	Pelouškova, H. - Káňa, T.: Paralelní korpus jako zdroj autentického jazykového materiálu pro výzkum i výuku jazyků. In: *Cizí jazyky*, Plzeň: Fraus, 2002. s. 43-45. Rychlý, P.: *Korpusové manažery a jejich efektivní implementace*. Disertační práce. Brno 2000. Sedláček, R. – Smrž, P.: A New Czech Morphological Analyser ajka. In: *Proceedings of the 4th International Conference TSD.* Berlin: Springer-Verlag, 2001. s. 100-107. Šulc, M.: *Korpusová lingvistika*. Praha: Nakladatelství Karolinum, 1999. Ungermannová, M.: *Manuál pro značkování a desambiguaci slovních tvarů v korpusu DESAM.* Interní materiál FI MU. Brno 1999.

Česko-německý paralelní korpus a jeho využití

Tomáš Káňa, Hana Pelouškova

1. Co je jazykový korpus

V posledním desetiletí se v souvislosti se zkoumáním i výukou jazyků i u nás stále častěji objevuje pojem **jazykový korpus**.

Nestor české korpusové lingvistiky, prof. František Čermák jej definuje jako „vnitřně strukturovaný, unifikovaný a obvykle i oindexovaný a ucelený rozsáhlý soubor elektronicky uložených a zpracovávaných jazykových dat většinou v textové podobě, organizovaný se zřetelem k využití pro určitý cíl, vůči němuž je pak také považován za reprezentativní.“ (Čermák, 1995).

Korpusová lingvistika, která je v současné době bezesporu podstatnou složkou lingvistiky komputační, se zabývá budováním korpusů, vytvářením jejich nástrojů a vytěžováním korpusů s cílem dokonalého poznání funkcí a struktury zkoumaného jazyka, popř. jazyků. Autentických jazykových dat elektronických korpusů využívá korpusová lexikografie (prakticky všechny nově vznikající slovníky Velké Británie jsou vybudovány na základě korpusových dat). Cenné informace poskytují korpusy rovněž pragmatice, logice, psychologii a psycholingvistice, sociologii a sociolingvistice, neurolingvistice, literární vědě, pedagogice a mnoha dalším disciplínám.

Doslova novou epochu efektivního mnohostranného poznávání jazyka, lidského myšlení a kultury otevřely elektronické jazykové korpusy ve srovnání s předchozími, pracně sestavovanými korpusy manuálními (kartotékami excerpt) zejména proto, že poskytují data:

- typická, nenáhodná a věrná
- aktuální
- neselektivní a objektivní
- dostatečná
- s pomocí počítače rychle přístupná a dále zpracovatelná

(srov. ČNK, 2000)

2. Existující korpusy

Počítače vstupují do lingvistiky počátkem 60. let dvacátého století a od této doby též vznikají jazykové korpusy. Prvními významnými a pečlivě sestavenými korpusy byly Brown Corpus americké angličtiny (1964) a Lancaster-Oslo/Bergen Corpus britské angličtiny (1978), oba v rozsahu zhruba jednoho milionu textových slov.
Od konce 80.let a zejména v devadesátých letech již vznikají rozsáhlé, stamilionové korpusy (Bank of English, British National Corpus, německý Mannheimský korpus…), jakož i různé korpusy specializované.
V současné době existují nebo se připravují velké korpusy mnoha jazyků. Existence národních korpusů a s nimi spojených kvalitních lexikografických projektů je podmínkou multilaterální spolupráce v rámci Evropy.

Z korpusů slovanských jazyků nejdále pokročil Český národní korpus (ČNK), jenž vzniká od roku 1994 na půdě Filozofické fakulty Univerzity Karlovy, v Ústavu ČNK, vedeném profesorem F. Čermákem.
ČNK zahrnuje v současné době asi 120 milionový synchronní korpus SYN 2000, téměř dvoumilionový diachronní korpus DIAKORP, mluvený synchronní korpus ORAL-PMK (Pražský mluvený korpus) v rozsahu 700 000 textových slov a mluvený korpus BMK (Brněnský mluvený korpus) v rozsahu 500 000 textových slov.
Informace o ČNK a přístupu k němu najdete na internetové stránce:
http://ucnk.ff.cuni.cz

Vznikající i existující korpusy je možné dělit podle nejrůznějších kritérií i specifických cílů, pro něž byly sestaveny. Rozlišujeme tedy korpusy malé a velké, synchronní a diachronní, mluvené (příp. nářeční) a psané, studijní (např. psané studenty cizích jazyků), korpusy sestavené ze specifických typů textů (literární, publicistické…), korpusy jednojazyčné a paralelní.

3. Paralelní korpusy

Jsou mnohem složitější jak technicky, tak i na vlastní tvorbu (viz dále), skýtají však dosud nebývalé možnosti studia dvou a více jazyků v aktuálním kontrastu.
Podívejme se nejdříve na otázku paralelních textů, tedy na základní kámen elektronických paralelních korpusů:

Paralelními texty rozumějme takové texty, jež dvěma nebo i více různými jazykovými kódy zachycují stejný textový invariant. Současné paralelní korpusy si vystačí s „dobrými překlady“. (Hodnocení překladů nechme na tomto místě stranou a zahrňme mezi ně všechny překlady, které jsou určeny širšímu publiku.) Bylo by jistě zajímavé porovnávat dvě autentické varianty jednoho textu, leč zdroje takových textů jsou velmi omezené, proto lze pochybovat o dostatečné reprezentativnosti dat korpusu, jenž by tyto originální texty obsahoval.

Existující i právě vznikající paralelní korpusy si můžeme proto představit jako rozsáhlou sbírku překladů a originálů těchto překladů (o výběru textů se zmíníme později).

Z výše uvedeného lze vytušit, že paralelní korpus nutně naráží na problémy, od kterých je jeho jednojazyčný kolega oproštěn. Tyto problémy jsou v zásadě trojí povahy: **obecné, lidské a technické**.

Obecně trpí paralelní korpus relativním nedostatkem materiálu. Nepřekládá se všechno a ne všechny překlady jsou dostupné. Dále se musí paralelní korpus vzdát obrovského potenciálního zdroje dat - tím jsou mluvené texty. (Přetlumočené texty – byť i tlumočnickými špičkami - neodpovídají přirozenému jazyku vzhledem k vysokému počtu interferenčních chyb.)

Druhým problémem je skutečnost, že paralely k výchozím textům (originálům) vytvářejí překladatelé různých kvalit. Výsledná data dotazů korpusového manažeru proto mohou být zkreslena, pokud korpus není dostatečně reprezentativní. Navíc musí uživatel překladatelovy poklesky eliminovat sám.

Příčinou třetího základního problému paralelních korpusů je nedokonalá technika. Zní to sice v současné přetechnizované době poněkud paradoxně, ale je třeba si uvědomit, že i ten nejsložitější počítač má stále ještě daleko k umělé inteligenci. Právě z toho důvodu, že stroje neznají sémantiku, nedovednou odhadnout nestrukturovatelné pragmatické složky v jazyce a bez předchozího zadání dovedou jen omezeně vytvářet nepravidelné kombinace, jsou strojové překlady doposud poněkud nemotorné a ze stejného důvodu se všechny paralelní korpusy potýkají s problematikou tzv. zarovnání. Zjednodušeně lze říci, že v současnosti pracují paralelní korpusy pouze na principu přiřazování textových sekvencí k zadanému dotazu (viz dále).

I přes potíže a nedostatky, s nimiž musíme při sestavování a užívání paralelních korpusů počítat i do budoucna, nelze elektronickým korpusům odepřít revolučnost, již svým rozsahem a širokým záběrem vnášejí do mnoha vědních i aplikovaných disciplin.

Paralelní korpus pracuje podobně jako jednojazyčný a lze ho tak i využívat. Navíc má uživatel možnost k požadovanému dotazu vyvolat příslušný jev ve druhém jazyce.
Na základě velkého množství dat je tak možné zkoumat vzájemné shody a odchylky jazykových systémů, a to ve všech jazykových plánech, odhalovat zákonitosti překladatelského procesu, zkoumat společné a rozdílné znaky textových útvarů (téměř) všech stylových rovin nebo ho můžeme použít jako pomůcku při výuce některého ze zastoupených jazyků.
Elektronické paralelní korpusy tak nabízejí autentický jazykový materiál lingvistům, lexikografům, překladatelům, literátům, učitelům, studentům, redaktorům a všem, kdo z jakýchkoliv pohnutek mají zájem o jazyky. Jelikož současná technika umožňuje také jednoduchou konverzi do Braillova písma, nabízejí elektronické korpusy své služby i nevidomým a speciálním pedagogům.

Existuje několik vícejazyčných korpusů - např. CRATER pro angličtinu, francouzštinu a španělštinu nebo paralelní korpus biblických textů na University of Maryland.
Mnohem více paralelních korpusů je jen dvoujazyčných. Z pochopitelných důvodů převládají korpusy s angličtinou jako jednou z paralel. Běžným internetovým vyhledávačem nalezneme např. paralelní korpus anglicko-norský, francouzský, turecký, slovinský, portugalský, a dokonce i čínský. Dostupné jsou korpusy anglicko-německé.
Čeština je zastoupena v několika paralelních korpusech. Za zmínku stojí multiligvální projekty TELRI (Platónova ústava v 17 jazycích), Multext East (román G. Orwella 1984 v 23 jazycích) a dva korpusy česko–anglické (česko-anglický korpus PDT –1.0 na Matematicko-fyzikální fakultě UK Praha a tzv. „Kačenka“ na FF MU v Brně).
Zajímavé je, že podle internetové rešerše se zdá, že paralelní korpus většinou iniciuje zástupce minoritního jazyka.

4. Česko-německý paralelní korpus

Myšlenka vytvořit česko-německý paralelní korpus se zrodila v roce 2000 na katedře německého jazyka a literatury Pedagogické fakulty Masarykovy univerzity v Brně. Vedly k ní rešerše potvrzující absenci tohoto instrumentu a samozřejmě jeho nezbytnost pro kontrastivní výzkum i výuku obou jazyků. Po zajištění základních technických podmínek začalo budování korpusu v červnu 2001.

Tato iniciativa se setkala s příznivým ohlasem v Ústavu Českého národního korpusu, který ji od samého počátku podporuje nejen cennými radami, ale i finančně, a zejména poskytováním textů, k nimž lze získat německé paralely.
Na projektu spolupracují rovněž Fakulta informatiky MU, Fakulta informačních technologií VUT a v rámci bilaterálního programu Aktion Österreich-Tschechische Republik se na rozšiřování korpusu dva roky podílel i Institut für Slawistik na Vídeňské univerzitě. Většinu technických prací (skenování, opravy skenů, čištění textů, zarovnávání textů, vytváření metainformací aj.) vykonávají studenti německého jazyka na Pedagogické fakultě MU.
Od roku 2005 se korpus buduje v rámci projektu InterCorp, který je součástí VZ Český národní korpus a korpusy dalších jazyků (MSM 00216208239). V rámci tohoto projektu vzniká více než 20 paralelních korpusů s češtinou.

Parametry
Budování korpusu směřuje k dosažení parametrů, které byly stanoveny s ohledem na cílové skupiny uživatelů (obecně filologové zabývající se jazyky ze synchronního hlediska – lingvisté, učitelé, studenti, překladatelé):

- **Synchronní –** Do korpusu jsou zařazovány pouze texty publikované po roce 1920. Tato současnosti poměrně vzdálená hranice umožní zpřístupnit širšímu okruhu uživatelů subkorpus textů, které již nejsou vázány autorskými právy. Samozřejmě preferujeme texty co nejsoučasnější, poskytující objektivní pohled na aktuální stav obou jazyků.

- **Dvojjazyčný –** Do korpusu jsou zařazovány výhradně české či německé originální texty a jejich překlady. Je vyloučena přítomnost dalšího jazyka. Cílový poměr českých a německých originálních textů by měl být 1:1. Aktuálně mírně převažují české originální texty (64%).
- **Obecný –** Z evidentních důvodů vzniká pouze korpus psaného jazyka. Proporce základních stylistických kategorií byly vzhledem k dostupnosti textů stanoveny v poměru: 50% beletrie, 25% publicistika, 25% odborné texty. V současnosti je zařazeno do korpusu 63% beletristických (87% tvoří romány a povídky), 17% publicistických (97% tvoří informativní texty) a 20% odborných textů z různých oblastí (převážně společenských věd, zastoupení přírodních věd je marginální).
- **Dostatečně rozsáhlý –** Zatím je zařazeno do korpusu 184 textů s více než 3,6 mil.

slov v české části. Velké množství materiálu je ve stádiu zpracovávání. Korpus se bude dále rozšiřovat.

- **Zarovnaný –** Zarovnání (alignment) je jednou z nejpracnějších procedur při tvorbě paralelního korpusu. Automatické zarovnávání pomocí dostupných alignerů se příliš neosvědčilo, neboť vykazovalo značnou chybovost (texty se „rozjížděly"). Proto jsou texty zarovnávány prostřednictvím programu ***Checker*** (makro programu MS Word), který rozsegmentuje texty na odstavce a přiřadí k sobě paralely do dvou sloupců tabulky. Bohužel ne všechny překlady respektují důsledné členění na odstavce. Chyby v zarovnání jsou odstraňovány manuálně pomocí makra ***WordFisher*** (umožňuje snadnou manipulaci s buňkami tabulek, jejich přesouvání, slučování, segmentování a mazání). Většina zařazených textů byla dále manuálně zarovnána na úrovni věty (příliš dlouhé odstavce některých textů výrazně snižují efektivitu práce s korpusem). Texty vhodné k zařazení do multilingválního korpusu ***InterCorp*** bude nutné upravit pro zarovnání v programu ***ParaConc***.
- **Morfologicky anotovaný –** Česká paralela korpusu byla označkována morfologickým analyzátorem „***ajka***" (vytvořen na Fakultě informatiky MU). Za velmi uspokojivé můžeme označit vyhledávání podle lemmatu (základního tvaru), méně uspokojivých výsledků dosahuje vyhledávání podle tagu (morfologické charakteristiky). O manuální desambiguaci korpusu či jeho části zatím neuvažujeme. Německá paralela byla označkována systémem ***TreeTagger*** (vyvinut na univerzitě ve Stuttgartu), který umožňuje vyhledávat podle lemmat a slovních druhů.
- **Opatřený „metainformacemi",** tj. vnější anotací textů – Informace o původu a zpracování textů jsou provázejí každou paralelu textu a obsahují tyto položky: *název dokumentu, název textu, hrubou a jemnou stylistickou charakteristiku textu, zdroj textu, médium, na němž byl text získán, jméno a pohlaví autora, u překladu i jméno a pohlaví překladatele, jazyk textu, jazyk originálu, rok vydání nebo zveřejnění textu, technické údaje o zpracování textu.* Tyto informace je možné vyvolat ke každému konkordančnímu řádku a bude možné podle nich vytvářet subkorpusy.

Veškeré informace o zařazených a zpracovávaných textech zahrnuje rovněž katalog textů, který je průběžně aktualizován. Všechny texty jsou archivovány na CD.

V současné době je korpus umístěn a spravován na serveru Fakulty informatiky MU a jeho uživatelským prostředím je korpusový manažer Bonito (autorem je Pavel Rychlý z FI MU).

Tento manažer využívá i ČNK. Vzhledem k tomu, že je většina textů chráněna autorskými právy, není korpus běžně přístupný. Verze korpusu z počátku roku 2005 existuje na CD.

Práce s ČNPK

Fakticky i technicky se ČNPK skládá ze dvou samostatných korpusů, které jsou vzájemně spřaženy. V zásadě se práce s ČNPK neliší od práce s ČNK. V pracovní ploše Bonita však nejdřív vybereme paralelu, v níž chceme primárně pracovat:
v pravém horním rohu: **cnpkcz** – česká část korpusu; **cnpkde** – německá část.
Dál je práce shodná s prací v korpusu DESAM a téměř shodná s prací s Českým národním korpusem. (Na některé odchylky upozorníme později.)

- **Dotaz**

Nejběžnější dotazy jsou :
Word (pozice - slovní tvar, kombinace písmen, otazník, vykřičník, číslo atd.), **Lemma** a **Tag**.

Word

Chceme zjistit, zda česká výpůjčka německého „Lump“, tedy lump, odpovídá vždy německému „Lump“.

Nový dotaz/new Query: lump

V online verzi je celkem 12x „lump“. Informace o textu, z něhož konkordanční řádek pochází získáme buď stisknutím pravého tlačítka myši (informace se objeví v dolním okně), nebo v nabídce *zobrazení/view* a dále *reference/ references*, kde jsou veškeré informace o textu (viz metainformace).

Dotaz můžeme rozšířit na vyvolání veškerých slov, která začínají na „lump“:

Nový dotaz/new Query: lump.* („.“ - „tečka“ = jakýkoliv znak; „*“ - „hvězdička“ = jakýkoliv počet znaků)

V on-line verzi je celkem 46x slov začínajících na „lump“. Konkordance lze vytřídit tak, aby každý tvar zůstal v okně pouze jednou:

Nabídka*: konkordance/ concordance → obecné třídění/ generic sort → vynechat duplicity/ make unique*

Tímto způsobem zredukováno na 11:

```
#------------------------------------------------------------
## Corpus  : cnpkcz
# Label   :
# Query   : lump.*
# Expanded: "lump.*"
# Steps:
#   > Query   : "lump.*"
## Size    : 11
# Context : 5 Positions to left, 5 Positions to right
## User    : kana
# Date    : Mon, 07 Nov 2005 11:32AM
##------------------------------------------------------------

1:              peníze . Pak zmizel , < lump>  . Nevěděl si asi rady
2:  dneska toho hodnýho od toho < lumpa>  , zejména dnes , v
3:                  tebe . Podívej , to < lumpačisko>  , si stouplo za křoví
4:              lidi , aby sloužili této < lumpárně>  - nikdy by se byl
5:            , že mají ještě nějakou < lumpárnu>  za lubem . Nepřítele v
6:                      : " Čichni si , < lumpe>  ! " Švejk si čichl
7:                 zdá se mi ... ňákej < lumpeproletariját>  !  , ale vy
8:        plejboje . Bočár zakročil , < lumpi>  tasili zbraň a - abych
9:                    a povídá : , Vy < lumpové>  , padouchově , neřádi ,
10:  Němci pomáhali Čechům vyřezat pár < lumpů>  i u nás . Jaksch
11:          nebo mít kázání k těm < lumpům>  . V tomhle případě musí
```

Vyvolat lze i všechna slova, v nichž se objevuje kombinace písmen „lump“.

Nový dotaz/new Query: .*lump.*

A takových nalezneme 49, po vynechání duplicit 14:

```
# Query   : .*lump.*

1:  na horskou túru vzal zbytečný < klumprt>  navíc . Byli mladí .
2:          peníze . Pak zmizel , < lump>  . Nevěděl si asi rady
3:    dneska toho hodnýho od toho < lumpa>  , zejména dnes , v
4:            tebe . Podívej , to < lumpačisko>  , si stouplo za křoví
5:       lidi , aby sloužili této < lumpárně>  - nikdy by se byl
6:        , že mají ještě nějakou < lumpárnu>  za lubem . Nepřítele v
7:                : " Čichni si , < lumpe>  ! " Švejk si čichl
8:            zdá se mi ... ňákej < lumpeproletariját>  !  , ale vy
9:    plejboje . Bočár zakročil , < lumpi>  tasili zbraň a - abych
10:                a povídá : , Vy < lumpové>  , padouchově , neřádi ,
11:  Němci pomáhali Čechům vyřezat pár < lumpů>  i u nás . Jaksch
12:              nebo mít kázání k těm < lumpům>  . V tomhle případě musí
13:              je , aby bylo písmeno < yxilump>  správně vyslovováno . " "
```

Paralelní text

Pod nabídkou (*zobrazení/ view*) je i příkaz *paralelní text/ paralel text*. Ten se dá vyvolat i jednodušeji - označením řádku a následným stisknutím tlačítka F9. Odpovídající pasáž

v druhé paralele (zde tedy německé) se objeví v novém okně. (Tento krok však v současné době nelze uskutečnit, pokud jsme odfiltrovali duplicitní tvary!)

Z výsledků vidíme, že české „lump“ většinou odpovídá německému „Lump“.

KWIC je zvýrazněno pouze v primární paralele. V ukázce je druhé vyznačení provedeno ručně:

cnpkcz: 393363--393367
Pak zmizel , **lump** .
cnpkde: 430626--430632
Dann verschwand er , der **Lump** .
---
cnpkcz: 973052--973086
Ten měl syna Bohuslava , a kam přišel prodávat prasata , nikdo od něho nic nekoupil a každý říkal : , To je syn toho probodnutýho , to bude asi také pěknej **lump** . ´
cnpkde: 1097072--1097117
Der hatte einen Sohn namens Bohuslav , und wenn der seine Schweine verkaufen kam , wollt niemand was von ihm kaufen , und jeder hat gesagt : , Das ist der Sohn von diesem Erstochenen . Das wird gewiß auch ein feiner **Lump** sein . '
---
cnpkcz: 994849--994864
Z těch **lumpů** by ani Radecký , ani ten váš princ Eugen Savojský nevychovali vojáky .
cnpkde: 1122411--1122423
Aus diesen **Lumpen** hätte nicht einmal Radetzky oder Prinz Eugen Soldaten gemacht
---
cnpkcz: 3449524--3449533
Ty seš , zdá se mi ... ňákej **lumpe**proletariját !
cnpkde: 3975931--3975939
Ich glaub , du bist ... ein Lumpenproletariat !
--

Nacházíme však i jiná řešení:

cnpkcz: 1713850--1713886
Vše sprosté sproštěno je execírky ? To že je válka ? Spíš v tom vidím mír . Muž dobrý v boj jde , **lump si může hopsat** . Mřít nemusí . Neb on to umí popsat .
cnpkde: 1925233--1925269
Nicht exerzieren müssen die Gemeinen ? Ist es ein Krieg ? Ich denk ' , es ist der Friede . Die Bessern gehen und **die Schlechtern bleiben** . Nicht sterben müssen sie . Sie können schreiben .
---
cnpkcz: 1753430--1753496
Ať pod vodou či na zemi , **je lump** , kdo zůstal v zázemí . Však že jsem všude zvyk si rvát se , mně není třeba k frontě dát se , já v boji cvik a říznost mám a v zázemí tu vytrvám , já rvu se čile den co den tím rváčů šik je rozhořčen . Jen do toho ! Vždyť já jsem Němec !
cnpkde: 1969525--1969595
Ob unter See , ob in der Luft , wen Kampf nicht freut , der ist ein **Schuft** . Doch weil das Schuften ich gewohnt , so schuft ' ich nicht bloß an der Front , ich kämpf auch schneidig und gewandt und halte durch im Hinterland , ich schufte früh , ich schufte spat , die Schufte das erbittert hat . Nur feste druff ! Ich bin ein Deutscher !

A zcela náhodou objevujeme i jiné zajímavé vztahy mezi německými výpůjčkami a jejich „ekvivalenty“ v němčině:

cnpkcz: 3491447--3491476
Nervózně strhávali kabel s větví jabloně a vztekle přehazovali na prsou automaty jako někdo , kdo ve vedru poznal , že si na horskou túru vzal zbytečný **klumprt** navíc .
cnpkde: 4027684--4027729
Nervös zerrten sie die Kabel von den Ästen eines Apfelbaumes und warfen zornig die Maschinenpistolen auf ihrer Brust von einer Seite auf die andere , wie jemand , der in der Hitze erkennt , dass er für eine Bergtour viel zu viel überflüssiges **Zeug** mitschleppt .
--
cnpkcz: 997942--997961
Ted ' pláčeš , a až se odtud vrátíš do **cimry** , zas budeš stejně takový lump jako předtím .
cnpkde: 1126010--1126034
Jetzt weinst du , und bis du von hier in die **Zelle** zurückkehrst , wirst du wieder grad so ein Lump sein wie vorher.

Vyvolat lze i některé pozice – např. „*čárku*”, „*vykřičník*”, „*středník*”.

Dospěli jsme k zajímavému zjištění, že v německé paralele ČNPK je o 7.000 více čárek, asi o tisíc víc vykřičníků (tj. asi o 6%), avšak zhruba o 2.000 méně středníků (tj. asi o 1/3).

Příklad ad hoc vybraných divergencí:

cnpkde:
" Schnell , Abe , lauf , lauf **!** " flüsterte Li .
cnpkcz:
" Honem , Abe , běž , běž **,** " šeptala Li .
--
cnpkcz: 318581--318606
Tak poslouchej , Dášo **;** když budeš pěkně sedět , povím ti pohádku . O čem pohádku ? No , třeba pohádku o psím ocásku .
cnpkde: 344146--344179
Also hör mal , Dascha **,** wenn du ein Weilchen fein stillsitzen willst , werde ich dir ein Märchen erzählen . Ein Märchen **,** wovon ? Na , etwa das Märchen vom Hundeschwänzchen .
cnpkde:
--
" Frau Oberköchin ! Frau Oberköchin **!** " mahnte der Oberkellner , der ihren Blick aufgefangen hatte .
cnpkcz: 1332141--1332160
" Paní vrchní kuchařko ! Paní vrchní kuchařko **,** " napomenul ji vrchní číšník , který zachytil její pohled .

Výše popsaným způsobem můžeme hledat v obou paralelách.

Lemma

Obě paralely korpusu jsou automaticky lemmatizovány. Syntax dotazu je poněkud složitější, ale stejná jako v ČNK:

[lemma="lump"]

#----------

```
# Corpus  : cnpkcz
# Query   : [lemma="lump"]
#------------------------------------------------------------
1. pomáhali Čechům vyřezat pár <lumpů> i u nás .
2.                          . Pak zmizel , <lump> . Nevěděl si asi
3.             toho hodnýho od toho <lumpa> , zejména dnes ,
4.                           " Čichni si , <lumpe> ! " Švejk si
5.                    mít kázání k těm <lumpům> . V tomhle případě
6.       držkovala jako na ňákýho <lumpa> , když tady před
7.                     , že dívka miluje <lumpa> a halunka , i
8.                       ve slovníku . " <Lumpi> , " řekl tatínek
#------------------------------------------------------------
# Corpus  : cnpkde
# Label   :
# Query   : [lemma="Lump"]
# Expanded: [lemma="Lump"]
#------------------------------------------------------------
1.                  helfen , ein paar <Lumpen> bei uns zu schlagen
2. seiner Schandschnauze , der <Lump> " , brummt er
3.                           tu ich , ihr <Lumpen> . Ich verlier Zeit
4.                        " Gesteh , du <Lump> , daß do nur
5.                  Hause gehn , ihr <Lumpen> , es ist schon
```

Chybovost stoupá přímo úměrně se složitostí morfologie lexému.

Česká paralela

Relativně spolehlivě je lemmatizována oblast jmen v české paralele. Jelikož korpus není ručně desambiguován, naprosto nejsou rozlišeny homonymní tvary v rámci jednoho slovního druhu, navíc není možné zohlednit krácení kmenové samohlásky u některých slov.
Obecně lze říct, že česká část je celkem uspokojivě lemmatizována. Chyby napříč slovními druhy nejsou příliš časté:

```
# Query   : [lemma="puk"]
1. Heine je a zůstane největší německej básník , dybyste < puk> - Pozdravil sem sekčního šéfa , on pozdravil
2.          Snažil jsem se na mokrých kalhotách vymačkat < puky> .
```

Problematická je lematizace, pokud je některý tvar homonymní s výrazem pro češtinu atypickým (např. předložka „*de*" v románských jazycích a české hovorové „*von de*" = „*on jde*")

```
# Korpus  : cnpkcz
# Dotaz   : [lemma="jít"]
1:              opatření . Eventuální vznikající poplatky < jdou>  k tíži nájemce . Žádáme Vás , abyste nám
2:              vévodství . Nejúspěšněji si vedl , pokud < jde>  o Wittelsbachy , Ludvíka Římana a Otu Líného
3: Pantaleonem Barbem , tajným písařem Lorenzem < de>  Monacis a admirálem Janem Barbadikem .
4:          obecenstvu dělat chuť na válku a na list , to < de>  jedno s druhým . Moc důležitý sou při tom
5:   hovoří o 100000 mužích . Ve skutečnosti mohlo < jít>  asi tak o 9000 až 10000 mužů , zatímco
6:            V roce 1408 se vydal na pouť do Santiaga < de>  Compostela a do Irska a jednomu notáři
7:                     Konec . Hnedka člověk ví , vo co < de>  . Pak to vobskakuje , mluví o Talleyrandovi
8:               hrála ve sporu první housle . Přitom mu < šlo>  o spoluúčast knížat a o likvidaci dominantního
9:      Všech svatých veřejně oznámil , že do Savony < nepůjde>  , odvolával se mimo jiné i na radu uherského
```

Přesto lze funkci lemma spolehlivě použít a ručně dotřídit (např. sloveso „*mít*"):

```
# Dotaz   : [lemma="mít"]
1:              Pavlendo , vůči ní jsem s . Strýčínku , < maj>  esenbáci děti ? Pročpak ne , Katko . Jsou
4:          téměř vždy ( např . " icebreakers " ) , jiné < mají>  své místo jen na některých kurzech ( např
5:    záležitostí člověka , nýbrž hlavně praktickou ; < majíc>  autoritu božskou , určuje mravnost . Na
6: České republiky a Spolkové republiky Německo < majíce>  na paměti Smlouvu mezi Českou a Slovenskou
20:  nesouhlasil s maminkou , která ji chtěla vrátit . < Mělas>  mu ji dát hned , podpořil mě i Edvín ,
27:             se víc peněz než za rok selské roboty . < Mít>  ještě kus roviny a postavit stáje tam dole
30:  propuštění z armády Jaksch podporovat odmítal . < Nemá>  prý v úmyslu , ujišťoval Beneše , " fungovat
32:       Dala bych rozsvítit . Proč se tak zhasíná ? < Nemaj>  dost proudu . Kdo ? Oni , odpověděla Alice
34:      Bernis cítil se být vždy tak prozíravým , že < nemaje>  materiálu proti obviněnému , neznaje ,
66:     řečiště . " Oh , oh ! " zajásal . " Massa Bob < nemíti>  pušku . Větev býti jeho ručnicí , jeho
```

Německá paralela

I v německé paralele můžeme použít dotaz lemma. Vzhledem k tomu, že v němčině je homonymie častější než v češtině, je i chybovost mnohem větší:

```
# Korpus  : cnpkde
# Dotaz   : [lemma="Band"]

1:  Williams , der Schauspieler Otto Sander und die < Band>  Rammstein tragen seine Mode . Sein Erfolgsrezept
2:          Lager des Leidenden mit den herrlichsten < Bänden>  der Weltliteratur umgeben ließ . Bald bekam
3:  Trommelfeuer aus und stürzte sich auf das lange < Band>  der Straße , als ob er es aufwickelte .
4:              knöpfte sie das Kleid auf , knüpfte die < Bänder>  los , legte die Wäsche sachte hin und streifte
5:       seines Werkes erschwert : Es umfaßt gut 350 < Bände>  , größtenteils planmäßig mitstenografierte
6:             Heften bereits neunzehn oder volle drei < Bände>  eines vierteiligen Gesamtwerkes und sie
7:            berühmte Klosterbibliothek mit 85 . 000 < Bänden>  und 1200 Handschriften , der prunkvolle
```

Podobně jako v české paralele, ani německé se nevyhýbá problém homonymie s cizími slovy:

```
# Dotaz   : [lemma="lesen"]
1:    polnisches Land ist " . Aus dieser Antwort < las>  Allen heraus , daß Molotov zwar die Zuständigkeit
2:  kleinen , gelben Zettel entgegen . " Nein " , < lese>  ich von ihnen ab . Der Mann scheint zusammenzuzucken
3:  Morris liebt Metropolen . New York , London , < Las>  Vegas . Sie inspirieren die Künstlerin
4:  moderierte . Auf die Couch fläzt er sich , um zu < lesen>  . Momentan eine Biografie über Muhammed
5:       Propeller ... " " Und was steht hier ? " " Das < liest>  du ein andermal , du kannst doch lesen
6:       Menge Pulver auf Vorrat . In wenigen Tagen < las>  Prokop bereits geläufig alle Rezepte des
7:        zurückgekehrt sind ! Haben Sie mein Inserat < gelesen>  ? " " Ja " , antwortete Prokop , " was
```

Paralelní text

Ke každému řádku je pochopitelně možné vyvolat odpovídající pasáž v druhém jazyce:

A tu auditor Bernis cítil se být vždy tak prozíravým , že **nemaje** materiálu proti obviněnému ,
cnpkde: 1128757--1128814
Auditor Bernfis glaubte so überaus scharfsinnig zu sein , daß er , ohne Material gegen den Angeklagten zu besitzen ,
--

Über den Mangel an Imageträgern braucht er sich nicht zu beklagen . Stars wie Boris Becker , der Sänger Robbie Williams , der Schauspieler Otto Sander und die **Band** Rammstein tragen seine Mode .
cnpkcz: 309615--309647
Na nedostatek propagátorů si tedy mladý návrhář rozhodně stěžovat nemůže . Jeho módu nosí takové hvězdy jako kupříkladu tenista Boris Becker , zpěvák Robbie Williams , herec Otto Sander a **kapela** Rammstein .
--
Langsam knöpfte sie das Kleid auf , knüpfte die **Bänder** los , legte die Wäsche sachte hin und streifte die Strümpfe von den kräftigen , schönen Beinen ; nachdenklich blickte sie zu Boden und spielte dabei , wie es

Kinder zu tun pflegen , mit den langen , makellosen Zehen .
cnpkcz: 454938--454974
Nespěchajíc rozepíná knoflíčky a rozvazuje **tkanice** , tichounce skládá prádlo , smeká zvolna punčochy se silných a dokonalých nohou ; zamyslí se , hledí k zemi a hraje si po dětsku dlouhými , bezúhonnými prsty chodidel ;
atd.

Tag

Česká paralela

Česká paralela je morfologicky označkována analyzátorem „***ajka***“, má tedy stejný systém značek jako DESAM. Tento se však liší od ČNK.[1]

Vzhledem k tomu, že - na rozdíl od DESAMU – není ČNPK ručně zjednoznačněn, přesnost výsledku dotazu podle tagu kolísá. V optimálním případě dosahuje chybovosti kolem 10%, v horších případech i nad 50%.

Dotaz: [tag="k1gInSc7.*"], tj. substantiva neživotná, v singuláru a instrumentálu.

Chybovost asi 10%.

Korpus : cnpkcz
Dotaz : [tag="k1gInSc7.*"]
1: . Jejich <zájmem> bylo společně posílit
2:Amerikáni přece nejezdí osobním <vlakem> : jedou si v rychlíku ,
3: Marx a Engels nakonec - pod <vlivem> Anglie - uznali proti
4: a která je bohužel nutná <vzhledem> k poměrně velkému zájmu o
5: jak se pohybovaly ty holky šťastným <smíchem> , a kolem mne už bylo
6:rovnou do kuchyní a odkud jsem vyjížděl s < vozíkem> a na tom vozíku
7: o pořadateli , najdete je pod příslušným < projektem> nebo kontaktujte
9: napětí mezi etickou normou a vlastním < zákonem> určité věcné
10: je také dána možnost zužitkovat stejným < způsobem> moč a močůvku .
11: Škarohlíd : Ani tato vlastnost ho < celkem> nestaví nad většinu
12: k jádru věci . Opět zamával < kloboukem> , otřel se kapesníkem ,
13: Začíná brněnským <mírem> z roku 1364 mezi jeho otcem císařem Karlem

Dotaz: [tag="k0"], tj. všechna citoslovce.

Chybovost: asi 50%.

Korpus : cnpkcz
Dotaz : [tag="k0"]
1: vyplnění bude automaticky zaslán na naši < e> - mailovou adresu . Poté se s vámi zkontaktujeme
2: oprávněné podepisovat , datum nar . povolání č . < ú> . peněžní ústav BSČ Vídeň , Potvrzení objednávky
3: postavit . Král se rozohnil : " Jene Huse ! < Hle> , dvě cesty jsou ti položeny , buď abys
4: jednání zúčastnil , prohlásil , ať král < probůh> raději od Zikmundova ďábelského plánu upustí

[1] Seznam značek je uveden v „*Manuálu pro značkování a desambiguaci slovních tvarů v korpusu DESAM*“ od Marty Ungermanové.

5: osídleno již v pravěku (asi roku 4000 před < n> . l .) . Postupem času se pravěké osady
6: žena i dcera zemřely v koncentráku . Pro < Fr> . Hálu : většina členů rodiny v koncentráku
7: korektúru hraníc , pokial ' ide o Slovensko na < juhu> proti Maďarom . Odstráňte menšiny . " Na
8: jedna přes druhou , Hamburg , Mnichov , < pche> ! Na to můžeš zapomenout , teď už je jenom

Německá paralela

V německé paralele je každému slovu automaticky přiřazen slovní druh a základní tvar. Tyto lze vyvolat v nabídce *zobrazení/ view → atributy/ attributes*, nebo klávesou F5. Systém označkování je opět specifický.[2]

Chybovost je zde značná: např. u slova „**rund**“ je ve všech případech uvedeno, že se jedná o adverbium, ačkoliv převážnou část dokladů tvoří adjektiva, části sekundární předložky nebo částice.

Korpus : cnpkde
Dotaz : rund
1: oder die Internierung in Kauf nehmen , was < rund/ADV/rund> die Hälfte vorzog . Im Februar 1940 hatte

2: 2:vorgeschlagenen britischen Militärstützpunkte < rund/ADV/rund> um Deutschland waren für ihn ein
3: Golfparadies auf Barbados Es ist klein , weiß , < rund/ADV/rund> , und versetzt die coolsten Spitzenmanager
4: Zimmer vorhanden ist . Alles an ihm ist < rund/ADV/rund> . Der Kopf , die Augen , deren Grau wie

Vyhledávání podle dotazu „tag“ [tag="VAFIN"] – všechna pomocná slovesa.

Korpus : cnpkde
Dotaz : [tag="VAFIN"]

die Konzilsväter beschwor ein Schisma zu vermeiden . " Ich < bin/VAFIN> der Kaiser und ich will lieber sterben als von neuem
vergiftet worden seien . Der Geschichtsschreiber Thomas Ebendorfer erzählt er < habe/VAFIN> als Kind in seinem Heimatort Haselbach gesehen wie man den
in Visegrád über eine Kandidatur zur deutschen Königswahl verhandelten . < Hätte/VAFIN> erst der neue Papst Sigismund den Vorschlag einer Kandidatur gemacht
wir diese Möglichkeit der neuen Begegnung haben . DER WEG < IST/VAFIN> DAS ZIEL Die alte Poststraße von Wien nach Prag führte
Aufenthalte in der Natur auch anderen übergeben . Am Anfang < waren/VAFIN> wir fünf Mitarbeiter , aber heute , nach 13 Jahren
oder per E - Mail erreichen . In dem Fall < würden/VAFIN> wir auch begrüssen , wenn Sie ihre Teamarbeitkurs - Forderungen

Paralelní text

[2] Informace o systému morfologického značení německé části korpusu jsou dostupné na adrese: http://www.ims.uni.stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Jako v předešlých dvou typech dotazů i k dotazu na tag lze pomocí tlačítka F9 vyvolat odpovídající textovou pasáž (zvýraznění provedeno ručně):

cnpkde: 119903--119925
" Ich **bin** der Kaiser und ich will lieber sterben als von neuem ein Schisma in der Kirche hinnehmen " betonte er
cnpkcz: 113027--113045
" **Jsem** císař a raději zemřu , než bych znovu trpěl nové schizma v církvi , " zdůraznil .
---
cnpkde: 10315--10340
Der Geschichtsschreiber Thomas Ebendorfer erzählt er **habe** als Kind in seinem Heimatort Haselbach gesehen wie man den todkranken Herzog in einer Sänfte nach Österreich zurückbrachte .
cnpkcz: 9583--9609
Historiograf Thomas Eberndorfer píše , že **byl** jako dítě ve svém rodném Haselbachu svědkem toho , jak smrtelně nemocného vévodu dopravovali na nosítkách zpět do Rakouska .
---

cnpkde: 35059--35077
Hätte erst der neue Papst Sigismund den Vorschlag einer Kandidatur gemacht , wäre es notwendig gewesen sein , Einverständnis einzuholen .
cnpkcz: 33442--33457
Kdyby Zikmundovi nabízel trůn nejprve nový papež , bylo by nutné opatřit si jeho souhlas .

- **Nastavení kontextu**

V každé paralele můžeme libovolně nastavit kontext KWICu. Postup je naprosto shodný s ČNK i DESAMem (*zobrazení/ view → kontext/ kontext*, nebo klávesová zkratka F7). Takto rozšířený nebo zúžený kontext nemá v žádném případě vliv na velikost paralelních textů. Rozsah obou paralel nastavit nelze. Ten je dán fixně ručním zarovnáním.

- **Další funkce**

Ostatní funkce korpusu jsou shodné s jednojazyčnými korpusy fungujícími v uživatelském prostředí Bonito (viz manuály Bonita).
Jejich přehled zachycuje následující tabulka (odpovídající „rozbalenému menu manažeru):

Manažer	**Korpus**	**Konkordance**	**Zobrazení**	**Výběr**	**Nápověda**
Připojení	Souhrnné informace	Uložení	Reference	Všechny	O aplikaci
Změna hesla	Statistiky	Tisk	Atributy	Žádné	Ujednání
		Smazání vybraných	Struktury	Inverze	Zobrazení vysvětlivek
		Redukce	Kontext	Vložit do schránky	Dokumentace
		Jednoduché třídění	Rozsah		
		Statistiky- Frekvenční distribuce Kolokace Rozložení	Stránka- Nahoru Dolů		
			Skok na řádek		
			Paralelní text		

- **Problémy a nedostatky**

První a základní nedostatek korpusu je jeho velikost. Korpus je již vhodný ke kontrastivnímu výzkumu uzavřenějších systémů jazyka. Zatím je však příliš malý k serióznímu výzkumu např. v oblasti lexikologie.
Korpus je v současnosti poměrně kvalitně zarovnán. Pouze výjimečně nacházíme neodpovídající si textové pasáže.

Větším problémem je rozsah paralelních textových sekvencí.

Neuspokojivá je taktéž chybovost automatické desambiguace v obou paralelách.

I přes tyto nedostatky se ČNPK stal nenahraditelným pomocníkem kontrastivního výzkumu češtiny a němčiny. Už jen proto, že je svého druhu jediný.

- **Využití**

Česko-německý paralelní korpus je zatím ve zkušebním provozu a slouží pracovníkům a studentům katedry německého jazyka a literatury na PdF MU. Je využíván jako přímá podpora výuky některých disciplin (syntax, lexikologie, překladatelství), osvědčuje se při sestavování výukových materiálů a testů. Již vyšlo skriptum z německé slovotvorby a připravují se další.

Především je však korpus zdrojem autentického jazykového materiálu pro disertační, diplomové, bakalářské, závěrečné a ročníkové práce.

Dosud poskytl korpus materiál pro tyto práce:

a) *Volný dativ v češtině a jeho ekvivalenty v němčině* (disertační práce)
b) *Die syntaktische und semantische Analyse der deutschen und tschechischen Präpositionen* (série ročníkových a diplomových prácí zabývajících se spojeními s předložkami *an, auf, bei, für, in, mit, durch, von, um* a jejich českými ekvivalenty)
c) *Infinitivkonstruktionen als Transformationen der deutschen Nebensätze mit der Konjunktion dass und deren äquivalente Strukturen im Tschechischen* (diplomová práce)
d) *Das Pronomen „Es", seine syntaktischen Funktionen und Äquivalente im Tschechischen* (diplomová práce)
e) *Übersetzung von Okkasionalismen im Werk „Fimfarum"* (diplomová práce)
f) *Die Frequenz der deutschen und tschechischen Satzbaupläne* (závěrečná práce)
g) *Gründe für unterschiedliche Länge der deutschen und tschechischen Te*xte (závěrečná práce)
h) *Der Ausdruck der Vorzeitigkeit in den deutschen und tschechischen Temporalsätzen* (ročníková práce)
i) *Die Stellung der Partikeln im Deutschen und im Tschechischen* (ročníková práce)
j) *Das Subjekt im Deutschen und im Tschechischen* (ročníková práce)
k) *Einige tschechische Ortsnamen und ihre deutschen Äquivalente im ČNKP* (bakalářská práce)
l) *Das Präfix und Präfixoid „Haupt-„ und ihre Äquivalente im Tschechischen* (bakalářská práce)
m) *Suffixoide (z.B. „–werk" und „-zeug") und ihre Äquivalente im Tschechischen* (bakalářská práce)
n) *Komposita mit der Basis „-maschine" und ihre Äquivalente im Tschechischen* (bakalářská práce)
o) *Překladatelské postupy při řešení překladu vlastních jmen* (zatím zadaná bakalářská práce)
p) *Zu einigen tschechischen Toponymen in elektronischen Korpora* (příspěvek na konferenci)

Zájem studentů o práci s korpusem i na korpusu je velký, průběžně jsou zadávána další témata.

- **Perspektivy**

V nejbližší době se tvůrci korpusu zaměří na:

a) rozšiřování a precizování korpusu (vyvažování proporcí, hledání možností efektivnějšího morfologického značení v obou paralelách)
b) vytvoření subkorpusu textů nepodléhajících autorským právům (starší beletrie, texty z internetu) a jeho zpřístupnění širokému okruhu uživatelů
c) výběr a úpravu textů vhodných pro zařazení do multilingválního korpusu InterCorp
d) zpracování uživatelské dokumentace korpusu (manuálu)
e) výzkum v oblasti kontrastivní lingvistiky
f) realizaci předmětu „Úvod do korpusové lingvistiky“ pro studenty německého jazyka na MU
g) sestavování dalších výukových materiálů
h) implementaci korpusu do výuky praktického jazyka na různých stupních škol