**Počítačová a korpusová lingvistika**

*1. Počítačová lingvistika (komputační – formální – kvantitativní lingvistika, NLP, language engineering). Vývoj ve světě a v ČR.*

*2. Počítačové porozumění přirozenému jazyku a počítačové zpracování jednotlivých rovin přirozeného jazyka (morfologická, syntaktická, sémantická, slovotvorná).*

*3. Korpusová lingvistika – empirická věda. Historie korpusové lingvistiky. Definice korpusu v moderním slova smyslu. První korpusy. Chomského kritika korpusové lingvistiky.*

*4. Typy korpusů (kritéria – časové hledisko, druh zachycené komunikace, rozsah, žánrové hledisko, účel, jazyk). Specializované korpusy. Žákovské korpusy.*

*5. Budování korpusů. Získávání textů. Přepis, OCR metody, konverze dat. Značkování korpusů (strukturní značky, atributy). Metatextové anotace. Sociolingvistické značkování korpusů mluvených a korpusu soukromé korespondence.*

*6. Lingvistické značkování. Tokenizace, lemmatizace, automatická morfologická analýza, disambiguace (statistické metody, pravidlová disambiguace, hybridní metody). Lingvistické problémy grafického zachycení korpusů mluveného jazyka (přepis a pravidla přepisu v mluvených korpusech budovaných v rámci ČNK). Stručná charakteristika olomouckého mluveného korpusu.*

*7. Korpusy v českém prostředí. ČNK – historie a současnost. Charakteristika jednotlivých korpusů pod ČNK (korpusy řady SYN, projekt DIAKORP a další). Charakteristika korpusů budovaných na FI MU (webové korpusy).*

*8. Korpusové manažery a další nástroje pro práci s jazykovými korpusy. Volba korpusu vhodného pro řešení lingvistického problému. Morfologické značkování českých korpusů (dva analyzátory – pražský a brněnský, poziční a atributový systém, shody a odlišnosti, porovnání tagsetů).*

*9. Značkování syntaktické roviny jazyka. Pražský závislostní korpus. Projekty VALLEX a VERBALEX.*

*10. Lexikální významy – slova a kolokace z hlediska automatických nástrojů, využití korpusů, elektronické slovníky (obsah, budování), lexikální databáze (WordNet, EuroWordNet, thesaury).*

# **1. Počítačová lingvistika (komputační – formální – kvantitativní lingvistika, NLP, language engineering). Vývoj ve světě a v ČR**

* ***komputační (počítačová) lingvistika* –** stojí na pomezí lingvistiky, matematiky a informatiky, někdy se ztotožňuje s matematickou lingvistikou, nebo se považuje za její součást, v užším významu zkoumání zpracování přirozeného jazyka pomocí počítačů, počítačových metod a informatiky
* ***matematická lingvistika*** – popis fungování přirozeného jazyka za pomocí metod matematiky a matematické lingvistiky
* ***kvantitativní lingvistika*** – zkoumání přirozeného jazyka za pomocí matematické statistiky, teorie pravděpodobnosti a teorie informace
* ***formální lingvistika***
  + předmětem je exaktní a explicitní zpracování přirozeného jazyka pomocí matematických/informatických metod
  + usiluje o co nejpřesnější popis jazykových dat a struktur ve všech plánech
  + snaha modelovat jazyk pomocí logické formy a metod matematické logiky
  + počátky jsou spojeny s vývojem popisu přirozených jazyků v 50. letech a rozvojem matematické teorie formálních jazyků a automatů
  + s rozvojem jazykových korpusů a vznikem internetu od 80. let vzrůstá množství jazykových dat v počítačové podobě a nutnost jejich zpracování
* ***NLP*** – počítačové zpracování přirozeného jazyka, zkoumá problematiku analýzy a generování textů mluvného slova
  + významné úkoly:
    - syntéza řeči
    - rozpoznávání řeči
    - generování přirozeného jazyka
    - strojový překlad
    - odpovídání na otázky
    - vyhledávání informací
    - extrakce informací – výpis strukturovaných nebo polostrukturovaných informací z nestrukturovaných strojově čitelných dokumentů, např. rozpoznávání pojmenovaných entit, koreference (jmenné fráze odkazující na identický objekt)
    - korektura z textu
    - výtah z textu – automatická sumarizace
    - dolování z textu – text mining, automatické zpracování velkého množství informací, např. určování druhu textů, shlukování textu, analýza sentimentu
  + **Typická struktura NLP systému:** vstupní věty jsou nejprve podrobeny **lexikální analýze** využívající slovníku, který obsahuje znalosti o významech slov. Pak **morfologické** a **syntaktické analýze** opírající se o množinu pravidel definujících přípustné syntaktické struktury, tedy o gramatiku: to vše v modulu, který se obvykle nazývá parser (analyzátor). Získané syntaktické reprezentace jsou pak **sémanticky interpretovány** a výsledkem jsou sémantické reprezentace – zde, jak patrno, v PK1.

**Jazykové inženýrství (Language engeneering)**

* nová disciplína, která představuje kombinaci lingvistiky a počítačové vědy a zaměřuje se hlavně na tvorbu programového vybavení pro zpracování PJ

**Aplikace v oblasti PJ**

1. Programy pro zpracování textů v PJ

* jazyková podpora na úrovni textových procesorů
  + **pravopisné korektory** (angličtině spell checkers)
    - kontrola překlepů a pravopisu; červené podtrhávání (*grmatický*)
  + **gramatické korektory** (grammar checkers)
    - všímá si syntaktický chyb; používá morfologický analyzátor v první fázi, jeho jádrem je syntaktický analyzátor
    - modré podtrhávání (***aby jste****, programy pracoval****i****, pravopisn****á*** *korektor*)
    - úspěšnost asi 70 procent
    - kontrola české gramatiky
      * gramatický korektor **T. Jelínka, P. Květoně, K. Olivy a V. Petkeviče**
      * v Microsoft Word
      * 2005, ÚJČ AV ČR
  + stylistický korektor
  + dělicí programy
* vyhledávací (fulltextové) programy založené na lemmatizaci (tj. morfologické analýze)
* programy pro strojový překlad z jednoho jazyka do druhého, obvykle jen pro určité typy textů a experimentální povahy, kvalita překladu nebývá vysoká
* prohlížecí programy (browsers) využívající jednoduché morfologické analýzy a klíčových slov, prohlížení e-mailu, dokumentů na WWW

2. Dialogově orientované aplikace, např. dotazovací systémy pro přístup k datovým bázím, automatizované systémy pro komunikaci (i hlasovou, telefonem) s klienty

* informační systémy na nádražích a letištích
* hlasové ovládání počítačů
  + operační systémy typu Merlin
  + systémy převádějící text na mluvenou řeč (*Text-to-Speech*, *TTS*)
    - v ČR např. *Demosthenes, AUDIS*
  + systémy pro rozpoznávání mluvené řeči (Automatic Speech Recognition Systems, ASRS)
  + diktovací systémy
    - *Via Voice* IBM; *Dragon* apod.

3. Porozumění příběhům **(story understanding)**

* systémy, které dovedou porozumět novinovým článkům a vytvářet z nich souhrny a abstrakty

**Vývoj počítačové lingvistiky**

* po 2. sv. válce vznik pomezních disciplín lingvistiky, mj. také matematická lingvistika – využívání metod přírodních věd (statistika, algebraické metody)
* 70. léta – **vznik frekvenčních slovníků – předzvěst korpusů**, 1961 – Frekvence slov, slovních druhů a tvarů v českém jazyce – Jelínek, Bečka, Těšitelová
* 40./50. léta – vznik kybernetiky, studium přenosu informace (Claude Shannon, Warren Weaver, Charles Hockett)
* rozvoj teorie informace – jak měřit množství informace, užití statistických metod, navázala kybernetika a strojová lingvistika
* Norbert Wiener – zakladatel kybernetiky, matematik
* 2. pol. 50. let – algebraická lingvistika – využití nekvantitativních matematických metod v lingvistice
* od 50. let 20. století rozvoj **strojového překladu**
* první počítače
  + Charles Babbage – děrné štítky pro řízení tkalcovských strojů
  + 0. generace – elektromagnetické relé – Zuse Z4
  + 1. generace – ENIAC – elektronky
  + 2. generace – tranzistory
  + 3. generace – integrované obvody – IBM 360
  + 4. generace – mikroprocesory (od 1981)
* **strojová lingvistika** – od 60. let 20. stol – formální popisy jazyka srozumitelné pro počítače, rozvíjí se jednotlivé oblasti počítačové lingvistiky
  + automatické zpracování informací a textu
  + uložení a opětovné nalezení informace
  + klasifikace textů
  + uložení dat – ASCII, OCR
  + information retrieval
* **dialogové systémy** – ELIZA – 1964–66 – Joseph Weizenaum, simulace dialogu pacienta s psychiatrem

**Vývoj počítačové lingvistiky v ČR**

* ***Ústav pro jazyk český ČSAV*** 
  + *Oddělení matematické a aplikované lingvistiky* (1961–1985)
    - založil Lubomír Doležel
    - Marie Těšitelová (Korpus věcného stylu), Jitka Štindlová, Eleonora Slavíčková (Retrográdní morfematický slovník)
* ***Filozofická fakulta UK***
  + Laboratoř algebraické lingvistiky (1968–1972)
  + Oddělení pro teorii strojového překladu (1957)
    - Petr Sgall, Jarmila Panevová, Eva Hajičová
  + Ústav teoretické a komputační lingvistiky (1990)
    - Vladimír Petkevič
  + Ústav Českého národního korpusu (1994)
* ***Matematicko-fyzikální fakulta UK***
  + Ústav formální a aplikované lingvistiky (1990)
* ***Filozofická fakulta MU*** *–* Karel Pala
  + Ústav českého jazyka
    - syntaktický analyzátor klara (Karel Pala, Klára Osolsobě, Stanislav Franc); využití Prologu a aparátu DC gramatik
    - gramatický korektor pro T602 (Osolsobě, Franc)
    - morfologický analyzátor Xantypa (Osolsobě, Franc)
    - morfologický analyzátor LEMMA (Osolsobě, Pavel Ševeček)
    - logická analýza přirozeného jazyka (Pavel Materna, Karel Pala, Jiří Zlatuška)
* ***Fakulta informatiky MU***
  + Centrum zpracování přirozeného jazyka (1998)
    - Internetová jazyková příručka (2008)
    - Sketch Engine
    - CZ accent (nástroj na doplňování diakritiky do textů psaných bez háčků a čárek)
    - morfologie: **Majka**
    - syntax: **Synt** a **Sent**
    - slovníky: DEBDict
    - [​další technické nástroje:](http://nlp.fi.muni.cz/projekty/justext/) 
      * [jusText](http://nlp.fi.muni.cz/projekty/justext/): *nástroj na odstraňování netextových částí webových stránek*
      * [​onion](http://code.google.com/p/onion/): *program na odstraňování duplicitních částí textu*
      * [​chared](http://code.google.com/p/chared/): *program na rozpoznávání kódování textů*
      * [​Gensim](http://radimrehurek.com/gensim/index.html): *systém, který umožňuje určit témata, o kterých se v píše v daném textu*

# **2. Počítačové porozumění přirozenému jazyku a počítačové zpracování jednotlivých rovin přirozeného jazyka (morfologická, syntaktická, sémantická, slovotvorná)**

* ***roviny zkoumání jazyka***
  + *fonetická* – zabývá se zpracováním signálů, jejich tříděním a klasifikací
  + ***fonologická*** – zabývá se funkcí hlásek, zvukovými rozdíly, které mají v jazyce schopnost rozlišovat význam, rozlišuje fonémy – nejmenší jednotky jazyka schopné rozlišit význam, spolu s fonetickou rovinou jde o znalosti podstatné pro systémy založené na rozpoznávání mluvené řeči
  + ***morfologická*** 
    - ohýbání, odvozování slov, studium vztahů mezi jednotlivými částmi slov
    - jak se slova skládají z nejzákladnějších jednotek – morfémů (nejmenší jednotka, která může nést význam)
    - kořeny, kmeny, kmenotvorné přípony, prefixy, sufixy, koncovky
    - algoritmický popis ohýbání je základem pro tvorbu analyzátorů a generátorů tvarů
  + ***syntaktická*** 
    - vztahy mezi slovy ve větě, správné tvoření větných konstrukcí, slovosled
    - jak lze spojovat slova tak, aby z nich vznikaly gramaticky správné věty
    - jak lze reprezentovat vztahy formálně reprezentovat
    - budování syntaktických analyzátorů a generátorů
    - reprezentace v podobě stromových struktur
  + ***sémantická rovina*** 
    - tektogramatická, význam výrazů z různých strukturních úrovní jazyka
    - co jazykové výrazy znamenají a jak se jejich významy kombinují tak, aby tvořily smysluplné věty
    - významy vět nezávisle na kontextu
    - tvorba sémantických analyzátorů – použití predikátové logiky 1. řádu, lépe transparentní intenzionální logiky TIL
  + ***pragmatická***
    - zabývá se promluvami, výpověďmi, individuální interpretací textu
    - jak se vět užívá v různých komunikačních situacích a jak užití vět ovlivňuje interpretaci jejich významu
  + ***kontextová, promluvová rovina*** – jak bezprostředně předcházející věty ovlivňují sémantickou interpretaci vět následujících
  + ***znalosti o světě*** 
    - obecné encyklopedické znalosti potřebné k vedení normální komunikace, common sense
    - složitý komplex znalostí, k nimž se řadí též znalosti o komunikačních záměrech, plánech a vírách ostatních uživatelů jazyka
    - řadí se sem i znalosti a soubory inferenčních pravidel označované jako zásady zdravého rozumu = **common sence**
  + ***jazykové metaznalosti*** – propojují znalosti o světě se znalostí daného přirozeného jazyka
* ***kroky analýzy přirozeného jazyka***
  + *tokenizace*
  + *morfologická analýza* – přiřezaní lemmatu a morfologických kategorií
  + *desambiguace* – zjednoznačnění, výběr nejpravděpodobnější interpretace (tagging)
  + *parsing* – syntaktická analýza
* ***reprezentace a porozumění***
  + klíčová složka spočívá ve vybudování reprezentace významu vět a textu
  + pro reprezentaci významu potřebujeme jiný nástroj než přirozený jazyk – formální jazyk, jehož prvky jsou symboly a na jehož výrazy lze aplikovat
  + reprezentace významu musí být přesná a jednoznačná
  + reprezentace by měla zachycovat intuitivní strukturu vět (výrazů) přirozeného jazyka
  + významy, které jsou parafrázemi, by také měly být k sobě vztaženy prostřednictvím svých reprezentací
  + reprezentace by měla být nezávislá na daném přirozeném jazyce
  + ***reprezentace morfologických struktur***
    - morfologická analýza slov textu
    - deklinační a konjugační vzory – reprezentace pomocí **konečných automatů** a **struktur trie**
  + ***reprezentace syntaktických struktur***
    - stavba vět a jazykových výrazů
    - **stromové struktury**
    - **závislostní grafy**
  + ***významy slov a vět***
    - sémantické vztahy mezi větnými členy
    - potřeba zachytit reprezentace znalostí – systémy založené na **PK1** nebo **TIL**
    - odvozování inferencí

**Morfio**

* nástroj pro studium slovotvorby v češtině
* vyhledává v korpusu podobně utvořená slova podle zadaného vzoru (sufixu či prefixu), umí je i porovnat

# **3. Korpusová lingvistika – empirická věda. Historie korpusové lingvistiky. Definice korpusu v moderním slova smyslu. První korpusy. Chomského kritika korpusové lingvistiky**

* empirie je [zkušenost](https://cs.wikipedia.org/wiki/Zkušenost) získaná [pozorováním](https://cs.wikipedia.org/wiki/Pozorování), případně pokusem ([experimentem](https://cs.wikipedia.org/wiki/Experiment)). Slovo se používá tam, kde chceme zdůraznit cílený a řízený způsob získávání zkušeností. [Empirické vědy](https://cs.wikipedia.org/w/index.php?title=Empirická_věda&action=edit&redlink=1) jsou vědy založené na opakovatelných a ověřitelných experimentech.
* v moderní době je **jaz. korpus** **hlavním zdrojem koncentrované jaz. empirie**, kterou jsme do určité míry (většinou malé) někdy schopni vyvažovat empirií a jaz. zkušeností osobní. Korpus je tak **východiskem k realizaci empirického zkoumání jaz.**

**Korpus**

Rozsáhlý soubor autentických textů (psaných nebo mluvených) převedený do elektronické podoby v jednom formátu tak, aby v něm bylo možné jednoduše vyhledávat jazykové jevy (zejména slova a slovní spojení – kolokace).

Korpus zobrazuje jaz. jevy **v**jejich **přirozeném kontextu** a umožňuje tak vytvářet na reálných datech podložený jaz. výzkum v rozsahu, který byl dříve nemyslitelný.

Jeho hlavní předností je vedle užití přirozeného jaz. materiálu i **schopnost vypovídat o frekvenci** (četnosti) jevů a jejich typickém užití, což je informace jen pomocí badatelovy intuice nezjistitelná. Jelikož do k. vstupují texty jako celek, poskytuje na rozdíl třeba od lístkového katalogu nevýběrové informace o všech typech jaz. jevů.

Termín se v lingvistice objevuje od 50. let a používá se označení pro velký soubor elektronických textů, zpracovávaný a vytěžovaný specifickými softwarovými nástroji v počítači (pro odlišení se tu také někdy mluví o *textových korpusech*).

Korpusje tedy záznamem a **relativně objektivním modelem jaz. empirie** a je a má být při svém zpracování v základním, výchozím smyslu zcela zbavený jakékoliv apriorní teorie a jiných vlivů, které jaz. nutně deformují a které by se do něj mohly promítat, ať je jejich autor míní sebelépe. Využívá ho (vedle jiných oborů) zvláště **empirická lingvistika** za účelem studia a poznání jaz. a v širším smyslu skrze něj i k poznání celkové dobové reality, kterou jaz. zprostředkovává a zachycuje.

Dnešní rozsahy největších **k.**se pohybují řádově už ve stamiliónech a miliardách slov a jimi nabízená informace je, na rozdíl od zdrojů tradiční lingvistiky založené na manuálních excerptech a archivech, tak bohatá, že ji často nelze, bez použití dalších nástrojů, filtrů a zvláštních metod, bez rizika zahlcení informacemi přímo používat.

**Korpus**se liší od běžného elektronického archivu, tj. prostého úložiště daných textů s omezenými možnostmi prohledávání. Oproti archivu k.vyniká nejen sofistikovaným ***korpusovým manažerem*** schopným zpracovat miliardy slov, řadou speciálních programů a obvykle i vnesenou (lingvistickou) **anotací** každého slova, ale i celkovým plánovaným zaměřením na určitý cíl se záměrem ho textově maximálně pokrýt. Obvykle se tu pak mluví o *reprezentativnosti* takového korpusu.

**Word** – slovní tvar

**Lemma** – základní tvar pro nějakou skupinu tvarů (nominativ u substantiv, infinitiv u sloves).

**Token** – výskyt slovního tvaru v korpusu, slovní tvar jako takový; textová slova, tj. výskyt slovních tvarů (slovních exemplářů), čísel, zkratek, speciálních znaků a interpunkčních znamének. Základní prvek korpusu (pozice)

**Tokenizace –** rozdělení na slova, rozčlenění textu na základní jednotky určené pro vyhledávání

**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)

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

**KWIC**

* *key word 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.

**Historie korpusové lingvistiky, první korpusy**

* ***aplikace korpusové lingvistiky***
  + vznik fonetických databází
  + studie z morfologie a syntaxe ovlivněné korpusovým výzkumem
  + vznik frekvenčních a kolokačních slovníků
  + studium diskurzu založené na korpusu
* cizojazyčná pedagogika (tvorba učebnic)
* ***hlavní světové korpusy***
  + *první generace*
    - **Brown Corpus** – 60. léta 20. stol. na Brownově univerzitě v USA, 500 vzorků po 2000 slovech, dodnes slouží jako standard pro tvorbu korpusů, Henry Kučera, W. Nelson Francis
    - Lancaster-Oslo/Berge Corpus – britský protějšek, stejná struktura, Geofrrey Leech
    - The Survey of English Usage – Randolph Quirk
      * na prvním korpusovém pracovišti University College London
      * vzorky psané a mluvené britské angličtiny
      * použit pro jednu z nedůležitějších korpusově založených gramatik – Comprehensive Grammar of the English Language
  + *druhá generace*
    - **British National Corpus (BNC)** – reprezentativní korpus psané a mluvené britské angličtiny z let 1991–94
    - **Bank of English (Cobuild) –** nereprezentativní synchronní korpus, 525 milionů slov (v roce 2005)
    - Birmingham Collection of English Text (Collins Corpus) – John Sinclair → Collins Cobuild English Language Dictionary – 1987, první slovník založený na současné, běžně užívané angličtině
    - Corpus of Contemporary American English – 1999–2011, největší americký nereprezentativní synchronní korpus, 425 mil. slov
    - **Prague Dependency Treebank (PDT)** – syntakticky anotovaná část ČNK, 2 mil. slov, specifická verze treebanků
  + *třetí generace* – velmi velké korpusy vzniklé či vznikající po roce 2000 a čerpající i z webu
* ***raná korpusová lingvistika – 50. léta***
  + strukturalistická tradice, metody založené na zkoumání souborů textů a empirii
  + shromažďování jazykového materiálu, rozsah je důležitým parametrem
  + problematika slovní jednotky a lemmatizace
  + zatím ještě nemluvíme o korpusech
  + počátky moderní lexikografie – excerpční lístky, výpisky z beletrie, novin, zapojení slova v kontextu
  + ***Chomského kritika*** – generativní lingvistika, korpusy nejsou v lingvistice potřebné, poskytují pokřivená data – některé věty se v něm nevyskytnou, adekvátnější je intuice

**Světové korpusy**

**První korpus - Brown**

* americká angličtina (1961)
* Brown University, 1964
* gramatické značkování, 1979
* 500 textů, 1 mil. slov
* W. N. Francis & H. Kučera
* první statistické charakteristiky angličtiny
* relativní četnosti slov a slovních druhů

**Susanne**

* autor Geoﬀrey Sampson, Sussex University
* kniha English for the Computer, 1995
* část korpusu Brown (1/4)
* nové gramatické značkování
* syntaktické značkování

**British National Corpus**

* britská angličtina, 10% mluva
* první velký korpus pro lexikografy
* vydavatelé slovníků (OUP) + univerzity
* 1. verze: 1991–1994, 2. verze: World Edition 2000
* ≈3000 dokumentů, 100 mil. slov
* gramatické značkování automatickým nástrojem

**Bank of English**

* britská angličtina
* COBUILD (HarperCollins), University of Birmingham
* 1991, dále rozšiřován
* 2002, ≈450 mil. slov

**Korpusová lingvistika v ČR**

* 1988 *Iniciativní skupina pro přípravu počítačových korpusů, textů a slovníků* (Pala, Čermák, Schmiedtová, Hajičová ad.)
* *Počítačový fond češtiny*, 1992
* *Skupina pro počítačový fond češtiny* – Čermák, Králík, Pala, Hajič, Hajičová, Sgall, Schmiedtová, Benko, Kučera
* 1993–95 *Počítačový korpus českých psaných textů* (GAČR)
* 1994 – založení **Ústavu Českého národního korpusu**
* 1995 – cesta do Velké Británie po centrech korpusové lingvistiky – Pala, Čermák, Petkevič, Schmiedtová

# **4. Typy korpusů (kritéria – časové hledisko, druh zachycené komunikace, rozsah, žánrové hledisko, účel, jazyk). Specializované korpusy. Žákovské korpusy**

* ***členění podle počtu jazyků***
  + *jednojazyčný*
  + *vícejazyčný*
    - např. **InterCorp** (ČNK)
      * zahrnuje přes třicet jazyků, které lze navzájem různým způsobem a ve větším počtu najednou srovnávat
* ***z hlediska tématu***
  + *obecné*
  + *specializované*
    - např. **KSK** (Korpus soukromé korespondence), korpusy jediného autora (Karel Čapek, Bohumil Hraba. Orwellův román 1984), korpus období komunistické totality a další
    - lexikograficky orientovaný korpus tvořící podklad Frekvenčního slovníku češtiny (FSC2000)
    - korpus LINK (korpus lingvistických textů)
    - vzniká i speciální korpus neslyšících, korpus esemesek, e-mailů
* ***z hlediska modu***
  + *mluvený*
    - ze své podstaty je synchronní
    - zachycení mluveného jazyka
    - zvukový záznam přepsán v podobě transkriptu a spojen s nahrávkou
    - **PMK** (tematicky nespecializovaná autentická mluva z oblasti Prahy a okolí)
    - **BMK** (autentický jazyk z Brna)
    - **OMK** (olomoucký; vychází z pražského a brněnského)
    - **ORAL2006** (mluvená čeština z celé oblasti českých nářečí, 1 milion slov), ORAL2008 (vyvážený podle sociolingvistických kategorií), ORAL2013 (připravovaný pro území celé České republiky)
    - korpusy vznikající z fonetických či prozodických důvodů
    - první mluvený korpus – London-Lund Corpus of Spoken English – Jan Svartvik, 100 vzorků po 5000 slovech, morfologicky anotovaný, vyznačeny prosodické (přízvuk, tón, intonace, frázování …) a paralingvistické (použití lidského hlasu) jevy
    - reprezentativnost a rozmanitost z hlediska demografického, situačního, geografického, časového, diskusního, tematického
    - anotační programy pro propojení zvukové nahrávky s odpovídajícím přepisem – ELAN, EXMARaLDA, Transcriber
  + *psaný*
* ***z hlediska časového záběru***
  + *synchronní* – záznam jazykového úzu v jednom období, zachycuje živý jazyk užívaný mezi žijícími mluvčími, žádný korpus není synchronní navždycky
    - kritéria synchronie pro psané korpusy SYN
      * autor beletristických textů je narozen po roce 1880 a dílo bylo vydáno po roce 1945
      * odborné texty vydány po roce 1989
      * publicistika nesmí být starší než 5 let
  + *diachronní –* soubor více korpusů z různých dob vývoje jazyka, snaží se zachycovat proměny jazykového úzu v čase,
    - **DIAKORP** – sedm století vývoje českého jazyka, texty jsou transkribovány a značkovány
* ***z hlediska účelu*** – mohou být různé
* ***z hlediska typu uložení***
  + v prosté podobě (formát ASCII)
  + v podobě označkované, případně syntakticky analyzované
* ***žákovské korpusy***
  + obsahují texty nerodilých mluvčích, kteří se jazyk teprve učí
  + většinou psané školní práce
  + součástí bývá *chybová anotace* – možnost klasifikovat jednotlivé odchylky od běžného úzu a jejich motivaci

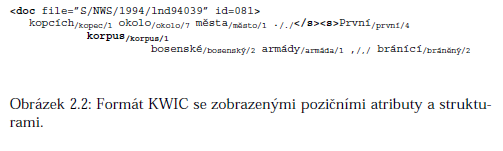
# **5. Budování korpusů. Získávání textů. Přepis, OCR metody, konverze dat. Značkování korpusů (strukturní značky, atributy). Metatextové anotace. Sociolingvistické značkování korpusů mluvených a korpusu soukromé korespondence**

* ***budování korpusů***
  + zdrojem je psaný i mluvený jazyk
  + psané texty lze získat přímo z nakladatelství, webu apod. v elektronické podobě, nebo pomocí OCR či manuálního přepisu
  + *TEI (Text Encoding Initiative)* – vydala doporučení pro společný výměnný formát, zásady kódování, znakové sady a navrhla společný značkovací metajazyk SGML
  + základním stavebním prvkem korpusu je token → *tokenizace* – proces rozdělení textu na pozice
* ***získávání textů***
  + *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)
* **zpracování textů, přepis (editace), konverze**
  + 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**
* ***obsah korpusu***
  + text
  + metainformace
  + struktura dokumentu – odstavce, nadpisy, verše, věty
  + značkování – informace o slovech/pozicích, morfologie, základní tvary, syntaktické vazby, …
* ***struktura korpusu***
  + *strukturní jednotky*
    - opus – text nebo ucelený soubor textů (kniha, číslo novin)
    - dokument – dokument v rámci opusů, každý opus má minimálně jeden dokument
    - věta
  + *strukturní atributy* – autor, název, díla, rok vydání apod.
  + *lingvistická informace* – poziční atributy
  + k zachycení se užívají značkovací jazyky, standardem je XML, často se však využívají různé formy jazyka SGML
* ***vertikála*** (vertikální text) – interní formát pro zachycení struktury korpusu a textů v něm, text je uspořádán do sloupců, první sloupec představuje originální text, další jsou vyhrazeny pro poziční atributy
* ***strukturní jednotky***
  + zapisují se do špičatých závorek
  + umožňují hledat jevy, které se nějakým způsobem vztahují k hranicím různých celků
  + *psané korpusy* – opus, doc (dokument), s (věta)
  + *mluvené korpusy* – doc (dokument – jeden ucelený rozhovor/sonda), sp (mluvčí), seg (technické členění na kratší úseky)
* ***strukturní atributy***
  + ke každé strukturní jednotce náleží jeden nebo více strukturních atributů
  + všechny jednotky mají shodě atribut id, který slouží pro jednoznačnou identifikaci jednotky
  + k jednotce opus se vážou – autor, název, nakladatel, místo vydání, rok, vydání, žánr, typ textu atd.
  + atributy mluvených korpusů
    - doc – rok nahrání sondy, počet mluvčích, typ promluvy
    - sp – číselné označení mluvčího, pohlaví, věk, vzdělání, nářeční oblast
* ***metainformace***
  + (autor, rok, publikace, pohlaví cílové skupiny)
* ***sociolingvistické informace mluvených korpusů***
  + pohlaví
  + věk (mladší 20–35 let, starší 35 let a více)
  + vzdělání (B – základní/středoškolské, A – vysokoškolské včetně započatého)
  + typ promluvy
    - formální – monolog vytvářený sledem odpovědí na otázky kladené nahrávajícím
    - neformální – dialogický soubor promluv dvou mluvčích, kteří se znají
* ***sociolingvistické charakteristiky korpusu soukromé korespondence***
  + pohlaví a věk (4 kategorie)
  + vzdělání pisatele i adresáta (3 kategorie)
  + teritoriální (nářeční) zázemí pisatele
  + typ vztahu mezi pisatelem a adresátem
  + rok napsání dopisu
  + díky nim je možné v případě potřeby pracovat pouze s vybranou částí textů

**Struktura korpusu**

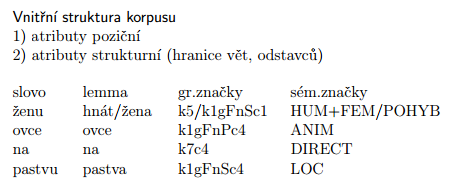
**Struktura**

* pomocí následujících struktur dokážeme pokrýt všechny možnosti značkování struktury textu, které obsahuje jazyk SGML. I když v korpusovém manažeru CQP lze uchovávat hodnotu pro struktury, nelze je používat v dotazech. Systém se omezuje pouze na existenci začátku nebo konce určité struktury
* používá se zápisu ve tvaru SGML, tedy jméno struktury uzavřené v úhlových závorkách, např. <s> jako začátek věty a **</s>** jako konec věty.
* typy struktury: plochá, stromová, prázdná. Pomocí nich můžeme pokrýt všechny možnosti značkování struktury textu, které obsahuje jazyk SGML.
  + **plochá struktura:** Struktury nejsou vnořovány, ani se nijak nepřekrývají. Každá struktura je dána svým začátkem a koncem. Mohou korpus rozčleňovat na části, tedy kde jedna struktura končí, tam další začíná – příkladem je doc. Nebo se jednotlivé intervaly vyskytují samostatně v textu – např. head, lang.
  + **stromová struktura:** struktury jsou dány začátkem a koncem, ale navíc lze stejné značky vnořovat do sebe. Například list, q.
  + **prázdná struktura:** značky tvoří pouhé „oddělovače“. Jsou dány jedním údajem: mezi kterými pozicemi se nachází. Příklad jsou značky pre a g.



**Strukturní atribut (hranice vět, odstavců)**

* **strukturní atributy** jsou prvky vkládané do korpusu mezi pozice, které korpus strukturují na různě velké souvislé celky (věty, dokumenty, knihy).
* strukturní atributy rozdělují korpus na menší celky, odrážející strukturu původního zdrojového dokumentu. Každý korpus může obsahovat své vlastní strukturní atributy, mezi nejčastější patří např. „opus“ nebo „doc“ (např. „**doc.autor**“ obsahuje informaci o autorovi původního dokumentu).
* **atributy struktur** k jednotlivým intervalům struktur přidávají další informace. Můžeme je brát stejně jako poziční atributy, tedy hodnotami jsou řetězce znaků, resp. množiny řetězců znaků, lze pro ně definovat dynamické atributy. Žádným způsobem nebudeme omezovat počet různých atributů k jedné struktuře. Každá struktura má samozřejmě svoji vlastní skupinu atributů, jejich jména se případně mohou shodovat.



**Poziční atribut**

* u každé pozice (slovo, číslo, interpunkční znaménko apod.) je v [dostupných korpusech](http://www.korpus.cz/bonito/korpusy.php) přiřazen jeden nebo více pozičních atributů
* **poziční atribut** je textová informace, která se vztahuje k dané pozici
* atribut může obsahovat například *informaci o slovním druhu slova, jeho základním tvaru* atd. Různé korpusy obsahují různé soubory atributů
* v korpusu SYN2000 najdeme poziční atributy následovně:
  + Zobrazení – Atributy;
    - základní slovní tvar (**lemma**)
    - morfologické značky (**tag**)
    - hledané slovo psané malými písmeny (**lc**)
    - slovní druh (**pos**)

# **6. Lingvistické značkování. Tokenizace, lemmatizace, automatická morfologická analýza, disambiguace (statistické metody, pravidlová disambiguace, hybridní metody). Lingvistické problémy grafického zachycení korpusů mluveného jazyka (přepis a pravidla přepisu v mluvených korpusech budovaných v rámci ČNK). Stručná charakteristika olomouckého mluveného korpusu**

* ***metody 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)
    - **informace lingvistické**, nákladné => omezuje se na morfologické značkování jednotlivých slovních tvarů (**tagování**), přiřazení slovnědruhové charakteristiky a **lemmatizaci** (přiřazení základního, slovníkového tvaru)
* ***poziční atributy***
  + informace, které se vztahují k jednomu tokenu (jedné textové pozici)
  + základní atributy
    - word – slovní tvar
    - lemma – základní tvar
    - tag – značka zachycující morfologickou informaci
    - lc – ekvivalent slovního druhu, zanedbává velikost písmen
    - pos – značka slovního druhu
* ***tokenizace*** – rozčlenění textu na jednotky – řetězce tvarů z obou stran ohraničené oddělovači
* ***lemmatizace –*** přidělení reprezentativní slovníkové podoby hesla každé formě v korpusu
  + lemma každého substantiva je nom. sg., u adjektiv je to nom. sg. mask, u sloves infinitiv
  + součástí procesu desambiguace
  + bývá ztížena homonymií
  + problematická je lemmatizace víceslovných spojení
  + prováděna lemmatizátorem, který bývá součástí taggeru
* ***disambiguace*** – zjednoznačnění, provádí ji tagger
  + odstranění homonymie, týká se všech jazykových rovin, nejčastěji se však uplatňuje na morfologické; *disambiguace* = jednoznačná interpretace slovního tvaru či skupiny slovních tvarů nebo věty na základě kontextu či mimojazykové situace
    - lexikální homonymie (*stát, žena, hnát*)
    - syntaktická homonymie (*Jím špagety s masem/se salátem/s použitím vidličky/se sebezapřením/s přítelem*.)
    - sémantická (*Jeřáb je vysoký. Viděli jsme veliké oko.*)
    - referenční (*Oni přišli pozdě. Můžeš mi půjčit knihu. Ředitel vyhodil dělníka, protože byl agresivní*.)
  + **metody disambiguace**
    - **stochastická (statistická, pravděpodobnostní)**
      * 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
    - **hybridní:** kombinace obou
  + *morfologická d.*
    - ruční
    - poloautomatická
    - plně automatická
      * **stochastická** (statistická) realizovaná na základě strojového učení, např. skrytého markovského modelu
      * založená na lingvistických pravidlech
        + vytvářené lingvisty
        + automaticky vyvozené z textů
      * hybridní
      * DESAMB, angličtina CLAWS
  + *syntaktická d. -* součástí syntaktické analýzy
  + *stochastická disambiguace*
    - nejprve se ručně označkuje větší množství textů (v řádu několika set tisíc slov) jako trénovací korpus
    - disambiguační korpus se naučí na trénovacím korpusu správné značkování
    - úspěšnost ČNK je cca 94 %, nejlepší programy 98 %
    - hlavní problém – nedostatek trénovacích dat, syntagramatická a slovosledná variabilita textu je příliš velká
  + *pravidly řízená disambiguace*
    - založena na syntaktických pravidlech
    - pravidla jsou ověřována na datech korpusu
    - jde o modelování jazykového systému, není tak závislé na trénovacích datech
* ***pravidla přepisu v korpusech ORAL***
  + přepisuje se věrně vše, co je nahráno včetně přeřeknutí
  + promluvy se dělí do segmentů – úseky odpovídající cca jednoduché větě, podřadnému souvětí nebo větnému fragmentu
  + používá se pauzová interpunkce
  + velká písmena jsou vyhrazena pouze vlastním jména, na začátku vět se nepíší
  + vyznačují se neukončené věty, přerušené výpovědi
  + kvůli anonymitě se kódují jména a příjmení
  + zachycují se příznakové rysy běžné mluvy, včetně regionálních, nespisovnou výslovnost
  + zachycují se případy různé spodoby znělosti (schoda/zhoda), odlišnosti v kvantitě
* ***olomoucký mluvený korpus***
  + vychází z koncepce PMK a BMK
  + duální systém transkripce
    - důsledně ortografický pro potřeby morfologické analýzy
    - fonetický přepis
    - mluvčí žije nebo žil v Olomouci nebo je v denním kontaktu s olomouckou mluvou
    - sociolingvistické proměnné – pohlaví, věk, vzdělání, zaměstnání, původ, délka pobytu v Olomouci

**Značkování**

* **morfologické:** přiřazení symbolů/značek SD každému výskytu slova v korpusu; odrazový můstek pro další výzkum)
  + značkování pro češtinu: **LEMMA, Hajičův morfologický analyzátor**, **Ajka (R. Sedláček)**
  + u nás potřeba značkování rozdělit do dvou fází:
    - zpracování morfologickým analyzátorem
    - desambiguace: manuální, program DESMB
  + **Majka:** 1999, Sedláček; schopen pracovat s 150 000 českými kmeny, dovede každému rozpoznanému slovnímu tvaru přiřadit jeho SD a odpovídající gramatické kategorie
    - založena na úplné plavidlové morfologické analýze češtiny
  + jiné: TreeTagger (enTenTen12), CLAWS (BNC, COCA), FreeLing (esTenTen11)
* **syntaktické:** značkování na vyšší úrovni
  + např. v London-Lund Corpus
  + u nás: Synt, SET, DIS/VADIS, IOBBER (polština)
  + jiné: MST Parser, MaltParser
  + **stromové banky (treebanks):** syntakticky analyzované subkorpusy; dlouhodobé vytváření; textové soubory tvořené větami, u nichž je vyznačena syntaktická struktura, např. ve tvaru syntaktického (složkového) stromu.
    - (Věděl jsem, (že (přijde)) a (že (mi (dá pusu))).
    - je-li stromová banka vytvořena, lze z ní automaticky odvodit *frázovou gramatiku*, v níž minimální podstromy interpretujeme jako *nekontextová* pravidla
  + Pražský závislostní korpus
* **sémantická** (word sense tagging/desambiguation /WSD/ - rozlišení sémantických významů slova, named entity recognition – rozpoznání jmenných entit /NER/)
  + u nás (čeština): DESAMB – desambiguace morfologických značek
  + jiné: WordNet, SuperSenseTagger – WSD, NER
* **koreference** (určení anafory)
  + u nás (angličtina): SARA
* **pragmatická** (označení mluvčího, komunikační situace)

# **7. Korpusy v českém prostředí. ČNK – historie a současnost. Charakteristika jednotlivých korpusů pod ČNK (korpusy řady SYN, projekt DIAKORP a další). Charakteristika korpusů budovaných na FI MU (webové korpusy)**

* ***ČNK (Český národní korpus)***
  + založen 1994 **F. Čermákem**
  + centrum korpusového výzkumu v ČR
  + synchronní korpusy **SYN** 
    - SYN2000: 100 000 000 slovních tvarů; časová etapa 1990–1999
  + synchronní mluvené korpusy **PMK** (Pražský mluvený korpus; nahrávky z r. 1988–1996) a **BMK** (Brněnský mluvený korpus, *1994-1999 zachycujících 294 mluvčích*) a **Oral**
    - 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.
    - ORAL-PMK (700 000 tvarů)
  + diachronní korpus **DIAKORP** (1 750 000 tvarů) – pokrývá období od 13. století do roku 1945
  + paralelní korpus **InterCorp**
  + zpřístupňuje data pro automatické zpracování přirozeného jazyka
  + vývoj nových nástrojů pro vytěžování korpusů (např. SyD)
  + rozvoj metodologie korpusové lingvistiky
  + žánrové složení korpusu psaného jazyka ČNK
    - 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)
  + dále viz <https://ucnk.ff.cuni.cz/doc/2002_cnk.rtf> a [https://wiki.korpus.cz/doku.php/cnk:struktura#korpusymluvene](https://wiki.korpus.cz/doku.php/cnk:struktura" \l "korpusymluvene)

**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))

Vznikl v Praze na **Ústavu formální a aplikované lingvistiky MFF UK** a navazuje na tradici pražské lingvistické školy. Závislostní pak znamená, že z hlediska syntaktického je zvolen **závislostní přístup**, kdy za hlavní člen věty je považován predikát (nejčastěji sloveso), který je rozvíjen dalšími, závislými členy (které mohou být rovněž rozvíjeny).

Pražský závislostní korpus je anotován na třech úrovních: **morfologické** (určení lemmat, slovních druhů a gramatických kategorií, jako jsou rod, číslo, pád, …), **syntaktické** (syntaktické informace – analytická funkce, závislosti jednotlivých uzlů) a **tektogramatické** (rozbor sémantiky, významu).

**Korpusy na FI:**

* Desam
  + 1996, ručně značkovaný (desambiguovaný)
  + ≈1 mil. slov
* Czes
  + periodika z webu, z let 1996–1998, další el. zdroje, webové zdroje (crawl)
  + ≈465 mil.
* \*TenTen
  + různé jazyky, ve spolupráci s LCL, UK
  + 1–20 mld. pozic
* Chyby
  + práce studentů předmětu Základy odb. stylu s vyznačenými chybami
  + ≈400 tis.

# **8. Korpusové manažery a další nástroje pro práci s jazykovými korpusy. Volba korpusu vhodného pro řešení lingvistického problému. Morfologické značkování českých korpusů (dva analyzátory – pražský a brněnský, poziční a atributový systém, shody a odlišnosti, porovnání tagsetů)**

**Nástroje pro práci s jazykovými korpusy**

* základem **konkordanční programy**, které třídí a počítají objekty nalezené v korpusu, což jsou v syrovém korpusu slovní tvary, interpunkce, případně další znaky
  + konkordanční programy neumí rozlišit víceznačnosti, proto potřeba gramatických analyzátorů
  + **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
* **gramatické analyzátory**: orientace na morfologii, syntax a v poslední době i na sémantiku
  + značkování (tagging)
  + značkovací programy (taggers) různé úrovně
  + taggery se snaží každému slovu v korpusu přiřadit jeho gramatickou značku (jeho SD, včetně relevantních gramatických kategorií)
* **korpusové manažery:** 
  + program umožňující efektivně pracovat s počítačovým korpusem, tj. vyhledávat podle zadatelných kritérií (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*
  + mají vlastní procesor, uživatelské rozhraní, lze zadávat vyhledávací dotazy, výstupem jsou konkordanční seznamy, výskyty slov a slovních tvarů v kontextech, kolokace, frekvenční údaje, statistické parametry. Nástroje na zpracování korpusů – uložení textu, editace/příprava textu, značkování, rozdělení do pozic (tokenizace), vyhledávání (konkordance), statistiky
  + např. **GCQP** nebo **Bonito**, který je vyvíjen a ověřován

**Jednotlivé korpusové manažery**

* systém **Manatee**, výkonný korpusový manažer; neposkytuje uživatelské rozhraní ani neřeší správu uživatelů => Bonito
* **Bonito 1:** klientská část Bonito komunikuje se serverovou částí ManateeSRV
  + výhoda: interaktivní prostředí, rychlá odezva, široká funkcionalita včetně podpory statistik, vytváření subkorpusů
  + nevýhoda: závislá na platformě; nutná instalace -> problémy s novými verzemi; zastarání uživatelského rozhraní, vzhled neodpovídá současným požadavkům
* **Bonito 2:** webové aplikace, v Pythonu; klientský počítač potřebuje webový prohlížeč
  + neinteraktivní GUI (při každé akci se musí načíst nová stránka), menší funkcionalita -> uživatelé radši využívají Bonito 1
* **Kondor:** webová aplikace, stále ve vývoji
* **CQP**
  + zadávání vyhledávacích dotazů funguje na bázi regulárních výrazů
  + výstup: konkordanční seznamy, výskyty slov v kontextech
  + lze vyhledávat kolokace (slovní spojení)
  + lze získávat základní frekvenční údaje ke slovům a kolokacím
  + lze počítat další statistické parametry jako MI a T-score
  + u značkovaného korpusu lze vyhledávat podle gramatických kategorií a lemmat a také podle strukturních značek

Word sketch engine

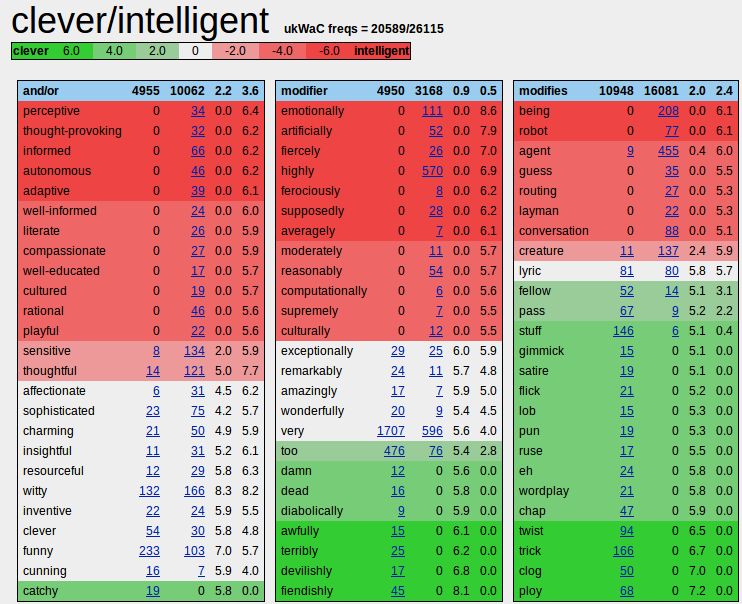
* 2003 Pavel Rychlý, [Adam Kilgarriff](https://en.wikipedia.org/wiki/Adam_Kilgarriff), [www.sketchengine.co.uk](http://www.sketchengine.co.uk/)
* webové rozhraní propojené s korpusovým manažerem Bonito 2
* vznikl díky spolupráci CZPJ a společnosti Lexicography MasterClass ve VB
* slouží k vyhledávání ve velkých jazykových korpusech, **nabízí přístup k 66 jazykovým korpusům** v různých grafických soustavách (azbuka, hebrejské písmo) a 42 jazycích
* vznikl díky spolupráci CZPJ a společnosti Lexicography MasterClass ve VB
* slouží k rychlému vyhledání a **rozbrazení kontextu zadaného slova** (substantiva, adjektiva, slovesa), přičemž se řídí gramatickými charakteristikami, jimiž jsou korpusy doplněny
* výsledkem po zadání slovesného lemmatu je seznam kontextových slov, které se k zadanému lemmatu váží, roztříděných do jednotlivých tabulek
* Sketch Engine je **korpusový manažer**, který nabízí uživateli kromě standardních funkcí jako je konkordance, třídění a filtrování také **funkci Word Sketch, funkci Tezaurus a Sketch Difference**
* Sketch Engine zobrazuje výsledky hledání i se statistickými parametry. Konkordance je zobrazení zadaného slova v jeho kontextovém okolí. U každého konkordančního řádku je také uveden textový zdroj, z něhož byla slova a slovní spojení převzata.
* Word Sketch je funkce, díky níž se zobrazuje seznam kolokací zvoleného slova. Seznam kolokací je rozdělen podle syntaktických vztahů do syntaktických kategorií
* Word Sketch:



* **tezaurus** 
  + !Sketch Engine (SkE) je software, který vyhledává slovní profily (*word sketches*), sdružuje je na základě gramatických relací a vytváří z korpusu tezaury (tj. slovník vrstvený hierarchicky, ve významových strukturách)
  + thesaurus: software zjistí, která slova se objevují ve stejných kolokacích jako jiná slova a na základě těchto dat generuje “distributional thesaurus”. Jedná se o automaticky vytvořený thesaurus, který umí najít taková slova, která se vyskytují v podobném kontextu jako hledané slovo



* **work sketch diferences**



**Předzpracování korpusových dat získaných z internetu**

**JustText**

* nástroj, který odstraní nežádoucí obsah
* odstraňuje netextové části webových stránek (html značky, styly, poznámky; negramatické věty: navigace, reklamy, tabulky, přliš krátké úseky)

**Onion**

* odstranění podobných odstavců
* odstraní duplicitní texty

**Unitok**

* tokenizace

**Chared**

* detekce znakové sady dokumentu
* rozpoznává kódování textu

**WebCrawler**

* druh počítačového programu, prochází internet, stahuje dokumenty (metainformace, obsah), ukládá části dokumentů v různých formátech k dalšímu použití
* k získání textových dokumentů pro ZPJ – SpiderLing (spuštění crawleru)

Anotace

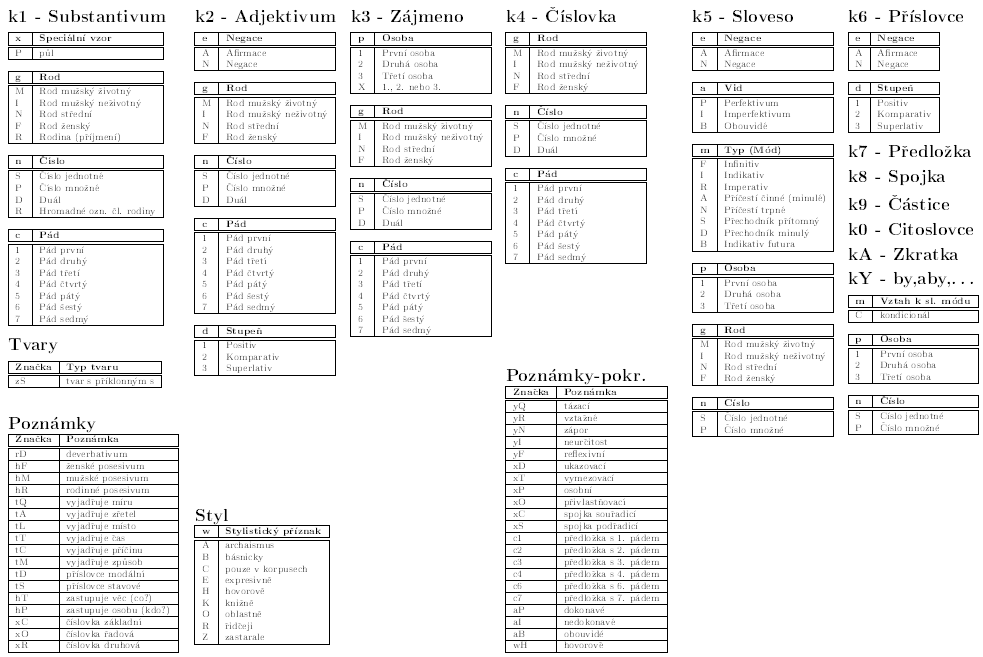
* ***anotace –*** přidávání lingvistických informací do korpusu
  + *morfologická* (Majka, TreeTagger, CLAWS pro BNC a COCA, FreeLing)
  + *syntaktická*
    - parsing – závislostní nebo složkové stromy
    - chunking – rozdělení na fráze
    - synt, SET, DiS/VaDis, IOBBER, MST Parser, MaltParser
  + *sémantická* – word sense desambiguation, named entity recognition
  + *koreference* – určení anafor (SARA)
  + *pragmatická* – mluvčí, komunikační situace
  + *editory anotací* – XML; Gate Brat, WordSmith, PhraseAnotator, SySel

Značkování v ČNK (poziční systém)

Celkem 15 pozic, 2 jsou volné pro případ nových kategorií.

* **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.**

Značkování brněnského korpusu (atributový systém)



rozdíl mezi pražskou a brněnskou značkou

**NNNS4-----A----:** substantivum, obyčejné, neutrum, singulár, akuzativ, afirmace

**k1gNnSc4:** substantivum, neutrum, singulár, akuzativ (nezachycena specifikace a afirmace)

**Výhody brněnského systému:** přehlednější, úspornější, snadno rozšiřitelný

# **9. Značkování syntaktické roviny jazyka. Pražský závislostní korpus. Projekty VALLEX a VERBALEX**

**Syntaktická anotace**

* *syntaktická*
  + parsing – závislostní nebo složkové stromy
  + chunking – rozdělení na fráze
  + **synt, SET, DiS/VaDis, IOBBER, MST Parser, MaltParser**

**Syntaktické značkování:**

* značkování na vyšší úrovni
* např. v London-Lund Corpus
* u nás: Synt, SET, DIS/VADIS, IOBBER (polština)
* jiné: MST Parser, MaltParser
* **stromové banky (treebanks):** syntakticky analyzované subkorpusy; dlouhodobé vytváření; textové soubory tvořené větami, u nichž je vyznačena syntaktická struktura, např. ve tvaru syntaktického (složkového) stromu.
  + (Věděl jsem, (že (přijde)) a (že (mi (dá pusu))).
  + je-li stromová banka vytvořena, lze z ní automaticky odvodit frázovou gramatiku, v níž minimální podstromy interpretujeme jako *nekontextová* pravidla
* Pražský závislostní korpus

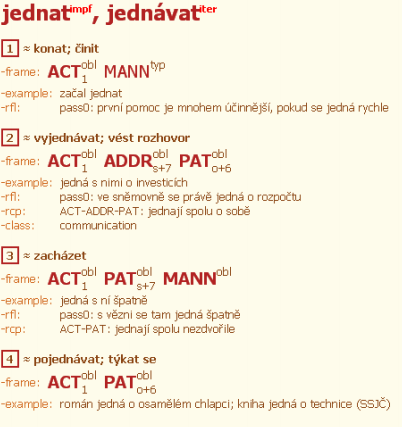
**PDT (Prague Dependency Treebank; Pražský závislostní korpus)**

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))

Vznikl v Praze na **Ústavu formální a aplikované lingvistiky MFF UK** a navazuje na tradici pražské lingvistické školy. Závislostní pak znamená, že z hlediska syntaktického je zvolen **závislostní přístup**, kdy za hlavní člen věty je považován predikát (nejčastěji sloveso), který je rozvíjen dalšími, závislými členy (které mohou být rovněž rozvíjeny).

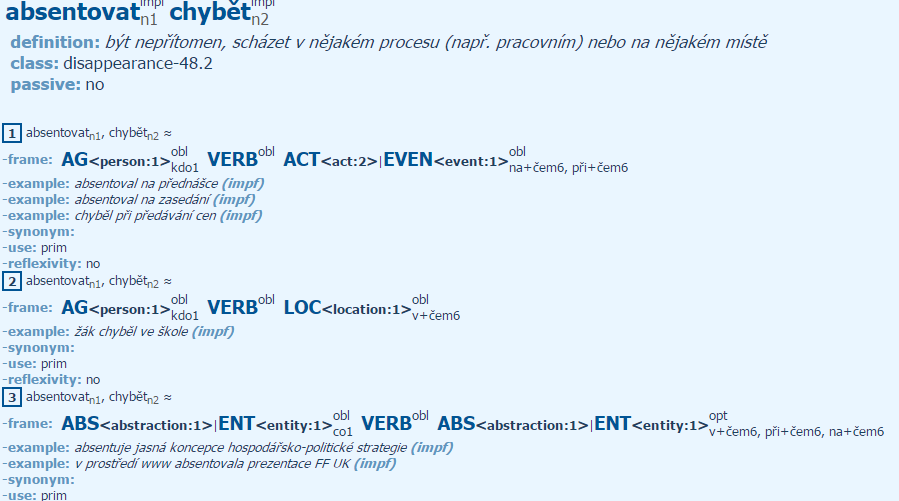
Pražský závislostní korpus je anotován na třech úrovních: **morfologické** (určení lemmat, slovních druhů a gramatických kategorií, jako jsou rod, číslo, pád, …), **syntaktické** (syntaktické informace – analytická funkce, závislosti jednotlivých uzlů) a **tektogramatické** (rozbor sémantiky, významu).

**VALLEX**

* Valenční slovník českých sloves (VALency LEXicon)
* **ÚFAL UK**
* 2001–07, tři elektronické verze, tištěná verze 2008
* valenční slovník jako zdroj syntakticko-sémantické informace
* valenční rámce / významy pro slovesa – komplexní zpracování sloves (všechny významy daného slovesa – top-down přístup)
* obsahuje 2 730 sloves a 6 460 slovesných významů
* přístupný prostřednictvím webového prohlížeče, v pdf formátu a jako xml soubor

**VerbaLex**

* **NLP FI, 2005**
* lexikální databáze
* speciální synchronní slovník, který je vytvářen pomocí počítačových nástrojů. Tento slovník pracuje s elektronickými zdroji dat
* tento slovník vychází ze tří základních zdrojů – BRIEF, VALLEX a Český WordNet
* organizace lexikálních dat v tzv. základních a komplexních valenčních rámcích
* zápis valenčních rámců ke slovesným synonymickým řadám (ne pouze k jednotlivým lemmatům)
* rozlišování významů polysémních sloves
* dvojúrovňové sémantické role
* obsahuje: **6256 slovesných synsetů**



**Základní valenční rámec**

* valenční doplnění na syntaktické úrovni (přímé a předložkové pády)
* valenční doplnění na sémantické úrovni (sémantické role)
* nejfrekventovanější idiomatická doplnění

**Komplexní valenční rámec**

* synonymie, číslování významů polysémních sloves
* definice významu synonymické řady
* homonymie (číslování odlišných významů, př. sladit, stát)
* možnost tvoření pasiva + tranzitivnost, intranzitivnost
* slovesný vid (slovesa dokonavá, nedokonavá, obouvidá)
* sémantické třídy sloves
* způsob užití slovesa (základní, přenesené, idiomatické)
* reflexivita (refl. tantum, reciprocita, syntakt. refl., absol. synon.)
* doloženo konkrétními příklady

# **10. Lexikální významy – slova a kolokace z hlediska automatických nástrojů, využití korpusů, elektronické slovníky (obsah, budování), lexikální databáze (WordNet, EuroWordNet, thesaury)**

**Lexikální analýza**

**Lexikální sémantika (starší název sémaziologie)**

* zabývá se významem lexikálních jednotek a významovými vztahy, základní metodologický postup je od formy k významu, odpovídá se tedy na otázku, jaké významy mají formy v jazyce
* význam slov a slovník spojení (kolokace)

**Lexikální analýza**

* identifikace jednotlivých slov a sousloví v textu dokumentu
* někdy se identifikace sousloví uvádí vzhledem ke své složitosti jako samostatná procedura
* **odstranění nevýznamových a nespecifických slov**
  + provádí se pomocí předem připraveného negativního slovníku; někdy se tato procedura považuje za součást lexikální analýzy
* **lematizace** (*stemming*)
  + redukce slov na jejich základní tvary, tj. kmen nebo kořen
  + opakem tohoto procesu (např. při vyhledávání) je derivace, kdy se k základnímu tvaru slova generují jeho inflexní tvary
* **srovnání slov**, resp. jejich kmenů nebo kořenů s termíny řízeného slovníku
  + jedná se o jednoduché srovnání výrazu vybraného z textu s termínem řízeného slovníku bez ohledu na strukturu řízeného slovníku; ta bývá zohledňována pouze při automatickém přiřazování
* **vážení** neboli stanovení vah termínů (*weighting*)
  + nejčastěji se provádí na základě frekvenčních metod

**Lexikální analýza**

* jednotlivá **slova** jsou v textu nejčastěji **rozpoznávána pomocí mezer**, i když u některých typů slov není tato identifikace zcela jednoduchá nebo jednoznačná. Obtížnější je např. určování zkratek, v nichž se tečka musí odlišit od větné tečky, problém představují také výrazy se spojovníkem, u kterých je třeba rozhodnout, zda je chápat jako jedno slovo nebo dvě samostatná slova.
* samostatný **problém** při identifikaci slov představují také **číslice**, u kterých je třeba stanovit, zda budou zpracovány jako samostatná slova nebo závislé prvky (např. *1. místo*) nebo zda budou z analýzy a dalšího zpracování zcela vypuštěny.
* podstatně složitější je **identifikace sousloví**, které jsou z hlediska sémantické nosnosti a selektivní síly podstatně významnější než jednotlivá slova. Pro identifikaci sousloví byla vyvinuta řada metod, zde uvádíme dvě nejobvyklejší:
  + **statistická identifikace sousloví** – u slov, která se v textu, popř. v celé databázi vyskytují často společně, lze předpokládat, že se jedná o sousloví. Při statistické analýze se vychází z frekvence výskytu daného sousloví (záleží na pořadí slov), současného výskytu slov (nezáleží na pořadí slov) nebo vzdálenosti slov v textu, přičemž tato vzdálenost může být určena buď počtem slov mezi slovy předpokládaného sousloví, nebo může být sledován jejich současný výskyt ve větě, odstavci nebo jiné ucelené části textu. Vzhledem k tomu, že častý současný výskyt dvou slov ještě nemusí znamenat, že se jedná o sousloví, není tato metoda bezezbytku úspěšná
  + **syntaktická identifikace sousloví** - identifikace sousloví probíhá jako v předchozím případě s tím, že mezi slovy v potenciálním sousloví je pomocí různých metod analyzována jejich syntaktická složka. Často se v tomto případě využívá existujících slovníků sousloví, popř. řízených slovníků.
* součástí identifikace sousloví bývá i normalizace jejich formy. Sousloví se v textu mohou vyskytovat v různých syntaktických (např. *indexace dokumentů x indexovaný dokument*), lexikálních (např. *intelektuální indexace x manuální indexace*) nebo morfologických (např. *pravidla indexace x indexační pravidla*) variantách, které je nezbytné pro potřeby indexace sjednotit. Používá se následujících metod:
  + **normalizace formy sousloví pomocí slovníku variant sousloví** - jedná se o jednoduchou metodu, která je však většinou z důvodu omezeného rozsahu slovníku vhodná pouze v rámci dílčího oboru
  + **vypuštění pomocných slovních druhů** (předložek, spojek, příslovcí apod.) ze sousloví a zanedbání pořadí zbylých složek sousloví
  + **syntaktická analýza sousloví v kombinaci s použitím kmene nebo kořene jednotlivých slov sousloví**

**Nástroje**

**Lexikální analyzátor** (*scanner*)

* je součástí [překladače](http://cs.wikipedia.org/wiki/Překladač)
* lexikální analyzátor rozdělí vstupní posloupnost znaků na [lexémy](http://cs.wikipedia.org/wiki/Lexém) – lexikální jednotky (např. identifikátory, čísla, klíčová slova, operátory, …). Tyto lexémy jsou reprezentovány ve formě [tokenů](http://cs.wikipedia.org/wiki/Token), ty jsou poskytnuty ke zpracování [syntaktickému analyzátoru](http://cs.wikipedia.org/wiki/Syntaktická_analýza)
* Úkolem lexikálního analyzátoru je také odstranění komentářů a bílých znaků ze zdrojového programu
* V praxi je lexikální analyzátor realizován pomocí [konečného automatu](http://cs.wikipedia.org/wiki/Konečný_automat)

Word sketch engine

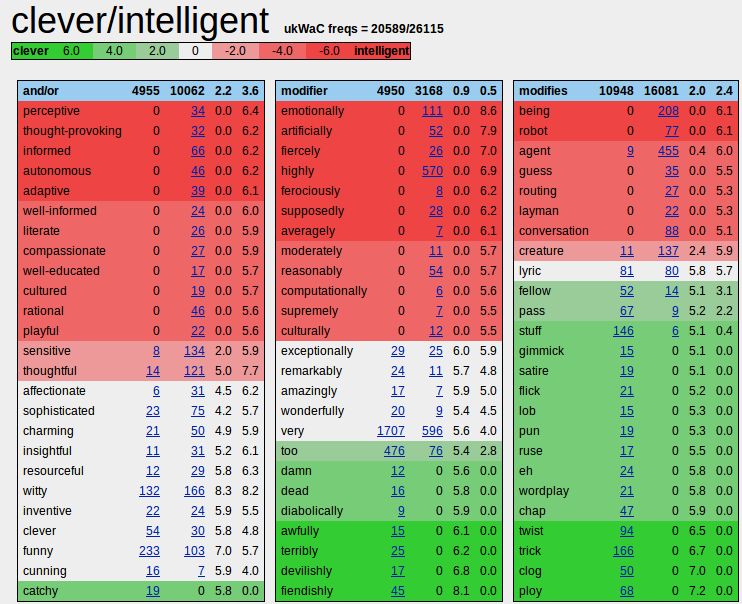
* 2003 Pavel Rychlý, [Adam Kilgarriff](https://en.wikipedia.org/wiki/Adam_Kilgarriff), [www.sketchengine.co.uk](http://www.sketchengine.co.uk/)
* webové rozhraní propojené s korpusovým manažerem Bonito 2
* vznikl díky spolupráci CZPJ a společnosti Lexicography MasterClass ve VB
* slouží k vyhledávání ve velkých jazykových korpusech, **nabízí přístup k 66 jazykovým korpusům** v různých grafických soustavách (azbuka, hebrejské písmo) a 42 jazycích
* vznikl díky spolupráci CZPJ a společnosti Lexicography MasterClass ve VB
* slouží k rychlému vyhledání a **rozbrazení kontextu zadaného slova** (substantiva, adjektiva, slovesa), přičemž se řídí gramatickými charakteristikami, jimiž jsou korpusy doplněny
* výsledkem po zadání slovesného lemmatu je seznam kontextových slov, které se k zadanému lemmatu váží, roztříděných do jednotlivých tabulek.
* Sketch Engine je korpusový manažer, který nabízí uživateli kromě standardních funkcí jako je konkordance, třídění a filtrování také **funkci Word Sketch, funkci Tezaurus a Sketch Difference**.
* Sketch Engine zobrazuje výsledky hledání i se statistickými parametry. Konkordance je zobrazení zadaného slova v jeho kontextovém okolí. U každého konkordančního řádku je také uveden textový zdroj, z něhož byla slova a slovní spojení převzata.
* Word Sketch je funkce, díky níž se zobrazuje seznam kolokací zvoleného slova. Seznam kolokací je rozdělen podle syntaktických vztahů do syntaktických kategorií
* Word Sketch:



* **tezaurus** 
  + !Sketch Engine (SkE) je software, který vyhledává slovní profily (*word sketches*), sdružuje je na základě gramatických relací a vytváří z korpusu tezaury (tj. slovník vrstvený hierarchicky, ve významových strukturách)
  + thesaurus: software zjistí, která slova se objevují ve stejných kolokacích jako jiná slova a na základě těchto dat generuje “distributional thesaurus”. Jedná se o automaticky vytvořený thesaurus, který umí najít taková slova, která se vyskytují v podobném kontextu jako hledané slovo



* **work sketch diferences**



*Kolokace a statistika*

**M-score**

* mutual information
* vzájemná informace, pravděpodobnost Px, Py, Px,y
* pravděpodobnost slova/výskytu v kontextu X
* vychází z teorie informace
* logdice: vzájemná informace, míra pevnosti spojení, míra M-Score; harmonizovaný průměr relativních četností, numerická úprava hodnot (hezké zobrazování)
* nevítanou vlastností MI-score je to, že je **velmi ovlivňováno frekvencí jednotlivých slov**. Nejvyšších hodnot totiž dosahují dvojice slov s nízkou frekvencí. Z tohoto důvodu umožňují korpusové manažery při výpočtu MI-score nastavit **spodní hranici frekvence** a pro slova s absolutní frekvencí pod touto hranicí se potom MI-score nepočítá
* hodnoty MI-score jsou převážně kladné (záporné hodnoty značí vzájemné odpuzování jednotek a jsou relativně řídké). Za relevantní bývá považována **hranice MI = 7** (pro stomilionový korpus), kdy je oprávněná domněnka, že se jedná o systémovou kolokaci. Tato hodnota ovšem samozřejmě závisí na velikosti korpus

**T-score**

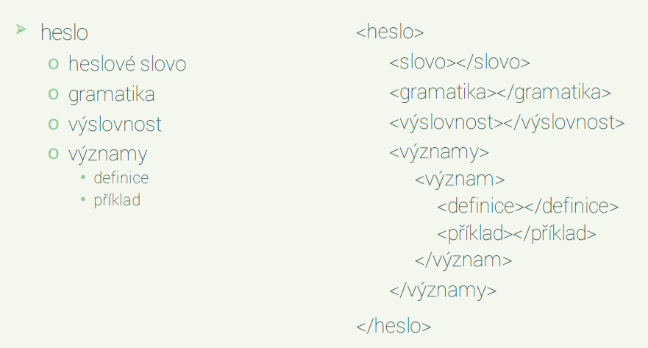
* vychází ze statistiky, z testování hypotéz
* je to T test, jestli je něco statisticky významné nebo ne, používají se hladiny jistoty
* míra kontrastu
* používá se ve statistice dost často
* minimální počet výskytů 5
* T-score vychází ze statistické metody testování hypotéz pomocí tzv. t-testu a bývá někdy označována jako míra kontrastu.
* v případě[**kolokací**](https://wiki.korpus.cz/doku.php/pojmy:kolokace) 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 naopak tím pravděpodobnější je, že jde o pevnější, ustálenější kombinace slov, tj. o kolokace

**Podstatný rozdíl mezi MI-score a T-score:**

* MI-score nachází silné kolokace s velkou relativní frekvencí, a tedy spíše výjimečné až náhodné
* T-score nachází kolokace nenáhodné, pravidelné a ustálené, ale nepříliš výrazné

**Elektronické slovníky**

* xml



* lexikání databáze
  + PraLeD (Pražská Lexikální Databáze)
  + DANTE (Database of ANalysed Texts of English)

**Dictionary writing systém**

* aplikace pro tvorbu slovníků (obvykle celý ➢ proces tvorby)
* často vlastní
* komerční
  + IDM DPS ‑ klient‑server (Windows)
  + iLex ‑ jádro a dokupované moduly, samostatně nebo klient‑server, mobily (Windows, Linux, Mac)
  + TLex ‑ online, offline (Windows, Mac)
* nekomerční (Glossword, Matapuna)
* DEB (Dictionary Editor and Browser)

**Slovníková makrostruktura**

**Slovníková mikrostruktura**

**viz okruh 11, ot. 7 (Počítačová lexikografie)**

**Wordnet a sémantické sítě**

* standardní způsob organizace lexikálního materiálu ve slovnících je **abecední řazení** (lexikografické uspořádání). Hledání v takových slovnících je pomalé. Slovníky obsahují prototypické informace založené na **genu proximum** (nadřazený pojem), neodkazují však k výrazům stejného typu, k hyponymům nebo sourozencům (coordinates).
* v poslední době se značná pozornost věnuje lexikální sémantice s cílem vytvořit lexikální zdroje, které by popisovaly jak významy lexikálních jednotek, tak jejich vztahy formálně (algoritmicky) a umožňovaly tak systematické využití v NLP
* dalším směrem je budování počítačových **lexikálních databází a elektronických verzí tezaurů, vznik sémantických sítí**. **Reprezentují faktové znalosti (pojmy + vztahy), znalosti jsou uloženy ve formě grafu**
  + k nejdůležitějším vztahům patří **podtřída** – vztah mezi třídami a **instance** – vztah mezi konkrétním objektem a jeho rodičovskou třídou

**Motivace**

* hledání ve slovníků je pomalé a časově náročné; používají také různá hyperonyma tam, kde by bylo vhodné pracovat s jedním
* prototypické definice ukazují vždy směrem nahoru: k nadřazeným pojmům, nikdo do strany k výrazům stejného typu (sourozencům) nebo dolů k hyponymům
* **=>> počátek** psycholexikologie je spojen se studiem slovních asociací s pokusy o modelování mentálního slovníku → **myšlenkou organizovat slovník konceptuálně** ⇒ vytvořit Wordnet
* WordNet: slovní síť – založený na psycholingvistických principech
* **synset: synonymická řada**
* Wordnet: člení slovník do **pěti kategorií:**
  + substantiva, verba, adjektiva, adverbia a funkční slova (synsémantika) – ta jsou pravděpodobně uložena odděleně od ostatní slovní zásoby a tvoří část syntaktické složky SZ
* členění se opírá o asociační experimenty - lidé měli reagovat na předložená slova prvním slovem, které je napadlo
* jednotlivé slovní druhy organizovány odlišně:
  + substantiva - tematická hierarchie
  + slovesa - na základě různých vztahů vyplývání
  + adjektiva, adverbia - n-dimenzionální hyperprostory (množiny n-tic)

**WordNet**

Slouží k reprezentaci faktových znalostí – pojmy + vztahy. Autorem je skupina kolem **G. A. Millera** z Princetonu, vznikl kolem r. **1960** pro reprezentaci významu anglických slov. **Znalosti jsou uloženy ve formě grafu**. V Evropě vznikl EuroWordNet. Miller usiluje o **poznání toho, jak je organizována naše lexikálních paměť a na jakých principech jsou budovány naše mentální slovníky**.

WordNet je založen na **psycholingvistických principech**. Ve verzi 3.0 obsahuje databáze 155 287 slov uspořádaných do 117 659 synsetů, čímž je pokryto 206 941 slovních významů (dvojic slovo-smysl).

Slovník člení do pěti kategorií: **substantiva, verba, adjektiva, adverbia a funkční slova** (sysnsémantika).

Slovní druhy jsou rozlišovány na základě jejich sémantické povahy – substantiva jako tematické hierarchie, sloves na zákl. vztahů vyplývání, adj. a dv. jako n-dimenzionální hyperprostory.

Pokouší se lexikální informace organizovat v termínech slovních významů a nikoli slovních tvarů. Přiřazení forem a významů je víceznačné – některým formám odpovídá více různých významů a zároveň některé významy mohou být vyjádřeny různými formami.

Lexikální paměť lze chápat jako organizovanou stromově, kde zákl. vztahem je **tranzitivní a antisymetrický významový vztah ISA** (is a kind of = je druhu), tzn. vztah hypero/hyponymie vedoucí od specifickému ke generickému = generalizace a specializace. WordNet je lexikální databází s hierarchickou strukturou umožňující prohledávání shora dolů a zdola nahoru stejnou rychlostí.

K editaci wordnetů byl na FI MU vyvinut nástroj **DEBVisDic**. Jako diplomová práce vznik VisualBrowser – nástroj na vizualizaci (sémantických) sítí.

Sémantické vztahy ve WordNetu

* **hyponymie/hyperonymie** – vztah podřazenosti/nadřazenosti (ISA-vztah), tranzitivní a antisymetrický, generuje hierarchickou (stromovou) reprezentaci pro substantiva
  + uskupují substantiva tak, že tvoří lexikální dědičný systém
  + popis významu substantivních synsetů je obvykle založen na nadřazeném výraz (termu) doplněném o rozlišující příznaky (differentia specifica)
  + synsety (synonimické řady) jsou propojeny ohodnocenými ukazateli (pointry)
  + každé slovo dědí všechny rozlišující příznaky všech svých nadřazených výrazů, např. činnosti, zvíře, výrobek, událost, pocit, jídlo, rostlina, proces, vztah atd. (původně bylo 25 tematických souborů), jestliže platí vlastnost pro třídu, platí i pro všechny její podtřídy a pro všechny její prvky
* funguje mechanismus vzorů a výjimek
  + vzor je hodnota vlastnosti u třídy nebo podtřídy, platí ta, která je blíže objektu
  + výjimka – u konkrétního objektu, odlišná od vzoru
* **synonymie** – nevysvětluje, co jednotlivé významy jsou, ale vyznačují jejich existenci a diferenciaci, významy spojené vztahem synonymie se seskupují do synonymických řad, které jsou zákl. organizačním prvkem sémantické sítě,, vynucuj si oddělení slovních druhů – nelze volně substituovat jednotky patřící k různým syntaktickým kategoriím
* **antonymie** – centrální organizující vztah pro adjektiva a adverbia
* **meronymie/homonymie** – vztah část – celek, tranzitivní a asymetrický vztah, vede také k budování hierarchických struktur

**Adjektiva** – člení se na dvě zákl. třídy

* deskriptivní – jsou organizována v termínech binárních opozic antonymních (*velký: malý*) a podobných významů (synonym)
* relační – *prezidentský, nukleární, zubní* – mají vztah k určitému substantivu nebo jsou s ním nějak spojena, nerozlišují škály, nemají přímá antonyma a nelze je stupňovat
* samostatné stojí skupina referenčně modifikujících adjektiv (*předchozí, údajný*) a adj. označující barvy

**Slovesa** jsou organizována na základě vztahu vyplývání (*prodávat: platit*) a jeho modifikací: troponymie (*chrápat: spát*) a kauzálních vztahů (*dát: mít*). 15 hlavních tříd, např. slovesa tělesných funkcí, poznání, komunikace, soutěžení, emocí, pohybu a další.

EuroWordNet

Samotný WordNet je vytvořen pro angličtinu, EuroWordNet v rámci dvou projektů pokrývá další jazyky (aj, holandština, italština, španělština a francouzština, němčina, čeština, estonština). Princetonsky WordNet rozšiřuje a upravuje.

Český WordNet

Obsahuje asi 28 tis. synsetů, což pokrývá 47.542 slovních významů (dvojic slovo-smysl).

Při jeho vytváření bylo použito **výkladového slovníku češtiny** (pracovní název pro postupně budovanou lexikální databázi češtiny s důrazem na maximální vnitřní konzistenci), **Lingea Lexicon 2.0** (oboustranný aj-čj a čj-aj slovník) a **Slovníku českých synonym** (Pala, Všianský, 1994).

Pomocné zdroje – seznam českých kolokací z textového korpusu ESO (FI MU), gramaticky i strukturálně značkovaný korpus DESAM (FI MU).