**NLP (Natural Language Processing)**

* počítačové zpracování přirozeného jazyka; počítačová lingvistika
* obor na pomezí lingvistiky, informatiky (umělé inteligence), popř. taky akustiky aj.; zkoumá problém analýzy či generování textů nebo mluveného slova, které vyžadují určitou (ne absolutní) míru porozumění přirozenému jazyku strojem
* zabývá se zkoumáním a zpracováním přirozeného jazyka pomocí počítačů a metod informatiky
* úkoly
* analýza přirozeného jazyka (morfologická, syntaktická, sémantická)
* generování přirozeného jazyka (vyvíjení, vytváření)
* syntéza a rozpoznávání řeči
* strojový překlad (Machine translation)
* odpovídání na otázky (Question answering)
* získávání informací (Information retrieval)
* korektura textu (Spell-checking, Grammar checking)
* extrakce informací (Information extraction)
* výtah z textu (Text summarization)
* určení typu dokumentace (Text Classification/Clustering)

**Korpusová lingvistika**

* lingvistická disciplína zkoumající jazyk pomocí elektronických jazykových korpusů a zabývající se i výstavbou těchto korpusů, jejich zpracováním a příslušnou metodologií
* jako obor se začala výrazněji vyhraňovat a rozvíjet v posledních dvou desetiletích 20. století, souvisí s rozvojem výpočetní techniky
* interdisciplinární spolupráce lingvistů s dalšími obory, především matematikou a informatikou
* texty jsou chápany jako produkty jazykového systému a jazykových schopností jejich tvůrců; skrze ně se dospívá k poznání obecnějších jazykových zákonitostí a pravidel
* lingvista je poprvé v historii zbaven nejistoty, zda nepracuje s příliš omezenou materiálovou základnou a zda pozorování a závěry, které z ní vycházejí, nejsou deformované
* charakteristika korpusových dat:
  + typická, nenáhodná a věrná ve vztahu k tomu, jak lidé užívají jazyk
  + aktuální, skutečně odrážející svou dobu
  + neselektivní a objektivní
  + dostatečná
  + s pomocí počítače snadno získatelná a rychle přístupná
* korpusy v zemích:
  + *Brown Corpus* (VB)
    - VB průkopnická země v budování velkých korpusů
    - konstituovala se zde nejvýznamnější korpusová disciplína – korpusová lexikografie
  + *Bank of English*
  + *British National Corpus*
  + *International Corpus of English* (cíl: mapovat všechny ve světě užívané varianty anglického jazyka)
  + *Frantext* (Francie)
  + dánština, němčina, italština, španělština, Maďarsko, Slovinsko, Litva
* monitorování korpusů: mezinárodně - organizace ICAME

**Korpus**

* základní prostředek výzkumu korpusové lingvistiky = korpus
* je nejlepší aproximací, nejlepší vzorek skutečného jazyka
* přesvědčení, že lépe než prostřednictvím korpusu nelze dnes jazyk při studiu uchopit
* strukturovaný, unifikovaný (často i označkovaný) rozsáhlý soubor jazykových dat, který je elektronicky uložený i zpracovaný
* skládá se zpravidla z jednotlivých textů a jako celek si činí nárok na reprezentativnost
* reprezentativnost: zdůrazňuje odpovídající zastoupení jednotlivých typů jazyka z hlediska míry, v jaké jsou mluvčími přijímány
* korpusy z celých textů

+ snižuje nebezpečí jednostrannosti malého korpusu

+ zvyšuje reprezentativnost malého korpusu

- dají se hůře studovat jevy celotextové povahy

* značkování:
  + *vnější anotace:* typ textu, autor, rok (lze pak vybírat texty z určitého roku, texty psané jen ženami apod.)
  + *vnitřní anotace:* strukturní informace (členění na kapitoly, odstavce, věty, slova) a informace lingvistické
    - lingvistické informace nákladné => omezuje se na morfologické značkování jednotlivých slovních tvarů (tagování), přiřazení slovnědruhové charakteristiky a lemmatizaci (přiřazení základního, slovníkového tvaru)

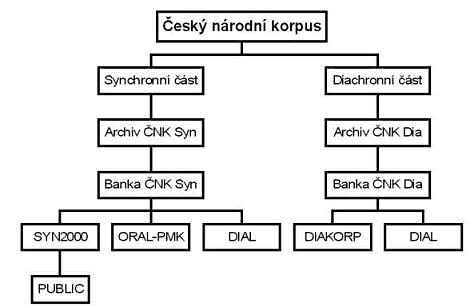
**Typy korpusů**

* malé, střední, velké
* obecně reprezentativní nebo budované k rozmanitým specifickým cílům
* základní členění: lingvistické
  + *psané, mluvené*
    - *mluvené korpusy*: nákladnost, velká investice a úsilí; neměl by se používat jazyk médií, protože je přechodné povahy, není spontánní; je formální, připravený; snaha vybudovat synchronní nářeční korpusy
  + *diachronní* (pokrývá stádia jazyka), *synchronní* (současné psané texty; převažuje)
    - diachronní korpus × korpus historický (jedno časové stádium vývoje jazyka)
* korpusy monitorovací, paralelní, studijní, testovací
  + *monitorovací:* otevřené novým textům a probíhajícímu jazykovému vývoji; v užším slova smyslu = pevná část korpusu plnící úlohu filtru určeného k registraci nových jazykových jevů a prvků, které v korpusu nejsou zastoupeny a které je třeba doplnit
  + *paralelní:* dva a více jazyky vytvářené z překladů, obsahující texty originální i jinojazyčné mutace; stále více aktuální; navíc cenné pro výuku studentů, překladatelů atd. (např. paralelní korpus G. Orwella ve 23 jazycích)
  + *studijní korpusy:* tvořeny hlavně texty psanými studenty cizích jazyků; analýza způsobů vyjadřování, odstínů i chyb => zkvalitnění výuky
  + *cvičné a testovací korpusy:* rozsahem omezené, většinou plně provedeno a manuálně opraveno značkování (vnitřní anotace), na základě jich se trénují a vylepšují různé anotační programy a lingvistické hypotézy

**Využití korpusů**

* cenné zdroje pro lingvisty, ale i pro další odborníky (literární vědce, sociology, psychology, pedagogy)
* i pro studenty, zájem jeví i laikové
* lingvisté:
  + výstavba popisů jazyka
  + tvorba a ověřování teorií
  + tvorba velkých aplikací založených na korpusových datech (hl. lexikografové)
* uživatelé:
  + konkordance (výpis řádků s výskytem zkoumaného jevu v textu) předkládá lépe podložená data a zasazena do dostatečného, libovolně rozšiřitelného kontextu
  + frekvenční a statistické charakteristiky příslušných dat => představa o tom, co je typické a co okrajové
  + můžeme-li porovnat s diachronním korpusem -> závěry o vývoji, tendencích atd.
* ukazuje škálovitou povahu jevů v jazyce
* variabilita jazyka

**Český národní korpus**

* kontinuální projekt, jehož produkty (jednotlivé korpusy) mapují a monitorují různé podoby českého jazyka s cílem zpřístupnit uživatelům co nejbohatší zdroj jazykových dat
* akademický, nekomerční projekt
* vytvářen Ústavem Českého národního korpusu (ÚČNK) na FF UK, který vede František Čermák
* založen 1994
* na tvorbě se podílí i další pracoviště, např. FI MU, Ústav českého jazyka FF MU
* problémy: konverze mezi různými formáty textových souborů, problematika způsobu anotace
* části:
  + SYN2000: 100 000 000 slovních tvarů
    - vychází z něj korpus PUBLIC (20 000 000, veřejně na internetu)
  + DIAKORP: 1 750 000 tvarů; diachronní psaný
  + ORAL-PMK: 700 000 tvarů, synchronní mluvený
  + složeny zásadně z celých textů (nikoliv jen ze vzorků)
* nekorpusové složky:
  + archivy (uložení surových, nezkonvertovaných a neoznačkovaných textů; Archiv ČNK Syn, ČNK Dia)
  + banky (uložení textů v konečném formátu SGML)

**Synchronní a diachronní korpus, hranice mezi nimi**

* v jazyce neexistují jasná a nepochybná kritéria pro stanovení časových rozmezí => tedy ani pro stanovení hranic jednotlivých druhů korpusů
* vodítko: to, co dnešní mluvčí pociťují jako živé a co už ne
* vymezení spodní časové hranice v ČNK:
  + novinové a časopisecké texty: začátek 1990 (v synchronním textu); hlavně kvůli tomu, že starší noviny plné dobového ideologického newspeaku; publicistika se navíc rychle mění
  + krásná literatura: 1990
    - doplňující kritéria vzhledem k přetiskům a čtení knih starších autorů; do synchronního korpusu se zařazují i
      * současně čtení starší autoři, kteří se narodili roku 1880 a později
      * knihy publikované od roku 1945
  + odborné texty: po roce 1989
* texty za těmito časovými hranicemi => diachronní korpus
* omyl domnívat se, že v korpusu jsou v elektronické formě přístupná celá česká literatura daného období

**SYN2000**

* časová etapa 1990-1999
* 100 milionů slov, vnější značkování; pokusy o značkování vnitřní (lingvistické)
* množství jazykového materiálu, jaké u nás dosud nikdy nebylo zpřístupněno k operativnímu prohledávání
* samotné texty mají rozsah 1–2 gigabytů; prosté manuální prohlížení přesahuje lidské možnosti;
* pro představu: průměrné knihy o 250 stranách by zaplnili 10 metrů knihovních regálů
* přečtení by zabralo 4 roky (8 h denně, 365 dní ročně)

**Diachronní korpus**

* cíl: vytvořit elektronickou materiálovou základnu pro výzkum vývoje českého jazyka od prvních dochovaných souvislejších záznamů (2. polovina 13. století) zhruba do poloviny 20. Století, tj. k hranicím synchronního korpusu
* zařazovány pouze dobově a útvarově autentické texty (u nichž lze vyloučit, že do nich byly vneseny prvky pozdějšího jazykového stavu nebo jiného jazykového útvarů – tj. opisy, přetisky, zkreslení nářečními prvky)
* postup budování pomalý a obtížný, většinu textů potřeba manuálně přepisovat, skenovat nebo manuálně korigovat
* odlišné vývoje českého pravopisu; texty vstupují v transkribované podobě (přepis z jednoho písma do druhého)
* DIAKORP v roce 2000 1 750 000 textových slov; součástí banka transliterovaných textů (asi 100 000 slov) a jazyková databáze (překlady starších českých slov, vysvětlivky obratů, …)

**Mluvený korpus**

* ORAL-PMK (Pražský mluvený korpus)
* dnešní rozsah 700 000 slov
* omezen na oblast Prahy a okolí, kde je míra širší reprezentativnosti nejvyšší (Praha ovlivňuje ostatní území nejen mediálně, ale i tím, že v ní pracují lidé z celé země)
* období 1988–1996; i přes svá omezení zdaleka největší a nejreprezentativnější záznam autentické mluvené češtiny
* odpovídá reálným možnostem (vybudování korpusu je pracnější a dražší)
* 300 magnetofonových nahrávek v upraveném přepisu
* reprezentativní zastoupení 4 hlavních sociolingvistických proměnných:
  + Pohlaví, věk, vzdělání, typ textu/nahrávky (formální/neformální)
* dosáhnuto vyvážené proporcionální podoby
* manuální přepisování, anotace
* využitelnost: lexikon, morfologie
* vzhledem k malému rozsahu méně vhodný pro studium syntaxe, ke zkoumání tematickému apod.

**Reprezentativnost ČNK**

* ideální korpus: zahrnuje z hlediska matematické statistiky neuskutečnitelné soustředění naprosto všech textů, které bychom chtěli mít k dispozici
* reálný korpus: to, co statistikové nazývají vzorek, výběr, sample
* reprezentativnost neroste lineárně s pouhým zvětšováním rozsahu, ale závisí především na tom, co sledujeme na pravděpodobnostních charakteristikách (rozložení) výskytů sledovaných prvků nebo jevů
* je vyloučeno, aby byl korpus naprosto všestranně reprezentativní
* neexistuje univerzální, všeobecně uznávaná představa o využití korpusů, staly se východiskem pro ČNK odhady možných očekávání ze strany dnes uvažovaných potenciálních zájemců o práci s korpusem
* na základě průzkumů => se stal výchozím bodem členění textových zdrojů ČNK poměr čtení knih, časopisů a novin; vyplynulo zastoupení novin **60 %**

|  |  |
| --- | --- |
| TYP TEXTU | Podíl [%] |
| IMAGINATIVNÍ TEXTY | 15,00 |
| krásná literatura | 12,04 |
| poezie | 0,81 |
| drama | 0,21 |
| próza | 11,02 |
| jiné imaginativní texty | 0,36 |
| přechodové pásmo | 2,60 |
| INFORMATIVNÍ TEXTY | 85,00 |
| publicistika | 60,00 |
| odborné texty | 25,00 |
| vědy o umění | 3,48 |
| sociální vědy | 3,67 |
| právo a bezpečnost | 0,82 |
| přírodní vědy | 3,37 |
| technika | 4,61 |
| ekonomie a řízení | 2,27 |
| víra, náboženství | 0,74 |
| životní styl | 5,55 |
| administrativa | 0,49 |

* druhá úroveň: vyvážený souhrn výsledků tří různých průzkumů týkající se porovnání zájmu o literaturu naučnou a krásnou (průzkum výpůjček ve veřejných knihovnách, průzkum struktury zájmu o žánry nových knižních edic ze strany čtenářů nových knih a o průzkum vztahu mezi fondy a výpůjčkami ve veřejných knihovnách); z výsledku odvozeno procentuální zastoupení naučné literatury (informativních nenovinových textů) – 25 %; krásné literatury (imaginativních textů) – 15 %
* korigovány z věcných hledisek:
  + vzhledem k požadavku co nejširšího záběru ČNK
  + vzhledem k omezené dostupnosti hovorového jazyka
  + vzhledem k univerzálnosti využití ČNK
* systém třídění a anotace se opírá o standardy vypracované skupinou TEI a skupiny EAGLES, jejich doporučení jsou podporována Evropskou unií
* vlastní sada charakteristik a vlastní škálování uvnitř těch charakteristiky (na návrh Čermáka, modifikována podle Deweyova desetinného třídění a dle zkušeností zahraničních korpusových projektů) => 11 kategorií a 112 rysy uvnitř těchto kategorií

**Získávání textů**

* pěti různými způsoby:
* *prostřednictvím smluv s nakladateli a vydavateli* (naprostá většina, více než 90 % SYN2000; více než 250 smluv, zavazují ČNK používat texty pouze pro budování korpusu a k nekomerčnímu užívání – rozsah citací jednotlivých textů omezen horní hranicí 100 slov)
* *využíváním textů dostupných na internetu* (příliš nevyhovují – jen zlomek spadá do oblastí, v nichž je potřeba korpus doplňovat; navíc jsou to nekvalitní texty, s množstvím jazykových chyb, překlepů a diakritiky)
* *skenováním* (7 % v SYN2000)
* *manuálním přepisem*
* *darem od autorů* (zanedbatelné množství)

**SGML (Standard Generalized Markup Language)**

* standardní jazyk určený k formálnímu popisu struktury dokumentů. Je definován v normě ISO 8879 z roku 1986
* vznikl v rámci projektu ODA (Open Document Architercture), jejímž cílem je poskytnout standardní architekturu pro vytváření, předávání, uchování a zpracování různorodých dokumentů v elektronické podobě
* zahrnuje různé standardy pro formáty dat, architekturu předávání zpráv, zabezpečení informací atd.
* zcela otevřený standard nezávislý na platformách, výrobcích nebo aplikacích. Soubory SGML jsou ukládány jako text ACII, což zajišťuje jejich použitelnost prakticky na libovolné počítačové platformě
* univerzální značkovací metajazyk, který umožňuje definovat značkovací jazyky jako své vlastní podmnožiny
* význam a přínos SGML se objevuje ve chvíli, kdy je dokument označen příslušnými značkovacími příkazy. Definování struktury a vnitřních vztahů v do té doby nestrukturované informaci otevírá zcela nové možnosti zpracování, publikování a opakovaného používání této informace.
* SGML databáze může obsahovat tisíce označkovaných dokumentů a využitím vyznačené struktury z ní lze na přání publikovat obsahově stejné dokumenty mnoha různých forem (například manuál k letadlu o několika tisících stranách; každý odstavec v dokumentu je označen z hlediska přístupových práv – volně přístupný, tajný, přísně tajný – podle toho jej pak mohou číst technici, piloti nebo personál)
* protože SGML uchovává informace o struktuře dokumentu, nikoliv o jeho formátování, může mít manuál různou podobu podle způsobu jeho použití (tisk, CD-ROM).
* SGML je metajazyk, který umožňuje definovat, jaké značky (tzv. elementy) lze v textu použít a jak spolu souvisí (např. uvnitř elementu, který označuje kapitolu, se mohou vyskytovat elementy vyznačující odstavce)
* SGML je vlastně metajazyk, který umožňuje definování dalších
* **DTD (Document Type Definition) =** definice přístupných elementů a vztahy mezi nimi; DTD bývá nejčastěji uložena v separátním textovém souboru. Samotný dokument se pak skládá z identifikace DTD, pomocí něhož je zapsán (tzv. prolog) a vlastního textu doplněného o elementy definované v DTD. DTD nedefinuje konkrétní význam jednotlivých elementů. To obstarávají další programy, které mohou dokument vytvořený na základě konkrétního DTD převést např. do PostScriptu, pokud chceme dokument tisknout, či do nějakého hypertextového systému, pokud chceme dokument prohlížet interaktivně. Jedinou skupinou programů, které jsou nezávislé na jednom konkrétním DTD, jsou tzv. *parsery*, které provádí kontrolu SGML dokumentů - tj. kontrolují, zda použití elementů v dokumentu odpovídá jejich definici v DTD.
* *Ukázka SGML* – chceme v elektronické podobě uchovávat sbírku básní (sbírka básní se skládá z několika básní, ta má název a několik veršů; každý verš se skládá z několika řádek)
  + celá sbírka je označena elementem book
  + konkrétním výskytům elementů v textu se říká *tagy*

*<book>*

*<poem><title>Cikáda</title>*

*<verse>*

*<line>Vzala ho za ruku</line>*

*<line>a vedla mezi zelené keře</line>*

*<line>Když stál proti ní</line>*

*<line>bledničkové nahý</line>*

*<line>dotkla se ho vlhkým prstem</line>*

*</verse>*

*<verse>*

*<line>Ani další noc</line>*

*<line>to nebylo dál</line>*

*</verse>*

*</poem>*

*<poem>*

*... následují další básně ...*

*</book>*

* definice příslušného DTD vypadá následovně:
  + za klíčovým slovem !ELEMENT vždy následuje definice nového elementu

*<!ELEMENT book - - (poem+)>*

*<!ELEMENT poem - - (title?, verse+)>*

*<!ELEMENT title - O (#PCDATA)>*

*<!ELEMENT verse - O (line+)>*

*<!ELEMENT line O O (#PCDATA)>*

* jelikož náš DTD umožňuje některé tagy vynechat, lze celou naši sbírku zkráceně zapsat takto:

*<book>*

*<poem><title>Cikáda*

*<verse>*

*<line>Vzala ho za ruku*

*<line>a vedla mezi zelené keře*

*<line>Když stál proti ní*

*<line>bledničkové nahý*

*<line>dotkla se ho vlhkým prstem*

*<verse>*

*<line>Ani další noc*

*<line>to nebylo dál*

*<poem>*

*... následují další básně ...*

*</book>*

* korektnost zápisu bychom mohli ověřit parserem, kterému bychom na vstup zadali naši sbírku básní a náš DTD

**výhody SGML**

* jednotný formát
* SGML není spojen s žádným operačním systémem nebo aplikací a tak je přenosný mezi různými systémy
* největší výhoda systémů založených na SGML je jejich **otevřenost.** Vše pracuje v rámci standardu, který je již definitivní a nehrozí tedy znehodnocení investic v důsledku nutnosti přechodu na nové verze, jako je tomu u komerčních textových procesorů.
* Jedná se o standard, který spravuje ISO (mezinárodní organizace pro standardy) => velmi stabilní, a i když existují možnosti pro jeho doplnění a změnu, ISO činí tento proces velmi obtížným. To se ukázalo být velkou výhodou, neboť je tak k dispozici známý syntax.
* další velkou výhodou je to, že při tvorbě SGML dokumentů nás **zajímá obsah** a nikoliv forma dokumentů. Konkrétní formu lze určit až v závislosti na požadovaném výstupu
* pokud je DTD dobře navrženo, je velmi usnadněna i automatická syntéza řeči (tagy vyznačují místa, kde se obvykle i v řeči dělají pauzy), což je velkým pozitivem při zpřístupňování informací nevidomým občanům.
* To, že je celý text rozčleněn a označen různými tagy, přináší i **výhody pro prohledávání dokumentů**. Jednak lze podobně jako u strukturovaných databází prohledávat jen určité položky (např. název a autora dokumentu) a jednak lze informaci o struktuře využít při fulltextovém prohledávání - slova obsažená v nadpisech a zvýrazněná odlišným typem písma mívají obvykle větší význam než ostatní text.
* díky **flexibilitě** SGML není problémem do textu vkládat i další méně obvyklé prvky jako jsou tabulky, matematické vzorce, notové záznamy atd.
* SGML nenabízí přesně určenou sadu tagů, ale syntax pro vytváření vlastních. Mnohá průmyslová odvětví vytvořila konsorcia, ve kterých jsou vytvářeny sady společných tagů pro výměnu informací za použití termínů a výrazů typických pro danou oblast
* *v současné době je asi nejznámější aplikací SGML jazyk HTML (Hypertext Markup Language),* který je používán pro tvorbu hypertextových dokumentů v rámci Internetovské služby World-Wide Web
* většina velkých softwarových firem (IBM) má své systémy na správu dokumentace rovněž založeny na SGML

**nevýhody**

* flexibilita: není lehké psát parsery
* nejrozšířenější WWW nepodporují SGML; pokud chceme něco dostat na web, nutná konverze do HTML, čímž dochází ke ztrátě významných informací (zejména o struktuře dokumentu), tím i k znehodnocení dat
* SGML je komplikovaný jazyk a obtížně se integruje do jednotlivých aplikací.
* SGML vyžaduje "parser", který je obtížné napsat a udržovat. Jelikož byl SGML vytvořen na počátku éry stolních počítačů, je příliš zaměřen na úsporu omezené paměti a diskového prostoru a proto využívá řadu minimalizačních pravidel a výjimek.
* Výsledkem této složitosti je fakt, že SGML je dražší než jednoduchý systém tagů jako HTML. Každý dokument musí obsahovat "definici typu dokumentu", takže vlastník dokumentu musí provést jeho analýzu, aby odkryl jeho strukturu.
* Po SGML zatím existuje jen omezená poptávka, takže je obtížné nalézt odborníky pro analýzu dokumentů a tvorbu definic typu dokumentu a je drahé si je udržet.
* Protože je standart tak složitý a existuje menší poptávka, nástroje podporující SGML jsou dražší než v případě HTML.

**Zpracování textů**

* každý text vstupující do ČNK je zaevidován a ve své původní podobě (tj. v té, v jaké byl získán od nakladatele, naskenován, přepsán apod.)
* poté uložen do textového archivu
* pro potřeby ČNK nutno získané texty konvertovat do jednotného formátu SGML, anotovat je a označkovat
* takto připravené texty uloženy do textové banky a dále se s nimi pracuje jako se soubory
* při konverzi se v co největší míře zachovává autenticita textu (původní členění, neopravují se chyby ani zjevné překlepy apod.); výjimka: odstraňování odstavců obsahujících cizojazyčný text, tabulky, vzorce apod.
* formát SGML, v němž jednotlivé soubory vstupují do banky, je přizpůsoben potřebám ČNK: používá vlastní DTD (document type definition – popis SGML dokumentu) a jednotné kódování češtiny (ISOLatin2, tj. ISO-8859-2)
* každý soubor je jednoznačně identifikován svým *jménem* (žádné dva soubory v bance nemohou být stejně pojmenovány) a opatřen *hlavičkou* se všemi relevantními technickými informacemi o konverzi
* hlavička souboru je pro běžného uživatele skrytá, technickým pracovníkům však pomáhá dohledávat a opravovat eventuální chyby
* každý soubor v bance je anotován a strukturován do dokumentů, odstavců, vět a slov
* při konverzi se dbá na to, aby bylo členění dokumentů na odstavce zachováno v původní podobě; členění se provádí plně automaticky při převodu z 2. mezifomátu do SGML
* **konverze**textových souborů probíhá ve třech hlavních fázích:

1. *převod textu z původní podoby do 1. meziformátu*
2. *převod textu 1. meziformátu do 2. meziformátu*
3. *převod textu z 2. meziformátu do formátu SGML*

* hlavním důvodem převodu textu do 1. meziformátu je to, aby všechny texty získaly jednotnou podobu a veškeré další zpracování mohlo probíhat jednotně, bez ohledu na původní formát textů. Výsledek tohoto dalšího zpracování (2. meziformát) se od 1. meziformátu na první pohled výrazně neliší: jde rovněž o ASCII text, navíc však obsahující hlavičku SGML s lingvistickou anotací textu, kterou většinou provádějí přímo lingvisté na PC
* posledním krokem konverze je převod z 2. meziformátu do formátu SGML. Tento proces je již plně automatizován, texty při něm procházejí nejprve *tokenizerem* (programem pro segmentaci textu do vět a slov), který je převede do SGML, a poté ještě několika *čisticími*a*kontrolními programy*.
* posledním krokem převodu z 2. meziformátu do SGML je **značkování (tagování)** textů, tj. přidávání morfologických značek a lemmat k jednotlivým slovním tvarům v textech

**Uložení korpusu v počítači**

* Ne každý soubor textů volně uložený v počítači je automaticky korpus. Text je nutné nějakým způsobem strukturovat. Podle toho rozlišujeme
  + archiv textů, tedy soubor textů různých formátů a kódování – nejhrubší podobu korpusů.
  + textové banky jsou organizovanější, texty uloženy v jednotném formátu a je na nich provedeno základní značkování.
  + Vertikální text
  + Binární data v aplikaci, pomocná data pro rychlejší zpracování (indexy, statistiky)
  + konečnou formou je pak použití korpusového manažeru, který umožňuje podrobnou práci s korpusem, umožněnou podrobnějším značkováním. Tento „finální“ značkovaný korpus se nazývá anotovaný. Je zde rozlišena logická struktura textů, typografický vzhled a především jazykové jevy.
* Než se však začne korpus upravovat do anotované podoby, je nutné přesunout data do počítače. K tomu existuje několik možností.
  + manuální přepis, časově velmi náročný
  + importovat text do počítače ze sázecích disket a pásek získaných od nakladatelství, nejjednodušší
  + OCR neboli skenování; nejrozšířenější
  + převod mluveného slova na psaný text
* následujícími typy uložení jazykových dat:
* *elektronické archivy* - volné kolekce celkově různorodých textů. Klasickým příkladem je *Oxford Text Archive - OTA*, který představuje rozsáhlou sbírku různých, většinou literárních textů, v různých formátech a různých jazycích: v OTA najdeme asi tisícovku literárních textů v 25 jazycích a různých formátech,
* *vlastní korpusy* tvořící relativně úplné celky, i tak ovšem značně různorodé a lišící se v řadě parametrů,
* *podle jazyků* - dnes už jen málo jazyků v Evropě nemá svůj korpus, v r. 1990 existovaly korpusy pro:
* angličtinu (220 000 000 slovních tvarů, 20 korpusů)
* francouzštinu (190 000 000 slovních tvarů)
* němčinu (27 500 000 slovních tvarů)
* holandštinu (60 000 000 slovních tvarů)
* italštinu (30 000 000 slovních tvarů)
* srbochorvatštinu (12 000 000 slovních tvarů)
* korpusy *dvoujazyčné*, paralelní: anglicko-francouzské, -italské, -dánské
* korpusy *obecné* a *specifické*, velké obecné korpusy obsahují subkorpusy jazyka psaného, mluveného, nářečí, synchronní - diachronní aj.

**Značkování**

* značkování (tagování) je proces, při němž jsou texty vstupující do korpusu opatřovány (anotovány) doprovodnými informacemi, které jsou formálně vyjádřeny prostřednictvím*značek*(*tags*). Tyto značky jsou trojího druhu: administrativní či správní (vnější anotace), strukturní a lingvistické (vnitřní anotace).
* *administrativní (správní) značky* zachycují administrativní údaje o každém textu soustředěné v tzv*. hlavičce* a obsahují zejména informace o původu, autorství, typu a zdroji textu
* *strukturními značkami* bývá hierarchicky členěn vlastní text (kapitoly, odstavce)
* *tokeny* jsou testová slova, tj. výskyt slovních tvarů (slovních exemplářů), čísel, zkratek, speciálních znaků a interpunkčních znamének
* *lingvistické informace* umožňují vyhledávat v koprusu nejen samostné slovní tvary a jejich konfigurace, ale i jiné lingvistické charakteristiky (morfologické, včetně slovnědruhové příslušnosti); tyto lingvistické značky jsou přiřazovány k jednotlivým slovům tzv. morfologickým analyzátorem (programovým modulem opatřujícím automaticky každé slovo v textu lingvistickou informací dvojího typu:
  + **LEMMATIZACÍ** je danému slovu přiřazena informace o jeho základním, slovníkovém tvaru zvaném lemma, popř. o více možných základních tvarech (tj. u slov víceznačných v rámci jednoho SD nebo v rámci více SD
  + **PŘIŘAZENÍ POTENCIÁNÍCH MORFOLOGICKÝCH INTERPRETACÍ** každé formě, tj. informace o její slovnědruhové příslušnosti a morfologických vlastnostech (např. o rodu, čísle a pádu podstatných a přídavných jmen, zájmen a číslovek, o stupni přídavných jmen a příslovcí, o osobě, čísle, slovesném a jmenném rodu slovesných tvarů atd.). Morfologická interpretace daného slova je formálně vyjádřena *morfologickou značkou* tvořenou maximálně 15 údaji, z nichž každý je reprezentován jedním znakem na dané pozici, přičemž význam jednotlivých pozic je jednoznačně stanoven.

**Automatická morfologická analýza a desambiguace**

* anotátor: náležitě zaučený a vycvičený pracovník, který anotuje texty
* desambiguace: procedura; zjednoznačnění homonymních forem; kontextové zjednoznačnění
* otázka, co je nejpřesnější a nejméně chybová automatická desambiguace korpusových textů; hrají tu roli faktory syntaktické i sémantické
* úspěšnost ovlivněna softwarovými nástroji, morfologickou a slovnědruhovou víceznačností jazyka, v němž jsou značkované texty psány a také kvalitou (chybovostí) samotných značkovaných textů
* čeština je jazyk s velmi vysokou mírou víceznačnosti jak morfologické, tak slovnědruhové: existuje v ní přes 1000 tříd systémové homonymie; pkl automaticky morfologicky desambiguovat český text je mimořádně složitý, a to i v porovnání s ostatními slovanskými jazyky; dosavadními prostředky toho není možné dosáhnout bez chyb
* dvě základní metody automatické desambiguace textů:
  + ***stochastická (statistická, pravděpodobnostní) desambiguace:*** model, který je založen především na pravděpodobnostech přechodu mezi jednotlivými značkami v morfologicky analyzovaném textu. Princip: nejprve ručně (správně) označkujeme větší množství textů (řády set tisíc slov), a vznikne tak trénovací korpus. Tagger (statisticky koncipovaný desambiguací program) se poté „naučí“ toto správné značkování, tj. učiní si představu o pravděpodobnostech přechodu mezi jednotlivými značkami a jejich četnostech, kterou uloží do svých vnitřních tabulek. Program pak tyto „naučené“ znalosti aplikuje na nedesambiguovaný korpus; v angličtině úspěšnost 97 a 98 %, v ČN 94 % (asi každé 16 slovo je špatně). Hlavní problém: nedostatek trénovacích dat, syntagmatická a slovosledná variabilita textů je příliš velká
  + ***pravidly řízená desambiguace:***Kvůli neúspěšnosti stochastického modelu zahájen vývoj metody disambiguace založené na syntaktických pravidlech. Podstatou je intuitivní formulace celé řady syntaktických pravidel, která odrážejí syntaktické konfigurace češtiny dané jejím vnitřním systémem. Jakmile je formulováno určité pravidlo, které vyplynulo z analýzy obecné chyby, ihned se počítačově implementuje a ověřuje na datech korpusu. Poněvadž tato metoda modeluje jazykový systém, není – na rozdíl od metody stochastické – závislá na trénovacích datech a vlastně je vůbec nepotřebuje. Pokud je možné formulovat nějaké pravidlo se stoprocentní jistotou, budou i data korpusu značkována správně, pokud ovšem není v textu korpusu chyba. Na chyby v textech (např. chybějící slovo či čárka, nesprávná morfologická analýza aj.) je pravidly řízený tagger velmi citlivý, dokáže však některé takové chyby i odhalit. Jelikož je vývoj této metody dosud na počátku, nelze ještě její úspěšnost exaktně kvantifikovat.

**Popis morfologických značek**

* morfologické značky jsou součástí výsledku morfologické analýzy, která pracuje s izolovanými slovními tvary (bez ohledu na kontext). Druhou součástí výsledku je tzv. lemma, které identifikuje příslušnou lexikální jednotku, někdy i včeteně jejího významu, ve smyslu jednoznačné identifikace slovníkového hesla. Morfologická analýza je obecně nejednoznačná; slovní tvary, brány izlované a bez ohledu na kontext, pochopitelně nemohou být v mnoha případech jednoznačně určeny ani z hlediska lemmatu, ani z hlediska morfologické značky.
* **struktura značky:** řetězec 15 znaků, každá pozice odpovídá jedné morfologické kategorii podle víceméně tradičního lingvistického pojetí. Hodnota, která nedává smysl (např. pád u sloves), je reprezentována pomlčkou
  + ***pozice 1 – slovní druh:*** A (adjektivum); C (numerál), D (adverbium), I (interjekce), J (konjugace), N (substantivum), P (pronomen), R (prepozice), T (partikule), X (neznámý), Z (interpunkce)
  + ***pozice 2 – detailní určení slovního druhu*:** slouží především k určení dalších relevantních morfologických kategorií, které jsou uvedeny na dalších pozicí
  + ***pozice 3 – jmenný rod:*** F (feminimum), H (femininum nebo neutrum), I (maskulinum inanimatum), M (maskulinum animato), N (neutrum), Q (femininum singuláru nebo neutrum plurálu), T (maskulinium inaminatum nebo femininum), Y (maskulinum – i. nebo a.), Z („nikoli femininum)
  + ***pozice 4 – číslo:*** D (duál), P (plurál), S (singulár), W, X (libovolné číslo)
  + ***pozice 5 –* *pád:*** čísla 1–7
  + ***pozice 6 – přivlastňovací rod:*** rody mužský neživotný a střední se nikdy nevyskytují samostatně; F (femininum), M (maskulinum animato), X (libovolný), Z („nikoli femininum“)
  + ***pozice 7 –*** ***přivlastňovací číslo:*** P (plurál), N (neutrál)
  + ***pozice 8 –*** ***osoba:*** 1–3; X
  + ***pozice 9 –*** ***čas:*** F (futurum), H (minulost), P (prézens), R (minulý čas), X
  + ***pozice 10 –*** ***grade:*** 1–3
  + ***pozice 11 –*** ***negace:*** A (afirmativ – bez negativní předpony), N (negace – tvar s negativní předponou „ne-“)
  + ***pozice 12 –*** ***aktivum, pasivum:*** A (aktivum), P (pasivum)
  + ***pozice 13 – nepoužito***
  + ***pozice 14 – nepoužito***
  + ***pozice 15 – varianta, stylový příznak apod.:*** d

**Technika**

* texty přicházejí do ÚČNK v mnoha podobách, které je třeba různými softwarovými nástroji zkonvertovat do jednotného formátu
* po konverzi jsou data uložena ve velkém počtu různě velkých souborů. Těm je třeba zajistit dostatek prostoru napevnémdisku, rozumnou dobu přístupu pro všechny oprávněné uživatele a přijatelnou spolehlivost počítačového systému
* z těchto dat se pro vlastní práci s *korpusovým manažerem* automatickygenerují binární soubory, jejichž formát je uzpůsoben rychlému vyhledávání na počítači. Tyto soubory jsou extrémně rozsáhlé a zároveň vyžadují rychlý přístup. Kdykoliv mohou být znovu vytvořeny a proto nemají vysokénároky na zabezpečení
* pro uložení a zpracování zkonvertovaných dat a jejich binárních verzí je vhodný serverový operační systém. Jak bylo uvedeno výše, data jsou zpřístupněna pro lingvistickou práci pomocí speciální sady programů, nazývané ***korpusový manažer*.** Jde o program, který musí splňovat dvě základní kritéria, z nichž prvním je *dostatečná rychlost při vyhledávání požadovaných lingvistických jevů a druhým uživatelsky příjemné rozhraní*
* protože výsledky hledání se obvykle dále zpracovávají, je pro následné zpracování k dispozici další samostatný program. Celý korpusový manažer tak tvoří tři samostatné moduly:
  + nástroje na vyhledávání (implementované v jazycích *C*, *C++* a *Perl*)
  + komunikační program, který zaznamenává výsledky zadaného vyhledání a dále je upravuje, popř. třídí (program je rovněž implementován v jazycích *C*, *C++* a *Perl*)
  + vlastní uživatelské rozhraní (implementované v jazyku *Tcl/Tk*), které umožňuje zadávat dotazy do korpusu, provádět další operace s daty a zobrazovat výsledky těchto akcí.
* všechny tři uvedené programy mohou být spuštěny na jednom počítači, ovšem obvykle jsou první dva spuštěny na serveru a poslední je provozován lokálně na osobním počítači uživatele
* počítače, na kterých se v současné době zpracovává ČNK, se dělí do dvou kategorií:
  + první jsou tzv. *pracovní stanice*, tj. počítače, na kterých pracují jednotliví uživatelé
  + druhou tvoří *servery* – centrální počítače sloužící všem uživatelům

**GCQP**

* speciálně vyvinutý vyhledávácí program, s jehož pomocí se pracuje v korpusu
* nejdříve využíván program CQP (Crpus query procesor), který byl vytvořen na univerzitě ve Stutgartu, umožňoval přístup pouze příkazovým řádkem v operačním systému UNIX -> pouze pro úzký okruh uživatelů. Vyvinuli tedy CQM (Karel Skoupý). Nevýhoda: nešlo zjistit, z kterého textového zdroje výrazy pocházejí; neumožňoval třídit konkordanční řádky ani neposkytoval statistické charakteristiky
* dnes GCQP (Grafické rozhraní pro CQP), autor Pavel Rychlý
* využívá možností grafických operačních systémů (Windows, Mac OS)
* výhody:
  + možnost vyhledávat posloupnosti několika slov
  + vyhledávání podle morfologických značek a základních tvarů (lemmat)
  + třídění konkordančních řádků
  + poskytování informací o tom, z kterých textů pocházejí jednotlivé konkordanční řádky s výskytem hledaného výrazu
  + možnost uložit označené konkordanční řádky na disk počítače, na němž uživatele pracuje
  + statistické funkce
  + možnost vytvářet složitější dotazy
* konkordanční řádek = výsledek „dotazu“, který se zadává do dotazového řádku
* v menu konkordance se dá dále pracovat s vyhledávanými konkordantními řádky – redukovat jejich počet, abecedně je třídy apod.
* P-filtr, N-filtr: slouží k redukci konkordančních řádků případě, že vyhledáváme podle části slova a v konkordančním řádku se nám zobrazí slovní tvary, které jsme vyhledat nechtěli
* regulární výrazy:
  + tečka: libovolný znak
  + hvězdička: libovolný počet opakování předchozího znaku nebo výrazu
  + plus: představuje 1 nebo více opakování předchozího znaku nebo výrazu
  + otazník: představuje žádný nebo jeden výskyt předchozího znaku nebo výrazu
* KWIC: absolutní frekvence hledaného výrazu; základní statistika, kterou získáme při každém vyhledávání a která udává počet výskytů hledaného výrazu x v celém korpusu
* Kontext: uživatelm zadaný počet pozic před či za hledaným výrazem
* Výsledky vyhledávání kolokací obsahují hodnoty tzv. MI-score a T-score.
* **Mi-score** – vzájemná informace (mutual information); je hodnotou, která udává pravděpodobnost toho, že dvě slova /či více/ se vyskytnou současně vedle sebe;
* **T-score** (míra kontrastu) je testem náhodného („random“) rozložení výskytů (předpoklad toho, že korpus je tzv. random sample; jde o složitou otázku)
  + čím vyšší je hodnota MI-score, tím pravděpodobnější je, že jde o pevnou, ustálenou kombinaci.
  + vysoká hodnota T-score naznačuje, že rozložení frekvencí není náhodné, tzn. že jsou zde tzv. clustering effects, např. to, že jde o pevnější spojení (slova jsou více u sebe, než by byla, kdyby bylo jejich rozložení nezávislé jedno na druhé, tj. zcela n
* frekvenční distribuce: spočítá frekvence slov (lemmat), morfologických značek a jejich posloupností v zadaném intervalu. Umožňuje zvolit více intervalů při jednom zadání. Frekvence jednotlivých intervalů lze sčítat, případně skrýt zobrazení výsledku některého ze zadaných intervalů. Na rozdíl od výpisu výsledků Kolokace je ve výpisu výsledku Frekvenční distribuce zobrazena celá kolokace.

**Budoucnost ČNK**

* Na základě ČNK vzniknou nové popisy češtiny
  + nová mluvnice, výkladový slovník, další typy speciálních slovníků, jazykovědné studie, různé příručky a učebnice. Konkrétním výstupem korpusu SYN2000 je mj. také *Frekvenční slovník psané češtiny*.
  + Před dokončením je frekvenční slovník pražské mluvené češtiny, který vznikl na základě PMK. Chystá se příručka cvičení a úkolů pro využití ČNK studenty. Před dokončením je nový manažer Bonito.
* Plánuje se:
  + SZ období totality; autorské slovník předních českých autorů; paralelní korpusy

**Co v korpusu nenajdeme**

* korpus není elektronická encyklopedie (přesto se spoustu věcí dá zjistit z novinových textů)
* korpus není výkladový slovník (přesto lze významy vyvodit z kontextu)
* korpus není seznam českých slov (nezachycuje izolované jevy)

**Co v korpusu najdeme:**

* zjistíme frekvence – např. které varianty jsou používanější (ačkoli, ačkoliv), srovnávat počeštěná a cizí slova (talent, nadání), stylistická pozorování (v jakém kontextu se co vyskytuje)

**Definice korpusu**

*Korpus je soubor počítačově uložených textů (v případě mluveného jazyka - přepisů záznamu mluvy), který primárně slouží k jazykovému výzkumu. Je to strukturovaný, unifikovaný (často i označkovaný) rozsáhlý soubor jazykových dat, který je elektronicky uložený i zpracovaný. Má čtyři základní rysy: vzorky a reprezentativnost; konečná velikost; strojově čitelná forma (MRF), standardní reference.*

**Jak se liší korpus od sbírky textů?**

*Korpus se skládá ze vzorků (samples), které jsou vybrány dle předem stanovených kritérií tak, aby reprezentovaly jazyk. Korpus je rozsahově i obsahově vymezen a omezen, je uložen v elektronické podobě a obsahuje standardní retence. Tato kritéria sbírka textů splňovat nemusí.*

**Typologie korpusů**

*Malé, střední, velké*

*Korpusy monitorovací, paralelní, studijní, cvičné a testovací*

*Lingvistické členění: psané (textové), mluvené; synchronní a diachronní*

**Známé korpusy**

*Brown: první korpus, obsahuje americkou angličtinu z roku 1961, vznikl na Brown univerzity v roce 1964; doplněno gramatické značkování v r roce 1979; asi 1 milion slov;*

*Susanne: autor Geoffrey Sampson, část korpusu Brown; nové gramatické a syntaktické značkování*

*Britis National Corpus (BNC): nejprestižnější projekt anglické lingvistiky, obsahuje britskou angličtinu, 10 % tvoří mluva; první velký korpus pro lexikografy; využívali ho vydavatelé slovníků a univerzity; obsahuje angličtinu z let 1991 až 1994; asi 100 milionů slov; gramatické značkování bylo prováděno automatickým nástrojem; pracuje s velkými a tematicky širokými oblastmi; dělí literaturu na informativní a imaginativní*

*Bank of English (BoE): britská angličtina, University of Birmingham; „hotový“ v roce 1991, ale stále jej rozšiřuují, v roce 2005 obsahoval asi 525 milionů slov; přístupný pro dotazování*

*Korpusy slovenské, maďarské, chorvatské, americké*

*Český národní korpus*

*Pražský mluvený korpus, Brněnský mluvený korpus*

*Korpus Desam**vytvořený na FI, obsahuje asi 1 milion slov, ručně značkovaný (desambiguovaný)*

**Český národní korpus**

*Vznikl v roce 1994, vytvářen Ústavem Českého národního korpusu na FF UK, vedený Františkem Čermákem; akademický, nekomerční projekt*

*Části: SYN2000 (100 000 000 slovních tvarů), DIAKORP (1 750 000 tvarů), ORAL-PMK (700 000 tvarů)*

**Žánrové složení korpusů psaného jazyka ČNK a pojetí synchronního korpusu**

*Publicistika: 60 %; po roce 1990*

*Odborné texty: 25 %; po roce 1989*

*Krásná literatura: 15 %, po roce 1990 (doplňující kritéria vzhledem k přetiskům a čtení knih starších autorů; do synchronního korpusu se zařazují i současně čtení starší autoři, kteří se narodili roku 1880 a později a knihy publikované od roku 1945)*

**Pražský mluvený korpus**

*První korpus mluvené češtiny, zachycuje autentickou mluvenou češtinu, hlavně obecnou a tematicky nespecializovanou. Materiály pochází z Prahy a jejího okolí, kde pracuje spousta obyvatel z celé ČR. Praha rovněž ovlivňuje zbytek republiky mediálně, proto v ČR není místo, které by bylo více reprezentativnější. Nahrávky pocházejí z roku 1988-1996.*

**Brněnský mluvený korpus**

*První mluvený korpus češtiny z oblasti Moravy. Zaznamenává autentickou tematicky nespecializovanou mluvu města Brna. BMK je elektronickým přepisem 250 anonymních magnetofonových nahrávek z let 1994-1999 zachycujících 294 mluvčích.*

***Rozdíl mezi PMK a BMK***

*BMK se pokouší nahradit tradiční interpunkci interpunkcí pauzovou. Striktně zachycuje simultánnosti dialogických promluv. Pauzová interpunkce je náročnější -> diference zápisu u přepisovatelů; zatím bez morfologického značkování. Praha: do jisté míry ukazuje skutečnost.*

**Korpusový manažer**

*Program umožňující efektivně pracovat s počítačovým korpusem, tj. vyhledávat podle zadatelných kriterií (slovní tvar, značka, lemma) ve formě KWIC, vyhledané informace třídit a statisticky zpracovávat, vytvářet subkorpusy, ukládat získané informaci, využívat standardních statistických metod pro vyhledávání kolokací atd. Program musí splňovat dvě základní kritéria, z nichž prvním je dostatečná rychlost při vyhledávání požadovaných lingvistických jevů a druhým uživatelsky příjemné rozhraní. Např. GCQP nebo Bonito, který je vyvíjen a ověřován.*

**Manatee**

*Manatee je knihovna napsaná v jazyce C++, výkonný korpusový manažer. Jedná se o aplikaci provádějící operace nad samotnými korpusy. Neposkytuje však uživatelské rozhraní ani neřeší správu uživatelů. K tomuto účelu slouží aplikace Bonito 1, Bonito 2 a nově aplikace Kondor.*

**Bonito1 a Bonito 2**

https://dip.felk.cvut.cz/browse/pdfcache/cokrtvac\_2011bach.pdf

*ÚČNK používá dva korpusové manažery – Bonito 1 a Bonito 2.*

*Bonito 1 je desktopová aplikace napsaná v jazyce Tcl/Tk. Skládá se ze dvou částí, jménem Bonito 1 se označuje jeho klientská část. Ta komunikuje se serverovou částí ManateeSRV přes TCP/IP. K přenosu dat aplikace implementuje svůj vlastní stavový textový protokol. ManateeSRV se stará o autentizaci a autorizaci a má přístup k databázi. K samotné práci s korpusy využívá knihovnu Manatee, komunikace s ním probíhá přes standardní vstup a výstup.*

*Výhody a nevýhody vyplývají z podstaty desktopové aplikace. Uživatelské rozhraní je interaktivní, s rychlou odezvou. Bonito 1 má také širokou funkcionalitu, včetně podpory statistik, vytváření subkorpusů, nebo dokonce graﬁcké tvorby dotazů. Nevýhodou je, že aplikace není platformově nezávislá (ale existuje verze pro OS Windows, Linux i Mac OS). Aplikaci je také nutno na klientském počítači nejdříve nainstalovat, z čehož plynou problémy s případným zaváděním nových verzí. Také GUI postupem času zastaralo a jeho vzhled neodpovídá současným požadavkům.*

*Z výše zmíněných důvodů vznikala nová verze klienta - Bonito 2. Bonito 2 je webová aplikace napsaná v jazyce Python. Na klientském počítači je tedy potřeba jen webový prohlížeč. Na straně serveru běží CGI skripty, které obsluhují příchozí HTTP požadavky a následně vygenerují HTTP odpověď - čistou html stránku bez jakékoli klientské logiky . Ke komunikaci s Manatee využívá Bonito 2 rozhraní vygenerované nástrojem SWIG. Struktura aplikace je naznačena na obrázku 2.2. Právě neinteraktivní GUI (při každé akci se musí načíst nová stránka) spolu s menší funkcionalitou způsobily, že uživatelé raději nadále využívají starší Bonito 1.*

**Kondor**

*Neúspěch Bonita 2 dal podnět ke vzniku nového manažeru s pracovním názvem Kondor. Jedná se také o webovou aplikaci. Jako programovací jazyk byl zvolen jazyk Java4, který je standardem po tvorbu podnikových aplikací. Na rozdíl od svého předchůdce však poskytuje interaktivní GUI podobné desktopové aplikaci, což je umožněno použitím technologie AJAX. Aplikace Kondor je stále ve vývoji, její základní komponenty včetně dotazování a správy uživatelů jsou již implementovány. Chybí však podpora pro výpočet statistických funkcí.*

*\*AJAX (Asynchronous JavaScript and XML) je obecné označení pro technologie vývoje interaktivních webových aplikací, které mění obsah svých stránek bez nutnosti jejich opětovného načtení. Z JavaScriptového kódu se na pozadí zašle HTTP dotaz na server a jeho odpověď je opět zpracována JavaScriptem. Na rozdíl od klasických webových aplikací poskytují AJAXové aplikace uživatelsky příjemnější prostředí, ale vyžadují použití moderních webových prohlížečů.*

**Značkovací jazyk**

*Značkovací jazyk je jakýkoli jazyk, který vkládá do textu značky vysvětlující význam nebo vzhled jednotlivých jeho částí. Text obohacený o dodatečné informace (o významu, struktuře, způsobu zobrazení jednotlivých částí textů), se původně používal jen pro formátování textu v nakladatelstvích - dodnes např. formátovací jazyk TeX (formátování knih do tisku). Dalšími jazyky jsou troff, PDF. Nejznámějšími značkovacími jazyky jsou HTML a XML, v nichž je vytvořena většina WWW stránek. Pro potřeby korpusové lingvistiky se používá SGML jazyk, dnes XML.*

*Pro potřeby KL se používal jazyk SGML, dnes XML.*

**Metody značkování a rozdíly mezi nimi**

*vnější anotace: typ textu, autor, rok (lze pak vybírat texty z určitého roku, texty psané jen ženami apod.)*

*vnitřní anotace: strukturní informace (členění na kapitoly, odstavce, věty, slova) a informace lingvistické, ty jsou nákladné => omezuje se na morfologické značkování jednotlivých slovních tvarů (tagování), přiřazení slovnědruhové charakteristiky a lemmatizaci (přiřazení základního, slovníkového tvaru)*

**Ruční značkování korpusu**

*Korpus značkují ručně tzv. anotátoři, zkušení a vyškolení pracovníci. Ruční značkování korpusu je nákladná věc, využívá se hlavně kvůli tzv. desambiguaci, tj. zjednoznačňování víceznačných jednotek. Dále se někdy používá při pročišťování korpusu – odstraňování překlepů. Dále se využívá automatického značkování, které provádějí počítačové nástroje. Ručního značkování se využívá u trénovacích dat, na jejichž základě pak probíhá automatické značkování.*

**Výhody anotovaných korpusů**

*Anotovaný korpus odrazový můstek pro další lingvistické výzkumy, nedocenitelný test pro lingvistickou teorii. Sestavování trénovacích dat, na nichž se počítačové programy učí automaticky anotovat slova. Na základě anotací se vyhledává v korpusu – např. výskyty pasiva, vidy apod. Anotovaný korpus poskytuje též výchozí statistická data pro pravděpodobnostní zpracování jazyka.*

*Anotovaný korpus přináší lingvistické interpretace na různých úrovních. Slouží k výzkumu jazyka (vytěžování korpusu pro lingvisty) a jednak pro aplikace NLP – budování modelů jazyka na základě dat získaných z korpusů.*

**Automatické nástroje pro značkování korpusu**

*Morfologické analyzátory = nejznámější morfologické analyzátory (značkovací programy - taggers pro angličtinu) zpracovávají data v korpusu tak, že každému slovnímu tvaru přiřadí jeho gramatickou značku (tag), tj. obvykle symbol slovního druhu (může jich být i víc). Obvykle se značkují vybrané části korpusu v rozsahu do 10 mil. slovních tvarů; vzniklé soubory jsou zhruba třikrát až čtyřikrát větší než původní, což znamená, že při jejich dalším zpracování vznikají časové problémy.*

*Probabilistický analyzátor CLAWS (autor R.Garside z Lancasteru): má vysokou úspěšnost, dosahuje jen 1.7% chyb. Celkově je CLAWS hybridní (vedle stochastického přístupu obsahuje i jednoduchá syntaktická pravidla) a pracuje s anotovaným lexikonem, jehož součástí je i seznam základních anglických idiomů. Značkování se provádí v několika fázích, používá se rovněž Viterbiho algoritmu (zpracovává pravděpodobnosti přechodu mezi větnými složkami). Probabilistický přístup je motivován tím, že je blízký psychologii člověka.*

*Analyzátor vytvořený J.Clearem v birminghamském COBUILDU: využívá pravděpodobnostního přístupu, je velmi robustní a jeho míra úspěšnosti je 95% - autor ji pokládá za dostačující.*

*Helsinský analyzátor: založen na tzv. constraint grammars a je 60krát rychlejší než ostatní (předpokládá ale užití dvoustupňového morfologického analyzátoru Kimmo od Koskenniemiho) - je zatím ze všech zjevně nejúspěšnější, pokud jde o zvládnutí více jazyků (dosud dovede pracovat s 5 jazyky).*

*Analyzátor D.Cuttinga et al. (je v public domain a dostupný v Internetu):  
Užívá skrytého Markovova modelu, je jazykově nezávislý, učí se od počátku na menších vzorcích, pracuje s vahami pravděpodobnostního výskytu, pracuje iterativně a ve fázi učení počítá s 18% předem označkovaného textu.*

*Mofrologické analyzátory Ajka a Majka*

*Taggit*

**Obecné zásady vytyčené G. Leechem, podle nichž mají být vytvářeny značky (7)**

*-> zachovat vratnost anotovaného korpusu do surového stavu (autor značek je interpretem, s nímž nemusí každý potenciální uživatel souhlasit, přičemž by mělo být technicky možné se případné nežádoucí interpretace zbavit a moci pracovat bez ní)*

*-> možnost extrahovat anotace z textu a uložit je zvlášť, aby bylo možné se k nim vrátit (formou nějaké relační databáze, nebo interlineárního formátu)*

*-> anotační schéma by mělo vycházet z teoretických východisek, která by měla být jasně formulovaná a přístupná každému konečnému uživateli korpusu. Mnohé korpusy byly anotovány ručně (existence subjektivních interpretací zaviněných osobou anotátora ve sporných případech). Značkování by pak mělo být doplněno komentáři, z nichž by byl důvod příslušné volby patrný.*

*-> mělo by být jasné, JAK a KDO anotaci provedl (JAK – ručně* × *automaticky* × *poloautomaticky, s postkorekcí* × *bez korekce; KDO – počítačový program, anotátor - člověk)*

*-> uživatel korpusu by si měl být vědom toho, že anotace nejsou nějakou nedotknutelnou neomylnou instancí. Anotace je pouze více či méně užitečným nástrojem. INTERPRETACE.*

*-> anotační schéma by mělo být založeno na široce schvalovaných a teoreticky nezatížených principech. Není na škodu i zjednodušující přístup.*

*-> žádné anotační schéma nemá právo být pokládáno za standardní. Je-li nějaké řešení uznávanější, děje se tak pouze z praktických důvodů.*

**Word, lemma , tag, tagování, token**

*Word je poziční atribut podporovaný každým korpusem. Jde o takový tvar slova, jaký se nacházel v předloze korpusu. Pro již zmíněnou konkordační řádku „Mýt v zimě auto, či nemýt.“ je hodnota atributu word na pozici -1 „zimě“.*

*Lemma je poziční atribut obsahující základní tvar slova. Např. pro podstatné jméno je to 1. pád jednotného čísla, pro sloveso inﬁnitiv, ... Pro již zmíněnou konkordační řádku „Mýt v zimě auto, či nemýt.“ je hodnota atributu word na pozici -1 „zima“*

*Tag je poziční atribut obsahující informaci o slovním druhu a dalších morfologických charakteristikách daného slova. Pro již zmíněnou konkordační řádku „Mýt v zimě auto, či nemýt.“ je hodnota atributu word na pozici -1 „NNFS6——A—–“.*

*Tagování je značkování na úrovni slovních druhů*

*Tokeny jsou testová slova, tj. výskyt slovních tvarů (slovních exemplářů), čísel, zkratek, speciálních znaků a interpunkčních znamének. Základní prvek korpusu (pozice)*

*Tokenizace rozdělení na jednotlivé pozice*

*Typ: slovní tvar jako takový*

*Př.: Token (počet všech unaků v korpusu, tj. i interpunkce) a type (počet všech různých slov) => např. krávy krávy, jak si vlastně povídáte => token = 7, type = 5)*

*Konverze převod textu do nějaké podoby*

**Co je to CQL?**

*Corpus Query Language je jazyk sloužící k dotazování v korpusovém manažeru. S jeho pomocí speciﬁkujeme, jaká slovní spojení v korpusu hledáme. V nejjednodušší formě může vypadat dotaz např. takto: [word="auto"], což znamená, že hledáme všechny výskyty řetězce „auto“ v korpusu.*

**Co je to KWIC?**

*KWIC - Keyword in context je konkrétní vyhledané spojení, které odpovídá zadanému dotazu (CQL). Pokud máme např. dotaz [word="auto"][][word="koně"] (tři slova, z nichž první je „auto“ a poslední je „koně“), pak KWIC může být třeba „auto přejede koně“ nebo „auto, koně“.*

**Konkordanční seznamy a jejich využití**

*Konkordance je množina klíčových slov nebo slovních spojení spolu s jejich kontextem vyhledaných jako odpověď na uživatelem zadaný dotaz. Konkordance je zpravidla zobrazována v tabulce se třemi sloupci. Uprostřed jsou vyhledaná spojení, odpovídající dotazu.*

*Konkordační seznam je výsledek dotazu na korpus; zobrazuje se ve formátu KWIC, kdy jsou hledaná slova se svými kontexty zobrazena přehledně pod sebou.*

**Frekvenční seznamy a jejich využití**

*Frekvenční distribuce“ slouží ke spočítání frekvence slov (lemmat), morfologických značek a jejich posloupností na zadaných pozicích (počítáno od KWIC). Jako výsledek tedy zobrazí tabulku se slovy (pozičními či strukturními atributy), která se nejčastěji vyskytují na daných pozicích.*

**Využití korpusů při tvorbě slovníků**

*Dříve sběr dat založen na lístkových katalozích sestavovaných z příkladů lexikálních jednotek. Korpusy umožňují vyhledávání zadané lexikální jednotky během několika vteřin. To umožňuje rychlé obohacování a rozšiřování slovníků; definice a výklad mohou být co nejbohatší až kompletní; příklady se mohou rychle řadit do smysluplných podskupin; při tvorbě slovníků se používá otevřených monitorovacích korpusů; ohromný význam u terminologických slovníků; souvýskyt + možnost jeho srovnání -> přesná definice termínu. Velké využití u frekvenčních slovníků, různých učebnic, jazykovědných studií, příruček. Na základě PMK vznikl frekvenční slovník pražské mluvené češtiny.*

**Desambiguace**

*zjednoznačnění víceznačných jednotek v kontextu*

**Jaký je rozdíl mezi kognitivně plausibilním a kognitivně neplausibilním systémem?**

*Kognitivně plausibilní systém usiluje o vytvoření poznávacího (kognitivního) modelu, pro nějž je relevantní, jakým způsobem člověk řeší nějaký úkol pomocí inteligence, a používá jej jako bázi pro stroj, který má tento problém (úkol) inteligentně řešit. Takový systém často používá úplnou sadu pravidel, které explicitně formulují znalosti, které člověk implicitně používá, ve formě tzv. báze znalostí (knowledge base).*

*Naopak systémy, které rezignují na kognitivní plausibilitu a jsou tedy kognitivně neplausibilní, se prostě snaží vytvořit model inteligentního chování, aniž by se přihlíželo k tomu, zda systém pracuje stejným způsobem jako člověk (inteligentně). Tyto systémy často používají surová kvantitativní data k vytvoření statistického modelu napodobujícího lidské chování.*

**Jaký je rozdíl mezi stochastickou disambiguací morfologicky označkovaného korpusu a disambiguací řízenou pravidly?**

*Stochastická desambiguace: model založený především na pravděpodobnostech přechodu mezi jednotlivými značkami v morfologicky analyzovaném textu. Princip: nejdřív ručně označkovat tzv. testovací data (tvaru přiřadit jednu správnou značku), z nich se pak program „naučí“ na trénovacích datech a učiní si představu o pravděpodobnostech přechodu mezi jednotlivými značkami a o jejich četnostech, kterou si uloží do vnitřních tabulek. Pak tyto poznatky aplikuje při desambiguaci korpusu, který není desambiguovaný.*

*Pravidly řízená desambiguace: stochastická metoda není 100% úspěšná, proto se začala vyvíjet metoda založení na syntaktických pravidlech. Podstata = intuitivní formulace celé řady syntaktických pravidel, která odrážejí syntaktické konfigurace čečiny dané jejím vnitřním systémem. Vývoj této metody je v počátcích. (viz výše). Přístup kognitivně plausibilní.*

**K čemu může sloužit kvantitativní analýza dat získaných z velkých jazykových korpusů?**

*NLP: stochastická desambiguace, automatická analýza založená na statistických metodách*

*Frekvenční seznamy, výzkum kolokací - lexikografie*

**Kolokace a korpusová lingvisitka**

*Kolokace jsou slova (nebo i jiné poziční atributy), která se statisticky významně často vyskytují v korpusu pohromadě. Statistika „Kolokace“ nám tedy zobrazuje, jaká slova se nejčastěji vyskytují v blízkosti KWIC spolu s hodnotami specializovaných kolokačních funkcí*

*T-score - míra kontrastu: Vychází ze statistické metody testování hypotéz pomocí tzv. t-testu.*

*V případě kolokací testujeme, zda zjištěné počty výskytů jednotlivých slov a jejich dvojic odpovídají náhodnému rozložení slov v korpusu. Čím vyšší je hodnota t-score, tím méně je pravděpodobné, že jde o náhodné rozložení slov a a naopak tím pravděpodobnější je, že jde o pevnější, ustálenější kombinace slov, tj. o kolokace.*

*Mi-score*

**NLP (Natural langue processing)**

*počítačové zpracování přirozeného jazyka. Je to obor na pomezí lingvistiky, informatiky (umělé inteligence), popř. taky akustiky aj.; zkoumá problém analýzy či generování textů nebo mluveného slova, které vyžadují určitou (ne absolutní) míru porozumění přirozenému jazyku strojem. Zabývá se zkoumáním a zpracováním přirozeného jazyka pomocí počítačů a metod informatiky.*

*Mezi hlavní úkoly patří: analýza přirozeného jazyka (morfologická, syntaktická, sémantická); generování přirozeného jazyka (vyvíjení, vytváření); syntéza a rozpoznávání řeči; strojový překlad (Machine translation); odpovídání na otázky (Question answering); získávání informací (Information retrieval); korektura textu (Spell-checking, Grammar checking); extrakce informací (Information extraction); výtah z textu (Text summarization); určení typu dokumentace (Text Classification/Clustering)*

**MRF (machine readable form)**

*soubor počítačově čitelných textů, složený ze souvislých textových úseků vybraných dle jistých pravidel tak, aby reprezentovaly jazyk jako celek v celé jeho pestrosti -> obsahuje standardní reference*

*anotace - ty zahrnují metatextové informace a interpretace jednotek, z nichž je text složen*

**OCR (optical character recognition)**

*optické rozpoznávání znaků je metoda, která pomocí scanneru umožňuje digitalizaci tištěných textů, s nimiž pak lze pracovat jako s normálním počítačovým textem. Počítačový program převádí obraz buď automaticky, nebo se musí naučit rozpoznávat znaky. Převedený text je téměř vždy v závislosti na kvalitě předlohy třeba podrobit důkladné korektuře, protože OCR program nerozezná všechna písmena správně.*

**TEI (Text encoding initiative)**

*\*1986. Aktivita sponzorovaná hlavními vědecky orientovanými asociacemi zabývajícími se využitím počítačů v humanitních vědách. ACL (Association for Computional Linguistics), ALLC (the Association for Literary and Linguistic Computing), ACH (the Association for Computers and Humanities). Je sponzorovaná EU a americkou vládou;*

*Cíl: vytvoření standardní implementace pro operace s počítačově čitelnými texty. Návrhla společný kódovací a značkovací metajazyk SGML – přijat pro svou jednoduchost, jasnost, formální přísnost a mezinárodní uznání; vydala doporučení pro společný výměnný formát, zásady kódování, znakové sady*

**DDL (Data Driven Learning)**

*Výuková metoda studentů, která se zakládá na vyhledávání v korpusu. Otec Tim Johns (univerzita v Birminghamu). Vytvořit takový postup výuky, který byl založený na induktivní metodě, tedy na aktivním vyhledávání jedné jazykové jednotky v korpusu. Studenti tak pomocí korpusu shromáždí data, které jim dávají příležitost objevit příklady užití jazyka. Z nich poté zobecní pravidlo. DDL je možné využít při skupinové nebo individuální výuce i mimo tradiční výuku, např. ve formě e-learningu nebo jako domácí cvičení.*

*Student je tedy naveden k tomu, aby sám na základě materiálu získaného z korpusu tvořil hypotézy o různých významech lexikálních, gramatických. Rozvíjí deduktivní přístup, což napomáhá dalšímu studiu jazyků.*

*Hlavní výhodou využití korpusu pro výuku jazyků je práce se skutečným materiálem, reálnými větami a slovními spojeními, které nejsou upravené pro didaktické účely.*

**DTD (Document Type Definition)**

*Definice typu dokumentu; jazyk pro popis struktury XML, případně SGML dokumentu. Omezuje množinu přípustných dokumentů spadajících do daného typu nebo třídy. DTD tak například vymezuje jazyky HTML a XHTML. Struktura třídy nebo typu dokumentu je v DTD popsána pomocí popisu jednotlivých značek (nebo též elementů) a atributů. Popisuje, jak mohou být značky navzájem uspořádány a vnořeny. Vymezuje atributy pro každou značku a typ těchto atributů. Poměrně starý a málo expresivní jazyk. Formální reprezentace, informuje uživatele nebo program o tom, které elementy text obsahuje, jak jsou tyto elementy kombinovány a obsahuje také sadu deklarací entit. Používá se programem SGML parser – kontroluje, zda je text otaggován ve formátu kompatibilním s TEI*

*<!ELEMENT clovek (jmeno, adresa\*)****>***

*<!ELEMENT jmeno (#PCDATA)****>***

*<!ELEMENT adresa (ulice?, cislo?, mesto)****>***

*<!ELEMENT ulice (#PCDATA)****>***

*<!ELEMENT cislo (#PCDATA)****>***

*<!ELEMENT mesto (#PCDATA)****>***

**PDT (Prague Dependency Treebank; Pražský závislostní korpus)**

*Pražský závislostní korpus (PDT) je probíhající projekt pro ruční anotaci velkého množství českých textů bohatou lingvistickou informací, sahající od morfologie přes syntax až po sémantiku/pragmatiku a ještě dále.*

*Představuje druhý největší, ručně označkovaný korpus na světě (hned za anglickým Penn Treebank (1992))*

**EAGLES**

*dozorčí skupina založená EU, jejím úkolem je sledovat a pomáhat různým evropským iniciativám*

*jde o vytvoření systému značek, který by na jedné straně zachytil všechny zvláštnosti, potřeby, specifika všech evropských jazyků - na straně druhé zachoval jednotu systému*