Seznam dat a nástrojů získaných v rámci projektu Centra komputační lingvistiky

- **Pražský závislostní korpus, PDT 1.0** (PDT, http://ufal.mff.cuni.cz/pdt)
  RIV/00216208:11320/01:00105063
  Pražský závislostní korpus, PDT 1.0 vydalo LDC v roce 2001 (katalogové číslo LDC2001T10, ISBN: 1-58563-212-0) obsahuje
  -- data:
  - anotovaná data: texty anotované na morfologické (1 974 301 slov / 116 885 vět) a analytické (1 507 372 slov / 87 898 vět) rovině, ukázka anotací na tektogramatické rovině
  - neanotované texty
  - česko anglický paralelní korpus

  -- nástroje
  - NetGraph (vyhledávání na stromech)
  - Tred (stromový editor, vyhledávání na stromech)
  - morfologický analyzátor
  - taggery (zjednoznačnění morfologické informace)

  -- dokumentace

- **Pražský závislostní korpus, PDT 2.0**
  RIV/zatím nepřiděleno
  Pražský závislostní korpus, verze 2.0 je stěžejním výsledkem práce Centra. Jde o obohacení korpusu PDT, verze 1.0 o anotaci na tektogramatické rovině. PDT 2.0 bude vydáno v LDC v roce 2005. PDT 2.0 obsahuje
  -- data:
  - texty anotované na tektogramatické rovině (49 192 vět)

  -- nástroje
  - nové, podstatně rozšířené verze nástrojů NetGraph (viz níž), Tred (viz níž), morfologický analyzátor, taggery

  -- dokumentace

- **Prague Arabic Dependency Treebank, PADT 1.0,**
  RIV/zatím nepřiděleno
  http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T23
  Závislostní korpus moderní standardní arabštiny vzniká s využitím bohatých zkušeností a nástrojů získaných při vytváření PDT ve spolupráci s Ústavem srovnávací jazykovědy FF UK a Linguistic Data Consorcium. Korpus je morfologicky anotován pomocí nástroje od Linguistic Data Consortium (LDC), University of Pennsylvania (anotováno 60 000 slov). V současné době se připravují podklady pro analytické značkování, dále se projekt soustředí na analytické značkování a na získání podkladů pro tektogramatický popis arabské věty.
  LDC2004T23, ISBN 1-58563-319-4

- **VALLEX 1.0,** http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/
  RIV/00216208:11320/03:00002609
  Valenční slovník českých sloves, verze 1.0 je souborem lingvistických dat a dokumentace, který je výsledkem snahy o formální popis valence českých sloves. Verze 1.0 slovníku obsahuje přibližně 1400 sloves, pro něž bylo vytvořeno na 4000 valenčních rámců (1000 nejčastějších sloves z ČNK a jejich vidové protějšky). Při budování VALLEXu je kladen důraz na skutečnost, aby byl slovník snadno a rychle čitelný pro člověka, i na možnost jeho využití v automatických procedurách. Proto je slovník k dispozici v několika formátech: HTML verze (umožňuje snadnou a rychlou orientaci ve slovníku a vyhledávání podle nejrůznějších kritérií), verze pro tisk a XML verze. Po zaregistrování je pro nekomerční účely volně k využití.

- **Český anotovaný korpus,** http://ckl.mff.cuni.cz/~sgd/CAC.html.
  RIV/zatím nepřiděleno
  Anotovaný korpus českého jazyka (o celkovém objemu 560 000 slov) vznikl konverzí původního korpusu anotovaného v Ústavu pro jazyk český AV v sedmdesátých letech. Konverzí vnitřního kódování a anotačních schémat (na morfologické a syntakticko-analytické rovině) získáváme korpus, který je „kompatibilní" s Pražským závislostním korpusem. Byla dokončena konverze vnitřního kódování a morfologického anotování.

- **Prague Czech-English Dependency Treebank, PCEDT 1.0,**
  RIV/zatím nepřiděleno
  http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T25
  Prague Czech-English Dependency Treebank (PCEDT) je paralelní, česko-anglický závislostní korpus, který byl v roce 2004 vydán v Linguistic Data Consortium (LDC, LDC2004T25, ISBN: 1-58563-321-6). Základ paralelního korpusu tvoří překlad přibližně jedné poloviny (24 tis. vět) textů pensylvánského PennTreebanku, verze 3 (vydaného v LDC v roce 1999), který je hlavním zdrojem trénovacích a testovacích dat pro parsery angličtiny. Česká část PCEDT je automaticky morfologicky, analyticky i tektogramaticky označkována, anglická část je automaticky převedena z frázové gramatiky do závislostních analytických i tektogramatických struktur. Vzorek pětiset paralelních vět, určený pro testování, byl navíc na tektogramatické rovině anotován ručně v obou jazycích. Testovací české věty byly přeloženy čtyřmi různými překladatelskými společnostmi do angličtiny a slouží jako referenční překlady pro automatickou evaluaci výstupů překladového systému. Dále budou součástí korpusu paralelní texty z Readers' Digestu (50 tis. vět), překladový česko-anglický slovník forem, nástroje pro automatické sestavení překladového modelu z paralelních dat a nástroje pro zobrazování a vyhledávání v závislostních strukturách.

- **Czech Broadcast News Speech**, vydáno LDC, 2004
  RIV/zatím nepřiděleno
  (katalogové číslo LDC2004S01, ISBN 1- 58563-280-5)
  řečový signál: 22,05 kHz, 16 bitů
  rozsah korpusu: cca 50 hod vysílání
  stanice: ČRo1, ČRo2, ČRo3, ČTV, Prima

- **Czech Broadcast News Transcripts**, vydáno LDC, 2004
  RIV/zatím nepřiděleno
  (katalogové číslo LDC2004T01, ISBN 1-58563-281-3)

- **Korpusy spontánních promluv projektu MALACH** (ZČU Plzeň)
  - Český korpus anotovaných výpovědí lidí přeživších holocaust:
  RIV/zatím nepřiděleno
  řečový signál: 44,1 kHz
  (stereo, 1. kanál - „řečník" poskytující výpověď,
  2. kanál - moderátor), 16 bitů
  počet řečníků: 346
  rozsah korpusu: cca 100 hodin anotované řeči
  počet slov přepisu: cca 0,7 mil. slov

  - Ruský korpus anotovaných výpovědí lidí přeživších holocaust:
  RIV/zatím nepřiděleno
  řečový signál: 44,1 kHz
  (stereo, 1. kanál - „řečník" poskytující výpověď,
  2. kanál - moderátor), 16 bitů
  počet řečníků: 410
  rozsah korpusu: cca 120 hodin anotované řeči
  počet slov přepisu: cca 0,8 mil. slov

  - Slovenský korpus anotovaných výpovědí lidí přeživších holocaust (stav k 31.12.2003):
  RIV/zatím nepřiděleno
  řečový signál: 44,1 kHz
  (stereo, 1. kanál - „řečník" poskytující výpověď,
  2. kanál - moderátor), 16 bitů

počet řečníků: 100
rozsah korpusu: cca 25 hodin anotované řeči
počet slov přepisu: cca 0,2 mil. slov

- **Old-Church Slavonic Corpus** (OCS), http://ckl.ms.mff.cuni.cz/~ribarov.
  RIV/zatím nepřiděleno
  Korpus staroslověnských a církevněslovanských textů je vytvářen na základě dříve zpracovaných rukopisů z Ústavu pro makedonský jazyk, Skopje, Makedonie. Tento korpus obsahuje cca 600 000 slovních forem, lemmatizovaných a morfologicky označkovaných pomocí základní množiny (27) značek. Některé slovní formy (dle příslušnosti) mají asociovaný překlad, případně i referenci k jiným zdrojům. Slovní zásoba pokrývá období od 12. do cca 17. století.

Nástroje vyvíjené v rámci jednotlivých projektů Centra:

- **TrEd**
  Grafický nástroj určený k anotaci a prezentaci stromových struktur rozšiřitelný prostřednictvím uživatelem definovaných maker. Zahrnuje též nástroje pro konverze souvisejících datových formátů, dávkové zpracování souborů a na rozložení dávkového zpracování mezi skupinu výpočetních strojů. Licence GPL, http://ckl.mff.cuni.cz/~pajas/tred.

- **Nástroj pro automatický převod analytických stromových struktur na tektogramatické**
  Automatické předzpracování přechodu mezi anotací na analytické rovině k anotaci na tektogramatické rovině - soubor procedur ve formě maker pro editor TrEd. Obsahuje například algoritmy pro vypouštění uzlů funkčních slov a interpunkce, spojení analytických tvarů sloves, spojení uzlů modálních sloves s významovým slovesem, přiřazení tektogramatických lemmat uzlům, přiřazení hodnot gramatémů na základě morfologických značek z analytické roviny; http://ufal.mff.cuni.cz/publications/year2001/MN+dodat.doc.

- **XSH**
  Univerzální nástroj na interaktivní i dávkové zpracování XML souborů prostřednictvím jednoduchého jazyka založeného na standardu XPath. Licence GPL, http://xsh.sourceforge.net.

- **NetGraph**
  Souběžně s Pražským závislostním korpusem (PDT) je vyvíjen nástroj Netgraph, program pro prohledávání PDT (a jiných korpusů podobného formátu). Netgraph má architekturu klient-server a umožňuje uživatelům vyhledávat v korpusu, umístěném na výkonném serveru, z kteréhokoliv bodu internetu pomocí uživatelsky přívětivého, ale přesto velmi výkonného grafického rozhraní. Přehledný, plně grafický dotazovací jazyk je každým rokem zesilován – v roce 2003 přibyly především relace jiné než rovnítko, negace a odkazy na hodnoty atributů jiných uzlů.
  V listopadu 2003 byl Netgraph v rámci oboustranné spolupráce instalován rovněž v Linguistic Data Corporation (LDC) na University of Pennsylvania ve Philadelphii v USA, kde slouží k prohledávání arabského korpusu, tamním pracovištěm vytvářeného.
  Netgraph je pro akademické účely volně k dispozici na internetu, včetně podrobné dokumentace – viz http://quest.ms.mff.cuni.cz/netgraph.

- **Syntaktické analyzátory češtiny** ("parsery")
  V CKL se paralelně vyvíjejí nástroje pro povrchovou syntaktickou analýzu (odpovídající analytické rovině PDT) založené na různých přístupech.
  - Statistický parser (tzv. Zemanův parser)
  Tento parser je založen na statistickém modelování závislostí mezi slovy. Balíček s parserem bude vyvěšen ke stažení na domovské stránce CKL a analýza bude také pokusně zprovozněna on-line prostřednictvím webových formulářů.
  - Pravidlový parser
  Tento parser je založený na automaticky získávaných pravidlech (tzv. rule-based přístup a jeho modifikace pro závislostní syntax), neobsahuje žádné před nebo post zpracování výsledných struktur.

- **Nástroje používané ve strojovém překladu**
  Nástroje jsou podrobně popsány v dokumentaci k Prague Czech-English Dependency Treebank, který bude vydán na CDROM v r. 2004 v LDC (viz výše bod 1.5).

- **Editor pro morfologickou anotaci spontánních promluv projektu MALACH**
  Vstupem editoru pro morfologickou anotaci jsou textová data zpracovaná českým morfologickým analyzátorem a taggerem. Program umožňuje snadnou vizuální kontrolu a případnou manuální korekci automaticky označkovaného textu. Jelikož byl editor vyvinut zejména pro anotaci spontánní řeči, lze v něm též opravit hovorové tvary češtiny na tvary spisovné, přičemž je současně automaticky vytvářen slovník obsahující původní nespisovné a opravené spisovné tvary.

- **Nástroj pro vytváření anotovaných korpusů ACT**
  V rámci vývoje technologií pro zpracování psaného slovanského kulturního dědictví byl za pomoci studentů vyvinut programový balík ACT (Annotated Corpora of Text) - jazykově nezávislý nástroj pro vytváření anotovaných korpusů s řadou speciálních funkcí pro zachycení jazykových víceznačností a variant. V rámci ACT je možné lemmatizovat, desambiguovat (s možností registrovat více správných variant), morfologicky značkovat, určovat reference k jiným zdrojům, určovat víceslovní celky nejrůznějších druhů, udržovat slovník lemmat, spravovat různé redakce slovníku, pracovat s překlady a asociovat text s jeho překladem. Je podporováno libovolné vyhledávání výskytů slov, včetně kontextových dotazů a předzpracovaných komplexních dotazů jako nejrůznější typy indexů, retrográdních indexů apod. V rámci ACT lze nalézt i prostředí pro zpracování lexikálních kartotéčních lístečků s cílem zpětné rekonstrukce původních excerpovaných textů. Licence GPL.