hodin s FF UK, kde jsem se seznámila s PhDr. Olgou Nádvorníkovou, PhD. Ačkoli mi o InterCorpu poskytla mnoho informací, na to, jak korpus využívat, jsem musela přijít sama, což bylo mnohdy zdlouhavé a komplikované. Mým cílem bylo tedy mimo jiné i usnadnit používání InterCorpu dalším zájemcům, nejlépe z řad středoškoláků. Jsem toho názoru, že právě tito studenti jsou ideálními uživateli-začátečníky korpusu, neboť už dosáhli věku, kdy rozumí složitějším problémům. Navíc jsou pořád ještě otevření novým poznatkům.

Jak již bylo výše uvedeno, v první části práce jsem shrnula základní teoretické poznatky o korpusové lingvistice a paralelním korpusu InterCorp. V této části jsem se pokusila zdůraznit všechny způsoby, jak využít korpus, aby si i laik mohl udělat přehled o tom, k čemu může být InterCorp užitečný. V praktické části jsem zkoumala, jak důležitá je pro studenty středních škol propagace korpusu. Zajímalo mě, jestli skupina, která bude mít k dispozici pouze psaný a komplikovaný návod, dosáhne stejných výsledků jako skupina, jíž jsem s prací pomáhala jak ústními radami, tak vysvětlením celé problematiky. Pomocí dotazníků jsem chtěla zjistit, jaké konkrétní aspekty tyto dvě skupiny odlišují a jestli by se rezervy, které vyplynou z vyhodnocení těchto dotazníků, daly eliminovat.

Z dotazníků vyplývá, že žáci, kteří byli během své práce motivováni, považují projekt za smysluplný. Naopak studenti pracující samostatně velkou chuť užívat InterCorp i nadále nemají. Tento kontrast je velice důležitý, ukazuje totiž, že projekt paralelního korpusu InterCorp by úspěch měl i u širší veřejnosti, konkrétně u středoškoláků, kdyby byl více popularizován.

Srovnání postojů obou skupin vůči obtížnosti úkolů také potvrdilo mou domněnku – ten, kdo nemusí hledat cestu sám, bude považovat úlohy za mnohem jednodušší, přestože z objektivního hlediska jde o totožnou úroveň obtížnosti. Přirozeně pak ti, kterým se bude zdát vyhledávání pomocí InterCorpu příliš složité, přestanou projekt využívat.

Dle mého názoru však tento problém lze řešit, proto jsem se také v dotazníku ptala, co konkrétně nové uživatele odrazuje od používání InterCorpu.

1. Co se týče designu vyhledávacího programu, na vyřešení tohoto dílčího problému už pracovníci Ústavu Českého národního korpusu pracují. Na přelomu března a dubna 2013 by mělo být spuštěno nové uživatelské rozhraní, které bude vizuálně vstřícnější a bude obsahovat více užitečných funkcí, lze tedy logicky předpokládat, že veřejnost zaujme daleko více.

2. Složitost zadávání dotazů do korpusu odrazuje také mnoho potenciálních uživatelů. „Používání InterCorpu rozumím, ale je komplikované.“ I tuto odpověď zvolilo několik respondentů. V tomto případě se však nejedná o chybu projektu jako takového. Pokud totiž chceme, aby InterCorp dokázal vyhledávat i velmi komplexní dotazy, nelze jej zjednodušit. Část studentů ovšem naopak uvedla, že nerozumí způsobu používání InterCorpu, což je vcelku logické. Během dvou vyučovacích hodin si totiž nemohou vštípit všechny důležité

aspekty, díky nimž může InterCorp fungovat. Dle mého názoru by žákům pomohl nový, zjednodušený návod, který by jim vysvětlil základní pojmy, jež musí při práci s korpusem znát. Samozřejmostí je také delší časový limit na pochopení fungování InterCorpu. Mimo jiné má velký vliv na rychlost proniknutí do problematiky i zázemí daného studenta (např. jestli se zabývá programováním, jazyky apod.). Jak se ukázalo při srovnání skupin S pomocí a Bez pomoci, ačkoli ve druhé skupině pracovali žáci zcela samostatně, v porovnání s první skupinou si nevedli špatně. Tento výsledek přičítám již zmiňovanému zájmu skupiny Bez pomoci o počítače, matematiku a programování, což velmi usnadnilo jejich práci.

3. Za nejdůležitější ovšem pokládám prvotní motivaci studentů pro práci s InterCorpem. Podstatný je počet studentů, kteří v dotazníku uvedli, že nerozumí, jak by jim mohl být korpus InterCorp nadále užitečný, a kteří zvolili možnost: „Projekt InterCorp mě vůbec nezaujal.“ Právě to, aby žáci žádný z těchto dvou výroků nevybrali, je důležitým předpokladem pro další práci s korpusem. Pokud student nerozumí, jak do korpusu zadávat dotazy, ale chápe, k čemu by mu mohl být dobrý, po čase se s ním naučí pracovat. Jenže pokud student nevidí žádné praktické využití tohoto projektu, popř. jej korpus vůbec nezaujal, logicky jej nebude používat. To potvrzují i reference od účastníků projektu. Jak jsem již výše zmínila, všem studentům zůstal k dispozici účet Jiřina Zelená, který jim může sloužit jako odrazový můstek pro další práci s korpusem, a soudě podle jejich následných komentářů opravdu InterCorp zkusili používat z vlastního zájmu. Dokonce i Mgr. Monika Peroutková, která dohlížela na průběh praktické části v obou skupinách, začala paralelní korpus používat pro ozvláštnění svých hodin.

Pro Ústav Českého národního korpusu by však mohli být potřební nejen odborníci, již využívají korpus pro svůj výzkum a práci, ale i laici, kteří pracují s korpusem spíše ze zájmu. Právě laici totiž mohou poskytnout jiný pohled na velmi složitou problematiku, což je pro Ústav jistě přínosné.

Výsledky dále naznačují, že návod na webových stránkách Ústavu Českého národního korpusu je sice velmi precizně sestavený, laikovi ale s používáním InterCorpu příliš nepomůže. Pokud tedy tvůrci stojí o masovější rozšíření projektu, bylo by dobré umístit na stránky nějaký stručnější návod, který by pomohl začátečníkovi s orientací v problematice. Stačilo by, aby se takový návod dotýkal jen základních věcí, např. jak se přihlásit do systému, jak vyhledávat nejjednodušší věci a jak zpracovávat výsledky.

Ústavu Českého národního korpusu jsem navrhla i další řešení problému popularizace, a to mnou vytvořený laičtější návod *(viz kapitola 7.2)*. Jeho výhodu spatřuji právě v tom, že na problematiku nahlížím z pohledu méně zkušeného uživatele – nejsem tedy ani nezkušený uživatel, ani tvůrce projektu. Jsem toho názoru, že návod bude pro začátečníka pochopitelnější než dlouhý a komplikovaný návod, po částech umístěný na různých odkazech webových stránek www.korpus.cz . Tuto domněnku potvrdili i respondenti dotazníků.

Abych nalákala zcela nové potenciální uživatele paralelního korpusu InterCorp, rozhodla jsem se natočit motivační video, které Ústav Českého národního korpusu umístí na své webové stránky. Cílem videa je propagace korpusu pro širší vrstvy. Inspirovala jsem se v protidrogových kampaních, kde se osvědčily takzvané peer programy (z angl. peer –

vrstevník). To znamená, že cílová skupina nejlépe akceptuje informace a postřehy od svého vrstevníka. Tento postup vyplynul z mých pozorování během praktických hodin – obě skupiny oceňovaly pomoc od někoho na stejné úrovni. Proto tedy hlavním propagátorem InterCorpu ve videu budu já jakožto spolužačka z jiné střední školy, která už jisté zkušenosti s projektem má. Motivační video ve formátu .wmv je součástí práce jako příloha na CD. *(Scénář motivačního videa viz kapitola 7.5, odkaz na motivační video viz kapitola 7.6)*

V dubnu 2013, tedy už po uzávěrce této práce, plánuji natočit minimálně ještě jedno video ve spolupráci s ÚČNK. Jeho cílem ale nebude motivace, nýbrž zjednodušení ústního návodu pro práci s InterCorpem. V průběhu praktické části jsem totiž přišla například na to, že velkým problémem pro naprosto nezkušené studenty je i samotné přihlášení do rozhraní. Plánuji tedy seznámit s fungováním korpusu toho potenciálního uživatele, který už předtím zhlédl motivační video a rozhodl se, že se chce s InterCorpem naučit pracovat. Druhé, návodné video by mu mělo práci značně ulehčit a zčásti samozřejmě opět motivovat uživatele-začátečníka.

Mezi výstupy mé práce tedy patří v první řadě motivační video, které, jak doufám, nadchne nové uživatele InterCorpu. Druhým přínosem práce je vytvořený návod. Ten nezahlcuje začátečníka zbytečnými odbornými termíny, jež si pokročilejší uživatel může najít na webových stránkách korpusu. Základní názvosloví je vymezeno, nicméně značně zjednodušeno a doplněno o praktické příklady. Kromě toho má práce nabízí i pracovní list, který slouží k prověření znalostí. Další uživatelé proto nemusí vymýšlet, jaké konkrétní výrazy vyhledávat, protože mnou vytvořené úkoly mu poslouží jako model.

Navíc mi bylo nabídnuto vystoupit s prezentací paralelního korpusu v Městské knihovně v Třebíči v rámci Francouzského klubu, což je program, který si klade za cíl propagovat francouzský jazyk, kulturu, historii i různé zajímavosti. Má přednáška by přinesla možnost seznámit s projektem nejen studenty třebíčského gymnázia, ale i potenciální zájemce z jiných věkových kategorií.

O popularizaci projektu se nyní značně snaží i samotný Ústav Českého národního korpusu. Na 6. září 2013 je naplánován jednodenní workshop pro všechny zájemce, který bude zaměřen právě na využití paralelního korpusu InterCorp.

Jak jsem již mnohokrát výše uvedla, jsem přesvědčená o tom, že paralelní korpus InterCorp by mohl být užitečný širší veřejnosti. Experimentem jsem dokázala, že korpus lze využít i na středních školách, proto doufám, že výsledky mého šetření přispějí ke zlepšení propagace projektu mezi středoškoláky po celé České republice. Videa, stručnější návod, ale hlavně osobní přístup rozšíří povědomí o projektu paralelního korpusu InterCorp, který může být užitečný komukoli, kdo má zájem o jazyky.

6 Seznam bibliografických citací

Literatura

CVRČEK, V. – KOVÁŘÍKOVÁ, D.: *Možnosti a meze korpusové lingvistiky*. Naše řeč 94, 2011, s. 113-133.

ČERMÁK, F.: *InterCorp: jeho povaha a možnosti*. Ústav Českého národního korpusu, Karlova Univerzita v Praze 2011.

ČERMÁK, F. (ed.): *Korpusová lingvistika Praha 2011 – 2 Výzkum a výstavba korpusů*. Nakladatelství Lidové noviny, Praha 2011.

ČERMÁK, F. – BLATNÁ, R. (eds): *Jak využívat Český národní korpus*. Nakladatelství Lidové noviny, Praha 2005.

ČERMÁK, F. – BLATNÁ, R. (eds): *Manuál lexikografie*. Nakladatelství a vydavatelství H&H, Praha 1995.

ČERMÁK, F. – KŘEN, M. (eds): *Frekvenční slovník češtiny*. Nakladatelství Lidové noviny, Praha 2004.

Francouzsko-český slovník. FIN Publishing, Olomouc 1998.

KOCEK, J. – KOPŘIVOVÁ, M. – KUČERA, K. (eds): *Český národní korpus – úvod a příručka uživatele*. ÚČNK FF UK, Praha 2000

NÁDVORNÍKOVÁ, O.: *Analýza predikačního potenciálu francouzských tvarů na –ANT*. Diplomová práce, Filozofická fakulta Univerzity Karlovy v Praze, ved. J. Tláskal, Praha 2003.

Internetové zdroje

URL: <http://ucnk.ff.cuni.cz/co_je_korpus.php> [Cit. 14.12.2012].

URL: <http://www.ucl.ac.uk/english-usage/about/history.htm> [Cit. 15.12.2012].

URL: <http://icame.uib.no/brown/bcm.html> [Cit. 15.12.2012].

URL: <http://www.corpus4u.org/forum/upload/forum/2005052811133696.pdf> [Cit. 15.12.2012].

URL: <http://en.wikipedia.org/wiki/Corpus_linguistics> [Cit. 15.12.2012].

URL: <http://ucnk.ff.cuni.cz/n_neref.html> [Cit. 17.12.2012].

URL: <http://ucnk.ff.cuni.cz/struktura.php> [Cit. 17.12.2012].

URL: <http://ucnk.ff.cuni.cz/struktura.php> [Cit. 26.12.2012].

URL: <http://www.korpus.cz/intercorp/?req=page:info> [Cit. 15.12.2012].

URL: <http://www.korpus.cz/intercorp/?req=page:releaseNotes> [Cit. 15.12.2012].

URL: <http://www.korpus.cz/intercorp/dokumenty/VZpara.pdf> [Cit. 15.12.2012].

URL: <http://www.korpus.cz/intercorp/?req=page:info> [Cit. 27.12.2012].

URL: <http://www.korpus.cz/intercorp/?req=page:info> [Cit. 27.12.2012].

URL: <http://www.korpus.cz/intercorp/> [Cit. 10.2.2013].

URL: <http://utkl.ff.cuni.cz/~rosen/public/pc_short_2008.pdf> [Cit. 15.12.2012].

7 Přílohy

7.1 Návod, kompilovaný z informací na webových stránkách ÚČNK, který sloužil jako prvotní zdroj informací pro skupinu „Bez pomoci“

Stručný návod pro práci s uživatelským rozhraním korpusu InterCorp - Park

Přístup ke korpusu InterCorp je možný s přístupovými údaji, které jsou platné i pro ostatní korpusy ČNK. Pokud zatím nemáte přístup ke korpusům ČNK, můžete si jej zřídit zdarma po registraci na stránce: **Registrace** nebo můžete použít odkaz v horní části přihlašovací stránky.

Po zadání uživatelského jména a hesla se otevře stránka se seznamem dotazů. Po přihlášení bude k dispozici pouze odkaz k založení nového dotazu. Pokud se na tuto stránku vrátíte později z navigačního menu, můžete zde přepínat mezi více aktivními dotazy.

uživatelské jméno: ***zelenajirina***
heslo: ***intercorp***

specifikace prohledávané části korpusu

Po založení dotazu budete přesměrováni na stránku se seznamem jazyků a textů, které jsou v současné době k dispozici. Nejprve je třeba levým tlačítkem myši označit alespoň dva jazyky pomocí zaškrtávacího okénka u každého z nich.

Pokud chcete hledat ve všech textech v jádru a nezajímají Vás kolekce automaticky zpracovaných textů ani výběr konkrétních textů, se kterými chcete pracovat, můžete v tomto okamžiku rovnou přejít k zadání dotazu.

Jestliže chcete kromě jádra ručně zkontrolovaných textů hledat i v kolekcích automaticky zpracovaných textů, můžete u jednotlivých balíčků nastavit **Zařadit**, pokud je daný balíček k dispozici pro Váš výběr jazyků. Tím přidáte k celému jádru i konkrétní balíčky a můžete opět jít přímo k dotazu.

Chcete-li z dostupných souborů vybrat na základě známých parametrů pouze některé, můžete použít filtr. Při jeho nastavování postupujte následovně:

- Nejprve se ujistěte, že máte vybrány jazyky, které chcete prohledávat
- Potom musíte zvolit jazyk, podle kterého budete chtít filtrovat. (To se provede opět levým tlačítkem myši, tentokrát ale poklepem přímo na nápis jazyka, podle kterého chcete filtrovat. Ten by se měl označit tmavší šedou barvou. Pokud zvolíte za jazyk filtru např. angličtinu, zobrazí se v seznamu textů informace o anglických verzích, a pokud zvolíte například filtr rok vydání do r. 2001, řídí se filtr rokem vydání anglické verze textu.)
- Potom můžete v oblasti filtrů zaškrtat nebo naopak odebrat typy textů, podle známých bibliografických údajů.
- Následně můžete použít filtr i na kolekce automaticky zpracovaných textů – volba: **Podle filtru**. (Pokud použijete volbu **Zařadit**, budou vybrány texty z jádra podle filtru, ale daný balíček bude zařazen celý, nezávisle na filtru. Pokud použijete volbu **Nezařazovat**, daný balíček se prohledávat nebude, a to bez ohledu na nastavení filtru.)
- Nakonec můžete ještě výběr podle filtru ručně doladit. Pokud si zapnete **Ruční výběr textů**, zobrazí se seznam konkrétních textů.

- Když použijete **Filtruj texty**, zobrazí se zaškrtnutí pouze u textů, které odpovídají výše nastavenému filtru, ale toto nastavení můžete ještě ručně upravit podle Vašich požadavků.

Přitom je potřeba dodržet určené pořadí operací, protože změna v předchozích krocích má obvykle za následek reset nastavení všech následujících kroků.

zadání dotazu

- hledání v jednom jazyce nebo ve více jazycích současně
- hledání podle slovního tvaru
- hledání podle posloupnosti tvarů (**Slovní spojení**)
- hledání podle výrazu jazyka CQL (analogicky jako v české části ČNK, viz např. **zde**)
- hledání podle lemmatu (základního tvaru) - pro některé jazyky
- hledání podle **morfosyntaktické značky** (tagu), je třeba zadat ve formátu CQL - pro některé jazyky
- možnost využít při zadání dotazu **regulární výrazy**
- možnost využít při zadání dotazu virtuální klávesnici

možnosti zobrazení paralelních konkordancí

- zobrazení strukturních značek (**Konkordance/Zobraz možnosti/Struktury**)
- zobrazení bibliografických údajů a identifikace konkordance (**Konkordance/Zobraz možnosti/Reference**)
- zobrazení lemmatu a/nebo morfosyntaktické značky, pro klíčové slovo nebo všechna zobrazená slova - pro některé jazyky (**Konkordance/Zobraz možnosti/Atributy**)
- "filtrování" výsledků na základě přítomnosti nebo nepřítomnosti zadaných výrazů v kontextu (**Konkordance/Zobraz filtr**)
- zobrazení celých vět (**Segment**) nebo řádků s hledaným výrazem ve středu (**Kwic**)
- zobrazení více jazyků ve sloupcích vedle sebe nebo v řádcích pod sebou (**Pohled: vertikální/horizontální**)
- zobrazení širšího kontextu (**zobraz kontext**)
- export konkordancí ve formátu tabulky (**Export: xls1, xls2**)

možnost návratu k zadaným dotazům a výsledkům

- nahoře v navigaci klikněte na **Domů**

Dotazovací jazyk korpusového manažeru Bonito

Michal Křen

I. Regulární výrazy

Ještě než začneme s výkladem o regulárních výrazech, je třeba přesně vysvětlit pojmy pozice, atribut a strukturní značka. ***Pozicí*** rozumíme libovolné jedno slovo nebo interpunkční znaménko, zjednodušeně řečeno cokoli, co lze najít korpusovým manažerem.

Korpusový manažer Bonito umožňuje při vyhledávání používat tzv. ***regulárních výrazů***. Zjednodušeně řečeno se jedná o vkládání určitých speciálních znaků se zvláštním významem do slov, která chceme vyhledat. S podobnou, i když velmi primitivní formou se můžeme setkat třeba při hledání souborů ve Windows (pomocí menu Start > Hledat), při kterém nemusíme znát přesně celý název souboru. Pokud si například u souboru *seznam.txt* nepamatujeme jeho příponu, stačí pro hledání

zadat **seznam.*** - hvězdička v tomto případě zastupuje libovolnou koncovku. Podobně lze postupovat i při vyhledávání pomocí korpusového manažeru, jeho možnosti jsou však daleko širší. Chceme-li tedy například v korpusu vyhledat všechny tvary slova *bobule*, nechceme je všechny vypisovat a nechceme ani používat lemmatizaci, je možné zadat dotaz **bobul.*** - tečka zde zastupuje libovolný znak a hvězdička libovolný počet opakování předchozího (tj. libovolného) znaku. Musíme však počítat s tím, že manažer vyhledá všechna slova začínající *bobul*, tedy například i adjektivum *bobulovité* apod. Sekvence .* zastupuje tedy libovolnou část slova a je zřejmě vůbec nejčastěji používanou součástí regulárního výrazu. Může se vyskytnout samozřejmě i na začátku nebo uprostřed slova. V regulárních výrazech lze kromě běžných znaků používat všechny následující speciální znaky:

- **tečka** (.) - *představuje jeden libovolný znak,*
- **interval** ({n, k}) - *představuje n až k opakování předchozího znaku nebo výrazu; je-li k vynecháno, odpovídá intervalu nejméně n opakování, pokud má interval tvar {n}, odpovídá mu přesně n opakování;*
- **hvězdička** (*) - *představuje libovolný počet (0 a více) opakování předchozího znaku nebo výrazu, totéž co* {0,}
- **plus** (+) - *představuje 1 nebo více opakování předchozího znaku nebo výrazu, totéž co* {1,}
- **otazník** (?) - *představuje žádný nebo jeden výskyt předchozího znaku nebo výrazu, totéž co* {0,1}
- **seznam** ([]) - *představuje alternativu - libovolný jeden znak z těch, které jsou uvedeny v seznamu uvnitř hranatých závorek; pokud je prvním znakem seznamu stříška (^), jde o negovaný seznam a představuje tedy libovolný jeden znak kromě těch uvedených uvnitř závorek; v rámci seznamu je možné používat také pomlčku (-) jako operátor rozsahu,*
- **svislá čára** (|) - *představuje také alternativu, ne ovšem mezi jednotlivými znaky, ale celými řetězci,*
- **kulaté závorky** - *libovolnou část výrazu je možné seskupit do kulatých závorek a ovlivnit tak prioritu jeho vyhodnocování,*
- **zpětné lomítko** (\) - *pokud některý speciální znak předchází zpětné lomítko, ztrácí tento znak svůj zvláštní význam; následuje-li však za zpětným lomítkem číslice, považuje se číslo za lomítkem za oktalový kód znaku v ISO Latin 2 - to je výhodné zejména pro uživatele v zahraničí, kteří nemají k dispozici českou klávesnici.*

Příklady regulárních výrazů najdete v následující tabulce:

Příklad	regulární výraz
všechny tvary slova *ptakopysk*	**ptakopys.***
slovo *kdy* s malým nebo velkým počátečním písmenem	**[kK]dy**
tečka jako interpunkční znaménko	**\.**
infinitivy předponových sloves od *nést*	**.+nést**
různě dlouhé varianty citoslovce *ratata*	**ra(ta)+**
pravopisnou dubletu: *diskuze* psané i se *s*	**diskuse|diskuze** *nebo* **disku[sz]e**
morfologické varianty slova *smích* (s vyloučením tvarů odvozených od slov *Smíchov* a *smíchat*)	**[Ss]mích[^oaá].*|[Ss]mích**
libovolná čísla skládající se ze tří nebo čtyř cifer	**[0-9]{3,4}**

II. Dotazovací jazyk korpusového manažeru

Celý dotazovací jazyk korpusového manažeru je vlastně rozšířením výše uvedených možností regulárních výrazů a jejich aplikací na vyhledávání v korpusu. Klíčem k němu je porozumění rozdílu mezi dotazy **bobul.***, **"bobul.*"** a **[word="bobul.*"]**. Ve všech případech jde o dotaz na jednu pozici. Podstatné však je, že máme-li nastavený jako implicitní atribut *word*, jsou všechny tři výše uvedené dotazy ekvivalentní, tj. dávají přesně týž výsledek. Není-li implicitním atributem *word*, výsledky stejné nebudou, protože první dva dotazy jsou příkladem zjednodušeného zápisu dotazu na implicitní atribut korpusu (tímto atributem je po spuštění manažeru vždy *word*, ale obecně jím být nemusí). První z těchto dotazů je však zjednodušen natolik, že je lze použít v případě víceslovných výrazů jen někdy, často je třeba hledané slovo uzavřít do uvozovek. Konečně poslední, třetí případ s explicitním uvedením atributu je naopak nejobecnější a funguje vždy stejně bez ohledu na nastevení implicitního atributu. Rozdíl mezi těmito třemi případy nyní demonstrujeme na příkladu hledání frazému *držet krok*. Chceme-li jej najít v této podobě, stačí zadat do manažeru dotaz **držet krok**.V tomto případě je však výhodné využít lemmatizaci, a tedy se ptát nikoli pouze na infintiv, ale na všechny tvary slovesa. K tomu je ovšem třeba již použít zápis**[lemma="držet"] [word="krok"]** nebo jednodušeji - je-li implicitním atributem *word* - **[lemma="držet"] "krok"**. Protože však jde o dotaz na dva různé atributy (nejen *word*, ale i *lemma*), není v tomto případě možné dotaz ještě více zjednodušit a napsat *krok* bez uvozovek. Uvedený dotaz je dále možné rozšiřovat a zobecňovat, přičemž uvnitř uvozovek je samozřejmě možné použít libovolný regulární výraz.

- Další možnosti uvádí následující tabulka:

Příklad	dotaz
výraz *držet krok*	**držet krok**
jeden z výrazů *držet krok* nebo *udržet krok*	**u?držet krok**
dtto, ale sloveso nemusí být v infinitivu	**[lemma="u?držet"] "krok"**
dtto, ale i substantivum může být v libovolném tvaru	**[lemma="u?držet"] [lemma="krok"]**
dtto, navíc je možný výskyt až pěti slov mezi nimi	**[lemma="u?držet"] []{0,5} [lemma="krok"]**
ekvivalentní s předchozím dotazem	**[lemma="držet" | lemma="udržet"] []{0,5} [lemma="krok"]**

Předchozí tabulka si zaslouží několik vysvětlujících komentářů. V první řadě je třeba zdůraznit, že vždy platí (s výjimkou výše popsaného zjednodušeného zápisu bez explicitního uvedení atributu), že každé pozici odpovídá hranatá závorka, ve které je určena podmínka, kterou musí splňovat určitý atribut (nulovou podmínku prázdných hranatých závorek splňuje libovolná pozice). Dále si všimněte, že některé ze speciálních znaků regulárních výrazů mají stejný nebo analogický význam i mimo regulární výrazy, jinými slovy fungují nejenom v rámci slov, ale i mimo něj. Těmito znaky jsou složené závorky označující interval ({n,k}) a jejich speciální případy, tj. hvězdička (*), plus (+) a otazník (?), a dále kulaté závorky a svislá čára (|), jak je vidět v posledním příkladu. Ta značí disjunkci, pro konjunkci lze dále použít ampersand (&) a pro negaci vykřičník (!).

7.2 Zjednodušený návod, který jsem vytvořila pro skupinu „Bez pomoci“

Návod na práci s paralelním korpusem InterCorp

1. Korpus je dostupný ze stránek http://www.korpus.cz/intercorp/. InterCorp je přístupný přes rozhraní Park. Kliknutím na Hledání v korpusu se dostanete na přihlašovací stránku, kde zadáte *stejné přihlašovací údaje* jako pro Český národní korpus.

2. V levé liště vyberete *korpusy jazyků*, s nimiž chcete pracovat. Můžete prohledávat buď jeden, nebo i více jazyků zároveň (proto korpus paralelní).

   Na této stránce si můžete vybrat i další *filtry* – například jazyk originálu, pohlaví autora či překladatele. Acquis je korpus právnických textů v jazycích EU, Presseurop a Syndicate jsou publicistické texty. Dále můžete texty filtrovat i ručně pomocí *Ručního výběru textů* – například pouze texty od René Goscinnyho.

3. Nyní je důležité vybrat, pomocí kterého vodicího ukazatele budete dotaz pokládat.
   Lemma je zobecněný tvar zastupující všechny ostatní tvary daného slova, například infinitiv u sloves, 1. pád u podstatných jmen, rod mužský u přídavných jmen apod. Zde je důležité zadávat jedině tyto základní tvary, jinak InterCorp nenajde žádný výsledek.

   Slovní tvar použijte tehdy, hledáte-li jednu konkrétní formu slova. Když zadáte například slovo *jít*, slovní tvar vám najde jen ty výskyty, kde je přímo slovo jít, kdežto lemma zobrazí i *šel*, *jdeme*, *nešlo* apod.

   Slovní spojení najde posloupnosti tvarů. Pouze tímto příkazem lze tedy vyhledávat spojení dvou a více slov – například byl rád (zadáte-li, že má korpus hledat toto spojení například v lemmatu či slovním tvaru, nenajde žádné výsledky).

   Do **CQL** musíte zadávat dotaz pomocí *regulárních výrazů*, kde:

- **tečka** (.) - *představuje jeden libovolný znak,*
- **interval** ({n, k}) - *představuje n až k opakování předchozího znaku nebo výrazu; je-li k vynecháno, odpovídá intervalu nejméně n opakování, pokud má interval tvar {n}, odpovídá mu přesně n opakování;*
- **hvězdička** (*) - *představuje libovolný počet (0 a více) opakování předchozího znaku nebo výrazu, totéž co* {0,}
- **plus** (+) - *představuje 1 nebo více opakování předchozího znaku nebo výrazu, totéž co* {1,}
- **otazník** (?) - *představuje žádný nebo jeden výskyt předchozího znaku nebo výrazu, totéž co* {0,1}
- **seznam** ([]) - *představuje alternativu - libovolný jeden znak z těch, které jsou uvedeny v seznamu uvnitř hranatých závorek; pokud je prvním znakem seznamu stříška (^), jde o negovaný seznam a představuje tedy libovolný jeden znak kromě těch uvedených uvnitř závorek; v rámci seznamu je možné používat také pomlčku (-) jako operátor rozsahu,*
- **svislá čára** (|) - *představuje také alternativu, ne ovšem mezi jednotlivými znaky, ale celými řetězci,*
- **kulaté závorky** ( )- *libovolnou část výrazu je možné seskupit do kulatých závorek a ovlivnit tak prioritu jeho vyhodnocování,*
- **zpětné lomítko** (\) - *pokud některý speciální znak předchází zpětné lomítko, ztrácí tento znak svůj zvláštní význam; následuje-li však za zpětným lomítkem číslice, považuje se číslo za*

lomítkem za oktalový kód znaku v ISO Latin 2 - to je výhodné zejména pro uživatele v zahraničí, kteří nemají k dispozici českou klávesnici.

Tyto regulární výrazy ovšem můžete využít i ve slovním tvaru. Ve slovním tvaru lze například zadat pouze .*ání, kdežto do CQL musíte zadat [word=".*ání"] .

4. Ukazatel CQL umožňuje zadávat nejkomplexnější dotazy. K tomu slouží příkaz *tag* neboli značka. Je však potřeba do dotazu napsat, že jde o tag, a to v tomto tvaru: **[tag=" "]** . Do mezery se pak doplní potřebné znaky. Stejným způsobem funguje i lemma. Pokud chcete zadat příkaz, v němž bude například kombinace lemmatu s jiným lemmatem nebo s tagem nebo s určitým slovem, bude tvar vypadat takto: **[lemma=" "] [tag=" "]** / **[lemma=" "] [word=" "]** apod.

CQL bude fungovat pouze tehdy, určíte-li, zda jde o lemma, tag nebo slovo. To znamená, že do hranatých závorek musíte napsat buď lemma, tag nebo word.

Jelikož každá značka je řetězcem 16 znaků, můžete daný výraz omezit až 16 kategoriemi. Chcete-li ovšem vyhledat například jen slovesa(v jakékoli osobě, čísle, čase, …), tvar bude vypadat takto**:** **[tag="V.*"]** (V jako verbum + libovolný znak opakovaný libovolně mnohokrát).

5. Výsledky se zobrazují v tmavě šedé liště, ve druhém řádku (např. 1-20 z 100 znamená, že je InterCorp našel 100 výskytů). *Tokens* jsou jednotlivé slovní tvary, nikoliv počet výsledků. Konkordance je úhrn výskytů tvaru hledaného slova v kontextu jeho užití. Pomocí konkordancí můžeme zobrazovat různé doplňující údaje o nalezených výsledcích.

- zobrazení strukturních značek (**Konkordance/Zobrazit možnosti/Struktury**)
- zobrazení bibliografických údajů a identifikace konkordance (**Konkordance/Zobrazit možnosti/Reference**)
- zobrazení lemmatu a/nebo morfosyntaktické značky, pro klíčové slovo nebo všechna zobrazená slova - pro některé jazyky (**Konkordance/Zobrazit možnosti/Atributy**)
- "filtrování" výsledků na základě přítomnosti nebo nepřítomnosti zadaných výrazů v kontextu (**Konkordance/Zobraz filtr**)
- zobrazení celých vět (**Segment**) nebo řádků s hledaným výrazem ve středu (**Kwic**)
- zobrazení více jazyků ve sloupcích vedle sebe nebo v řádcích pod sebou (**Pohled: vertikální/horizontální**)
- zobrazení širšího kontextu (**Zobrazit kontext**)
- export konkordancí ve formátu tabulky (**Export: xls1, xls2**)

6. Pomocí InterCorpu tak můžete získat spoustu jazykových dat, která nejsou takto přehledně dostupná ani z internetu, ani z žádných příruček. Využít se dají například v překladatelství (vhodnost či nevhodnost překladu apod.), stejně jako při studiu odborné terminologie. Na rozdíl od internetu se také dá snadno a spolehlivě citovat.
InterCorp, projekt Ústavu Českého národního korpusu Filozofické fakulty Univerzity Karlovy, byl spuštěn roku 2005 v rámci a od té doby se vyvíjí. V současné době obsahuje téměř 550 milionů slovních tvarů v 27 jazycích.

7. Přejeme hodně štěstí, nápadů a zábavy při práci s InterCorpem!

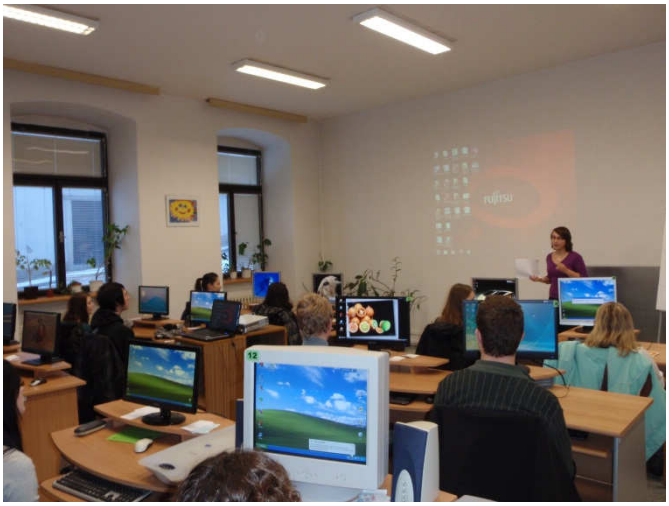
7.3 Morfologické značky pro francouzskou sekci InterCorpu

Morfologické značky v InterCorpu – sekce francouzštiny

ABR	abbreviation	(zkratka)
ADJ	adjective	(přídavné jméno)
ADV	adverb	(příslovce)
DET:ART	article	(člen)
DET:POS	possessive pronoun (ma, ta, ...)	(přivlastňovací zájmeno)
INT	interjection	(citoslovce)
KON	conjunction	(spojka)
NAM	proper name	(vlastní jméno)
NOM	noun	(podstatné jméno)
NUM	numeral	(číslovka)
PRO	pronoun	(zájmeno)
PRO:DEM	demonstrative pronoun	(ukazovací zájmeno)
PRO:IND	indefinite pronoun	(neurčité zájmeno)
PRO:PER	personal pronoun	(osobní zájmeno)
PRO:POS	possessive pronoun (mien, tien, ...)	(přivlastňovací zájmeno)
PRO:REL	relative pronoun	(vztažné zájmeno)
PRP	preposition	(předložka)
PRP:det	preposition plus article (au,du,aux,des)	(předložka se členem)
PUN	punctuation	(interpunkce)
PUN:cit	punctuation citation	(interpunkce – citace)
SENT	sentence tag	(tázací dovětek)
SYM	symbol	(symbol)
VER:cond	verb conditional	(sloveso, podmiňovací způsob)
VER:futu	verb futur	(sloveso, budoucí čas)
VER:impe	verb imperative	(sloveso, rozkazovací způsob)
VER:impf	verb imperfect	(sloveso, imperfektum)
VER:infi	verb infinitive	(sloveso, infinitiv)
VER:pper	verb past participle	(sloveso, příčestí minulé)
VER:ppre	verb present participle	(sloveso, příčestí přítomné)
VER:pres	verb present	(sloveso, přítomný čas)
VER:simp	verb simple past	(sloveso, minulý čas jednoduchý)
VER:subi	verb subjunctive imperfect	(sloveso, subjunktiv imperfekta)
VER:subp	verb subjunctive present	(sloveso, subjunktiv přítomný)

7.4 Fotografická dokumentace průběhu praktické části

Obr. 6: Praktická hodina v internetové studovně Městské knihovny v Třebíči (4. února 2013) – úvodní motivační řeč

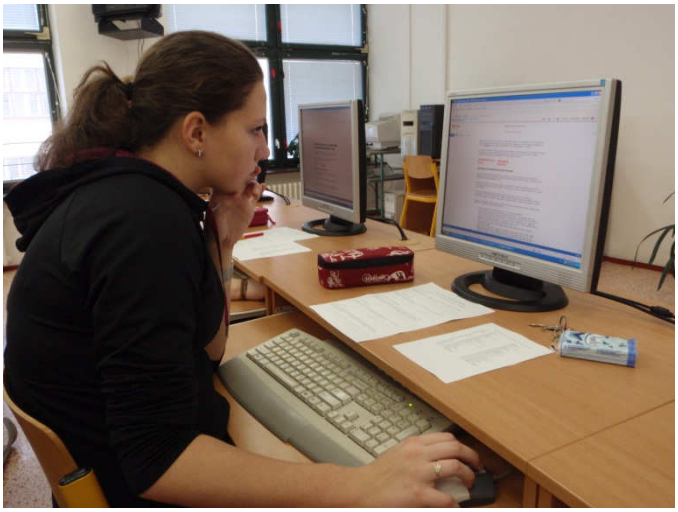


Obr. 7: Praktická hodina v internetové studovně Městské knihovny v Třebíči (4. února 2013) – práce studentů s korpusem

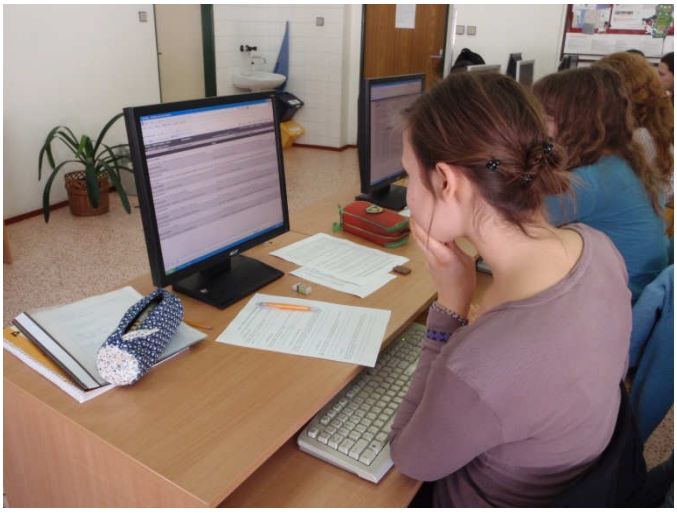
Obr. 8: Praktická hodina v internetové studovně Městské knihovny v Třebíči (4. února 2013) – kontrola úkolů

Obr. 9: Praktická hodina v počítačové učebně Gymnázia Třebíč (5. února 2013) – soustředění studentů při práci

Obr. 10: Praktická hodina v počítačové učebně Gymnázia Třebíč (5. února 2013) – práce s návodem



Obr. 11: Praktická hodina v počítačové učebně Gymnázia Třebíč (5. února 2013) – vyhledávání výsledků ve webovém rozhraní Park



7.5 Scénář motivačního videa

MOTIVAČNÍ VIDEO - INTERCORP

6 studentů sedí okolo stolu a usilovně pracují na nějakém úkolu.

TEREZA (*mírně naštvaně*): Achjo, jak mám sakra přeložit větu *She made me wait*.

KAMILA (*pokrčí rameny*): Já nevím.

ALEŠ (*sebejistě*): Já to přeložil jako *Dělala mě čekat*. Ale je to trochu zvláštní.

ALŽBĚTA *(vzhlédne od počítače)*: *(ironicky)* To teda je. Mám lepší nápad.

Všichni se podívají na Alžbětu.

ALŽBĚTA: Říká vám něco InterCorp?

Všichni kroutí hlavou.

ALŽBĚTA: To je paralelní korpus. ***(nechápavé výrazy)*** No, stručně řečeno je to taková databanka různých textů – románů, novin i třeba odborných knih. A pomocí něj se dají vyhledávat řešení problémů, jako je třeba ten náš problém make sb do sth.

MARKÉTA: No, je mi to trochu jasnější, ale jak se tam dostaneme? Nemusíme nikam jezdit, že ne?

ALŽBĚTA (*smích*): Ne, nikam jezdit fakt nemusíme. Stačí mít připojení na internet.

Střih (záběr na přihlašovací stránku)

ALŽBĚTA: Tak, a teď normálně zadám do prohlížeče www.korpus.cz/intercorp. Když chcete s InterCorpem pracovat naplno, musíte být registrovaní. Ale nebojte, ta registrace nic nestojí – je to pro každého z nás zdarma. Já tam teď zadám svoje údaje a zkusíme vyhledat to make sb wait.

Střih

(nechápavé výrazy se přemění ve výrazy chápavé)

ALEŠ: Teda, fakt super.

KAMILA: Fajn, takže z toho vyplývá, že *She made me wait* znamená *Nechala mě čekat*?

ALŽBĚTA: Přesně tak. Korpus vám ukáže, jak ten náš problém řešili jiní – tedy profesionální překladatelé. V našem případě to znamená, že ti najde úplně všechny vazby *make sb wait*, které náš korpus obsahuje. Pak už je na vás, jak s výsledky naložíte – kterou z variant zvolíte (protože jsme našli třeba i překlad přinutila mě čekat, což má trošku jiný význam).

MARKÉTA: A dá se využívat ještě nějak jinak?

ALŽBĚTA: Určitě! Já osobně InterCorp používám, když hledám, jak přeložit nějaké slovíčko. Třeba výraz pro všechny. Slovník mi řekne, že je to buď *for everybody*, nebo *for everyone*. Jak si ale má člověk vybrat? Já si to vyhledám v InterCorpu a pak se rozhodnu, co je vzhledem ke kontextu, v němž se používá, vhodnější, rozumíte?

VERONIKA: Já mám třeba občas problém při poslechu. Když posloucháme něco těžšího, tak třeba rozumím jen jedné větě s nějakým klíčovým slovem. Ráda bych si uměla odvodit, o čem je řeč. Neříkej mi, že s tímhle mi InterCorp taky dokáže pomoct.

ALŽBĚTA: Budeš se divit, ale taky to dokáže. Když máš větu ***(malou chvíli přemýšlí)****: I saw a seller/cellar*. Tak *seller* (psáno *seller*) znamená prodavač, *cellar* (psáno *cellar*) znamená sklep. A korpus ti vyhledá, které z těch slov se používá častěji.

TEREZA: Už to chápu, ty zjistíš, že třeba *cellar-sklep* se používá častěji, tím pádem je pravděpodobnější, že řeč je o sklepě?

(*chápavé výrazy*)

ALŽBĚTA: Bingo! A teď si můžete zkusit vyhledávat cokoli, co vás napadne – InterCorp je schopen vyhledat jakoukoli jazykovou hříčku nebo problém, tedy pokud je k nalezení v textech, které korpus obsahuje. Proto ho ale v Ústavu Českého národního korpusu neustále rozšiřují, abychom jednoho dne našli opravdu téměř všechno.

(všichni se vrhají k počítači a překotně do něj ťukají)

ALŽBĚTA (*obrací se do kamery*): I já jsem si dřív myslela, že InterCorp můžou používat jen lingvisté, lexikografové, překladatelé a tak. Ale zjistila jsem, že ten stejný InterCorp může být užitečný i „obyčejným“ lidem, třeba středoškolákům, jako jsme my. Jak už zaznělo, opravdu nemusíte nikam jezdit. Ale můžete to rozjet na www.korpus.cz/intercorp!

7.6 Odkaz na motivační video

Motivační video paralelního korpusu InterCorp je dostupné z URL: http://www.youtube.com/watch?v=TCeWW5u6v-Y [Cit. 9.3.2012]. Bude také umístěno na stránkách Ústavu Českého národního korpusu.

Obr. 12: Úvodní scéna motivačního videa