**POČÍTAČOVÉ ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA**

**1. Techniky word sense disambiguation a algoritmy.**

* ***Word Sense Desambiguation*** – tj. lexikální desambiguace; jedná se o název z oblasti AI pro automatickou disambiguaci (jedná se o AI-úplný problém, tj. problém řešitelný pomocí algoritmu)
* ***techniky desambiguace***: ***(1) pravidlové techniky*** (DES, DES1), ***(2) statistické techniky*** (bigramy, trigramy), tím, že se jedná o klasifikační úlohu, využívají se též metody strojového učení, kt. jsou dvojího typu: (a) strojové učení s učitelem, využívá se ručně anotovaná trénovaná data (bayesův klasifikátor, rozhodovací stromy,; (b) strojové učení bez učitele (tzv. shluková analýza); algoritmy shlukování (3) hybridní (kombinace obou)
* úspěšnost metod WSD lze vyhodnocovat měřením podílu automaticky přiřazených významů, kt. se shodují s ruční anotací
* DESAM
* algoritmicky řešitelná úloha – 2 různé přístupy: (1) porozumět zpracovávanému textu (báze znalostí), (2) algoritmy strojového učení: klasifikační a shlukovací algoritmy

***Lexikální disambiguace (WSD)***

* zjednoznačnění významu; nalezení významu slova v daném kontextu
* probíhá ve 2 krocích: (1) stanoví se všechny významy daného slova v textu; (2) určí se nejvhodnější význam na základě kontextu
* sémantické rámce? tagger?
* Word Sense Desambiguation (WSD) – název v oblasti AI pro automatickou disambiguaci
* metody:
  + hloubkové – metody založené na znalostní bázi využívající slovník/tezaurus
    - sémantické sítě (rámce) – tj. znalostní báze sloužící k reprezentaci vztahů mezi jednotl. slovy; WordNet
    - znalostní systémy: WordNet, paralelní korpusy
    - jak to funguje: porovnávají neznáme slovo se znalostní bází, přičemž berou v potaz další slova v testovaném dokumentu (kontextové okno), kt. porovnávají s významy definovanými ve znalostní bázi–> Yarowski přišel na to, že velikost kontext. okna může značně ovlivnit výsledek disambiguace (50 slov v kontext. okně)
    - metody založené na znalostní bázi –> přesnost odhahu významu může ovlivnit i popis významů slov ve znalostní bázi
  + povrchové – metody založené na strojovém učení,

**http://nase-rec.ujc.cas.cz/archiv.php?lang=en&art=7974**

**2. Regulární jazyky a regulární výrazy, CQL.**

* ***regulární jazyky* –** typ formálního jazyka v Chomského hierarchii gramatik a jazyků, je generovatelný regulární gramatikou (konečněstavovou, neboť má konečnou paměť) a rozpoznatelný tzv. konečným automatem; r. j. modelují jednoduché struktury jazyka, stačí na modelování všech menších struktur než je věta (zejm. slova, nominální fráze apod); regulární gramatika (typu 3, levou stranu každého pravidla tvoří neterminál a pravou stranu terminál, nebo terminál a za ním neterminál
* ***regulární výrazy*** – pojem pochází z teorie formálních jazyků, jedná se o speciální znaky pomocí kterých můžeme popsat jistou množinu textového řetězce řetězců odpovídající hledanému výrazu, jsou velmi podstatnou součástí dotazovacího jazyka (k dotazování v databázových systémech, má přesně definovanou syntax) ; vyskytují se téměř ve všech programovacích jazycích, textových editorech apod., nejčastěji se využívají pro vyhledávání či manipulaci částí textu;
* ***CQL*** – dotazovací jazyk užívaný v korpusovém manažeru (aplikace pro vyhledávání v korpusech) Manatee (serverová část) – KonText (klientská část), jehož základem je dotaz na jedno slovo (pozici) v korpusu: [atribut="hodnota"], atribut poziční (word/lemma/tag; tj. dodatečné info o tokenu=pozice, 3 tokeny – řekl, že), hodnota – hledaný výraz/regulární výrazy

**3. Strojový překlad – pravidlové a statistické přístupy, existující systémy, vyhodnocování.**

* ***strojový překlad*** – tj. obor pc lingvistiky zabývající se návrhem, implementací a aplikací automatických systémů (programů) pro překlad textů s minimálním zásahem člověka; patří mezi AI-úplné problémy
* ***klasifikace dle přístupu*** – (1) pravidlový strojový překlad, (2) statistický, (3) hybridní
* ***pravidlové systémy***: (1) strojový překlad založený na pravidlech (; (2) strojový překlad založený na znalosti jazyka

**4. Morfologická analýza přirozeného jazyka a desambiguace, přístupy.**

* ***morfologické anotace*** – tj. lemmatizace + morfologická analýza + desambiguace (jedná se o nejnižší rovinu zpracování PJ)
* ***morfologická analýza*** (AMA) – provádí ji morfologický analyzátor, kt. každému tokenu přiřadí jeho lemma a morfologické interpretace (tagy), program zjednodušeně řečeno porovnává vstupní slovo se slovy uloženými ve strojovém slovníku; v případě, že má slovo více než 1 možnou interpretaci, přiřadí mu interpretace všechny (na úrovni lemmatu i tagu) –> proto je nedílnou součástí strojového zpracování textů morfologická disambiguace neboli zjednoznačnění, kt. vybere ze všech interpretací interpretaci jedinou
* ***princip AMA***: morfologický (strojový) slovník
* ***čeština – systémy pro AMA***: existují 2 systémy – ***(1) pražský*** (poziční, Hajičův, 16 pozic, kt. nabývají různých značek; specifický význam jednotl. značek vyplývá podle toho, na jaké pozici se daná značka vyskytuje) a ***(2) brněnský*** s morfologickým analyzátorem Majka: nepoziční, atributivní, tag je dvojice atribut–hodnota bez ohledu na pořadí; užit v korpusu czTenTen12 – morfologická analýza pomocí majky, disambiguace pomocí DESAM
* obecně k anotaci: vnitrotextová/vnětextová (autor, rok vzniku textu)
* ***morfologická disambiguace*** –

**5. Syntaktická analýza přirozeného jazyka, syntaktické stromy závislostní a složkové, algoritmy analýzy.**

* ***úkolem*** syntaktické analýzy je rozpoznat, zda vstupní řetězec je větou v daném jazyce, pokud ano dá na výstup informaci o její struktuře (např. derivační strom); tj. syntakt. analýzou získáváme syntakt. strukturu
* ***větná struktura*** (tj. strukturovaný popis větných členů) – ***3 možné způsoby notace: (1) závislostní strom*** (PDT – Pražský závislostní strom, v centru stojí sloveso; musí existovat pravidla, kt. stanovují mezi jakými slovy vedou hrany –> např. PDT má 3 sady vět: věty pro učení, věty pro testování a věty pro vyhodnocení), ***(2) složkový*** (výsledek bezkontextové analýzy; vychází z poznatku, že věta se člení primárně na část podmětovou a přísudkovou a každá z nich se pak člení na další své složky, např. brněnský synt); ***(3) parciální syntaktická analýza*** (nepotřebujeme celý strom, zajímají nás pouze některé vztahy – word sketches –> tj. pro vyhledávání častých kolokací v korpusech)
* to, že je nějaká věta větou definujeme pomocí gramatik nebo pomocí pravděpodobnostního ohodnocení (statistické odhadování podoby stromu); (1) bezkontextová gramatika: CKY, chart parser; (2) závislostní přístup: algoritmus hledání nejmenší kostry na statisticky naučené vazby; DCG (Prolog) – podobné běžným přepisovavím pravidlům, ale neterminály začínají malými písmeny
* ***čeština – syntaktická analýza***: 2 hl. přístupy – (1) brněnský (synt, jehož základem je bezkontextová gramatika a algoritmus pro analýzu bezkontextových jazyků head-driven chart parser – tabulkový analyzátor vhodný pro nejednoznačné gramatiky (jakým je kupříkladu gramatika PJ); vyvíjený v laboratoři NLP FI MU; výstupem je složkový strom); (2) pražský (PDT) – tyto analyzátory fungují na principu pravděpodobnostní analýzy, využívají tzv. trénovací fráze – při kt.
* ***ad struktura věty***: hloubková struktura (týká se spíše významu) x povrchová struktura (spíše zápisu): 1 povrch. reprezentaci věty může odpovídat více hloubkových (věta je významově víceznačná), obdobně 1 význam se dá vyjádřit různě

**6. Budování a indexace korpusů, webové korpusy, statistické nástroje pro korpusy, jazykové modelování.**

* **indexování** – klíč (identifikátor záznamu, např. v relační databázi), index (seřazená posloupnost hodnot pro 1 klíč, s ukazateli na záznam); pro rychlé vyhledávání

**7. Počítačová lexikografie – systémy pro editaci slovníků, značkování slovníkového hesla.**

* ***počítačová lexikografie* –** se zabývá tvorbou slovníků na základě korpusů; reprezentace dat, TEI, LMF (značkovací jazyk pro tvorbu lexikální zdrojů), Dictionary Writing Systems (aplikace pro tvorbu slovníků)
* **tvorba slovníku:**  lexikální databáze, slovník
* **DEB**
* **struktura slovníku** – makrostruktura (heslář), mikrostruktura (1 záznam ve slovníku)
* ***lexikální databáze*** – tj. strukturovaná jazyk. databáze (na základě korpusů v podle době, gramatické údaje, valnce, vzory); podklad pro budování slovníků, PraLed (Pražský Lexikální Databáze), DANTE

**8. Paralelní korpusy, zarovnání.**

* ***paralelní korpus* –** tj. vícejazyčný (obsahuje zpravidla originální text a jeho překlad, jedná se v zásadě o synchronní korpus); slouží ke srovnání jazyků, základ. předpokladem při jejich tvorbě je zarovnání
* ***zarovnání*** – paralelní korpusy vyžadují zarovnání (jednotce (slovu, častěji větě) v 1 jazyce odpovídá jednotka v druhém jazyce); program aligner (často na základě pravděpodobnosti přiřazuje)
* ***InterCorp*** – paralelní synchronní korpus; referenční (texty se nemění, jsou tedy dohledatelné, dobré např. při citacích); má 2 části – jádro (beletristické texty s ručním zarovnáním) a kolekce (texty získané ve více jazycích, automaticky zarovnané); každý cizojazyčný text má českou verzi (čj – pivot, tj. česká verze textu – originál/překlad), dostupný před KonText

**9. Strojové učení – princip, techniky, vyhodnocování, křížová validace.**

* ***strojové učení*** (machine learning) – tj. podoblast AI (umělé inteligence), zabývající se algoritmy a technikami, kt. umožňují pc systémům (agentům) "učit se" **(**učení obecně – je klíčové pro neznámé prostředí; učení znamená využít vjemy z prostředí, adaptuje svůj vnitřní stav a vyvozuje další akce); máme vstupní a výstupní veličiny a snažíme se najít takový model, kt. co nejlépe postihne závislost výstupní veličiny vzhledem ke vstupním veličinám
* ***algoritmy strojového učení dle typu učení***: ***(1) s učitelem*** (je známa výstupní veličina, predikce – regrese, klasifikace), ***(2) bez učitele*** (není známa výstupní veličina, pouze vstupní, shlukování), ***(3) posílené*** (kombinace předchozích 2)
* zákl. druhy:
* ***křížová validace*** – tj. 1 ze způsobů testování přesnosti modelu,
* ***shluková analýza*** – statistická metoda, pomocí kt. je možné objekty klasifikovat na základě jejich podobnosti či příbuznosti; hierachická (objekty uspořádány ve stromové struktuře) / nehierarchická (objekty zařazeny do předem stanoveného počtu skupin, shluků, tak, aby si objekty byly vzájemně podobné

**10. Reprezentace znalostí, jazyková znalost, znalost o světě, encyklopedická znalost, common-sense.**

* ***reprezentace znalostí*** – způsob, jak vyjádřit určité znalosti počítačově zpracovatelnou formou (za účelem odvozování); znalostní báze (jsou zde uloženy info o pozorování světa a vlastních akcích/info o doméně; je to množina vět, do kt. přidává další znalosti pomocí akce TELL a získává z ní info pomocí akce ASK); ***nástroje pro reprezentaci znalostí*** (pro popis objektů v KB): (1) deklarativní schémata: znalosti vyjádřené pomocí poznatků (ontologie), sémantické sítě, jazyk predikátové logiky I. řádu, fuzzy logika; (2) rámcová schémata – rámce
* ***vyvozování znalostí*** – zpracovává znalosti uložené ve znalostní bázi (KB) a provádí odvození nových závěrů (odpovědi na dotazy, odvodit akci vyplývající z dodaných znalostí apod.); vyvozování nových znalost = hledání důkazu (algoritmus konstrukce důkazu: dopředně a zpětné zřetězení, rezoluce, logické programování); tj. komponenta zvaná inferenční stroj (jedná se o algoritmy nezávislé na doméně)
* ***2 užitečné předpoklady***: (1) předpoklad uzavřeného světa: cokoliv o čem nevím, že je pravda –> nepravda; (2) předpoklad jednoznačných pojmenování (různá jména označují různé objekty)
* ***jazykové znalosti*** – údaje o jazyce (gramatika, pravděpodobnost výskytů jednotl. sekvencí řeč. segmentů)
* znalost o světě –
* encyklopedická znalost
* **common-sense znalosti** –
* **využití:** např. expertní systémy (něco jako inteligentní pc program, kt. užívá znalosti a inferenční procedury k řešení problému; logický agent – pc program využívající znalosti
* ***logický (znalostní) agent* –** 2 komponenty: znalostní báze (KB), inferenční stroj
* ***sémantická síť*** – tj. reprezentace faktových znalostí, znalosti ukládá formou orientovaného grafu: uzly – pojmy (objekty, třídy) a hrany: vztahy mezi nimi, nejdůležitější vztahy: (1) podtřída (vztah mezi třídami); (2) instance (vztah mezi konkrétním objektem a jeho rodičovskou třídou); objekty jsou orientovány do hierarchie tříd, relace jsou tranzitivní, dědičnost: jestliže určitá vlastnost platí pro nějakou třídu, pak platí i pro celou podtřídu, jestliže určitá vlastnost platí pro nějakou třídu, pak platí i pro všechny prvky této třídy, WordNet – sémantická síť 100 000 pojmů zachycující synonyma, antonyma
* ***sémantické rámce*** – varianta sémant. sítí; velice populární pro reprezentaci znalostí v expertních systémech; rámce – univerzální struktury, kam se ukládají všechny informace relevantní pro daný pojem (obdobně jako sémant. sítě podporují dědičnost), rámce obsahují: objekty, sloty a hodnoty slotů

**11. Dialogové systémy – dialogové strategie, typy dialogů, značkovací jazyk pro popis dialogu, využití dialogových systémů, komunikační agenty (chatbots).**

* ***dialogový systém*** – systém komunikující s uživatelem pomocí dialogu v PJ (disponuje dialogovým (hlasovým/textovým) rozhraním), alternativně komunikace pomocí DTMF (tónová volba; SMS), multimodální komunikace (řeč+obraz, řeč+text apod.); dialog – tj. komunikace 2 účastníků (pc a člověk)
* ***struktura dialog. systému*** (typický hlas. dialog. systém) – několik modulů vykonávající různé fce: (1) rozpoznávání řeči, (2) porozumění jazyku (sémantická analýza), (3) generování odezvy, (4) pc syntéza řeči, (5) dialogový manažer (ústřední modul)
* ***značkovací jazyk pro popis dialogů***: ***SSML*** – značkovací jazyk určený pro zvýšení kvality syntézy řeči (vytváří standard pro značkování prozodických jevů mluvené řeči, popisuje např. výslovnost cizích slov, výslovnost zkratek apod.), ***VXML*** (VoiceXML) – jazyk pro popis dialog. strategií (jeho cílem je zlepšit vývoj webových aplikací –> zpřístupnit obsah pro dialog. systémy; definuje 2 druhy dialogů: formulář (interpretovány pomocí algortitmu FIA): pro získání hodnot položek a menu: poskytující uživateli sadu možností a odkazů na pokračování dialogu)
* ***dialogová strategie*** – tj. postup, kt. k dané promluvě vytvoří následující promluvu; využívá znalost stavu dialogu; je vlastností každého účastníka dialogu; hodnotící fce (přiřazuje danému dialogu reálné číslo popisující úspěšnost dialogu z hlediska dané strany) –> z hlediska hodnotící fce: (1) kooperativní dialog, (2) nekooperativní, (3) dialog s nulovým součtem
* ***využití dialogových systémů***:
* ***komunikační agenti (chatbots)***: chatbot (chatbotter) – tj. druh konverzačního agenta (pc program), kt. simuluje inteligentní konverzaci s cílem nerozpoznat, že se nejedná o člověka; jednodušší chatboti využívají k vyhledávání odpovědí klíčová slova (Eliza), kvalitnější zahrnují složitější analýzy komunikace (viz Turingův test)
* ***příklad – chatboti***: ELIZA (otázka obsahující slovo "matka" –> vygeneruje "řekni mi víc o své rodině" apod.) a PARRY, A.L.I.C.E., D.U.D.E.
* ***Turingův test*** – tj. kritérium inteligence pc; systém, kt. se chová jako člověk, Loebnerova cena za "nejlidštější program, zahrnuje: NLP, reprezentaci znalostí, vyvozování znalostí, strojové učení, (pc vidění), (robotiku) + soutěž Chatterbox Challenge

**12. Automatické rozpoznání emocí – techniky, jazykové zdroje.**

* ***emoce*** – z hlediska pc zpracování se na emoce pohlíží jako na součást člověka, kt. má určité fyziologické projevy (Descartes: 6 základních –> hněv, radost, strach, překvapení, zklamání, smutek)
* ***způsoby měření emocí*** – (1) fyziologické: detekce změny tlaku krve+ tepu, detekce galvanický odpor kůže (2) fonetické: emoce mohou ovlivňovat výslovnost jednotl. slov (celé promluvy)+můžeme tak odhalit ironii, vtip, nadsázku apod. –> využívá se u telefonních automatů, kt. mají řešit zákazníkův problém (je ve stresu, spěchu apod.); dále existují i (3) vizuální metody detekující pohyb očních víček (ospalý/nudí se) –> např. automobily vybavené zazřízením proti řízení unavených lidí
* Yale Face Database

**1. Techniky word sense disambiguation a algoritmy**

## Word sense disambiguation

= zjednoznačňování významů slov, nalezení významu slova v daném kontextu, úkolem je zjistit, jaký význam (z inventáře významů) má slovo ve vstupním textu

Jde o **klasifikační úlohu**

* jednotlivé významy tvoří třídy
* podle kontextu se rozhodujeme, do kterých tříd slovo na vstupu patří
* předpoklady: významy jsou diskrétní a je jich konečný počet, máme nějaký inventář významů
* jde o AI complete problém (těžký problém, k jeho řešení je třeba zvládnout všechny ostatní problémy v dané oblasti)

**Aplikace** **WSD**:

* strojový překlad
* inteligentní vyhledávání – information retrieval (IR)
* inteligentní korektor překlepů

Přístupy k WSD jsou **hloubkové** (zahrnují znalosti o jazyce i o světě) a **povrchové** (bez dalších znalostí, počítají s okolím). Hloubkové přístupy potřebují zpracovat a formalizovat příliš mnoho dat, jsou funkční jen v omezených doménách, protože tam je „svět“ malý a experty se vyplatí platit. Povrchové přístupy mají navrch, jsou relativně levné.

Největší slabinou WSD je inventář významů, proto existují jednak snahy vytvořit dobré inventáře a jednak snahy se jim úplně vyhnout (např. projekt Hyperlex). Všechny algoritmy pro WSD pracují s kolokacemi a určitým oknem, ve kterém kolokace sledují. Ono okno může zásadně ovlivňovat průběhy algoritmů. Není žádná doporučená velikost okna. Hlavním důvodem je to, že různá slova mají různý „dopad“ na význam promluvy.

#### Algoritmy WSD

**Algoritmy založené na znalostech**

* reprezentant: ***Leskův algoritmus*** (1986) = slovo *w*, jehož okolí sdílí nejvíc slov s definicí (nebo s příklady užití) *i*tého významu, má význam *si*.
  + Pro každý význam si slova *w*:
    - nastav váhu v na 0: *v*(*si* ) := 0
  + najdi množinu slov *O* v okolí slova *w*
  + pro každé slovo *oj* z okolí *O*
    - pro každý význam si
      * pokud se *oj* nachází v denici n. př. užití *Di*
        + přičti *v(o)* k *v(si )*: *v(si) := v(si ) + v(o)*
  + vyber si s nejvyšším *v(si )*: *return max(v(s*i*))*
  + váha slova *v(o) = idfo*
  + *idft* je nižší, čím běžněji se slovo vyskytuje

**Algoritmy strojového učení** – způsobují změnu stavu počítačového systému tak, že zefektivní schopnost přizpůsobení se

* s učitelem – pro zadaný vstup máme i správný výstup (trénovací data)
* bez učitele – pro zadaný vstup neznáme správný výstup
* kombinace – pro část vstupu máme i správný výstup
* typické úlohy strojového učení jsou klasifikační úlohy
* pravidlové – jsou typické pro lidi, intuitivní, příkladem je hra „Myslím si zvíře“ (protihráč se snaží zvíře uhádnout pomocí otázek, na které dostává odpovědi ano/ne – postupná redukce možných správných odpovědí podle toho jaké vlastnosti jsou vyloučeny zápornou odpovědí, pokračuje se dokud nezůstane (ideálně) jedno zvíře)
  + rozhodovací seznamy – je jednodušší na implementaci

if (zvíře má chobot) then output(slon)

if (zvíře má pruhy) then output(zebra)

if (zvíře má ploutve & zvíře není ryba) then output(žralok)

* + rozhodovací stromy – přehlednější i při vyšší složitosti

savec?

žije ve vodě? žije na souši?

žije v moři? žije v řece? býložravec? masožravec

* matematické
  + pravděpodobnostní: naivní Bayesovský
    - předpokládá nezávislost znaků, je rychlý
    - počítá s podmíněnou pravděpodobností určité vlastnosti (máme černé zvíře → spočítá na kolik procent je možné, že toto zvíře bude kráva, vlk, slon, ...)
  + maximální entropie
  + podobností: k-NN ve vektorovém prostoru
* promluvové
  + one sense per discourse
  + one sense per collocation (Yarowsky) – závisí na první volbě kolokací, způsobu určení pravděpodobnosti, prahu pro pravděpodobnost a správnosti předpokladu one-sense-per-discourse

1. vezmi všechny výskyty slov *w* z korpusu včetně jejich kontextů
2. pro každý možný význam slova vytvoř malou sadu příkladů (ručně nebo pomocí kolokací)
3. vytvoř rozhodovací seznam s pravděpodobnostmi pro další slova vyskytující se v kontextech
4. aplikuj tento seznam na celý korpus (s prahem pro pravděpodobnost)
5. nově zařazená slova obsahují další slova v kontextech
6. algoritmus můžeme upravit pomocí zařazení předpokladu one-sense-per-discourse
7. opakuj kroky 3–6
8. jakmile množiny přestanou narůstat, zastav
9. systém je nyní natrénovaný i na jiný korpus
   * redundance atributů

#### Common Sense

* sdílená znalost, zahrnuje taková fakta, která zná většina z nás, ale také odvozovací schopnosti, které lidé užívají při aplikaci znalostí
* nemá pevnou hranici, má velký rozsah
* není nutně vědecká znalost (někdy jde i proti ní)
* tvrzení common sense jsou příliš obyčejná, než aby je někdo někam psal
* bez common sense není možné úspěšně modelovat porozumění
* common sense najdeme ve výkladových slovnících, encyklopediích, korpusech, sémantických sítích, specializovaných kolekcích (CyC, OpenMind, ConceptNet)

**2. Regulární jazyky a regulární výrazy, CQL**

**3. Strojový překlad – pravidlové a statistické přístupy, existující systémy, vyhodnocování**

## Využití statistiky ve zpracování přirozeného jazyka

prezentace proseminář III, skripta matika

Ke zpracování přirozeného jazyka existují dva přístupy:

* **pravidlový** – občas kriticky selhává, jestliže jsme na něco nevymysleli pravidlo, je potřeba mnoho pravidel, těžko je možné vymyslet pravidla na všechno – přirozený jazyk je vágní, mnohoznačný a vyvíjí se, při tomto přístupu se používají formální gramatiky
* **statistický** – selhává často, ale jen málo, je potřeba mnoho dat ke statickému zpracování (čím víc dat, tím lepší výsledky)

**4. Morfologická analýza přirozeného jazyka a desambiguace, přístupy**

**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.

#### Ceské morfologické analyzátory

##### Ajka

* Radek Sedláček, FI MU Brno
* http://nlp.fi.muni.cz/projekty/ajka/
* značky jsou řetězce dvojic **atribut–hodnota**
* napsaný v C
* využívá struktury **trie**
* 390 000 základních tvarů, 6 300 000 různých slovních tvarů, 15 000 různých značek, slovník 3.13MB
* rychlost analýzy – cca 18 000 slov/s
* v současnosti nový nástroj majka od Pavla Smerka, na principu konečných automatů, s novým mechanizmem vzorů

##### Pražský morfologický analyzátor

* Barbora Hladká, Jan Hajič a jeho tým, UFAL MFF UK Praha ´
* http://ufal.mff.cuni.cz/czech-tagging/
* používá **poziční značky**
* “free” část napsaná v Perlu, menší slovník (cca 76 000 základních tvarů, 6 000 koncovek)

**5. Syntaktická analýza přirozeného jazyka, syntaktické stromy závislostní a složkové, algoritmy analýzy**

**Syntaktická analýza** – struktura větných frází, popisuje, jak vypadá gramatický správná věta, většinou pomocí pravidel gramatiky. **Syntaktický analyzátor** – nástroj, který analyzuje vstup na základě gramatiky a na výstup dává různé info, např. derivační stromy.

**Syntax** – charakterizace dobře utvořených kombinací slovních tvarů do věty nebo fráze. Syntaktická analýza se provádí pomocí gramatických pravidel, výstup ze syntaktické analýzy (např. derivační strom) tvoří často vstup pro analýzu sémantickou.

#### Základní termíny

* **fráze (phrase)** – jednotka jazyka větší než slovo, ale menší než věta např. jmenná fráze, slovesná fráze, adjektivní fráze nebo příslovečná fráze
* **lexikální symbol, lexikální kategorie (lexical category) –** tzv. pre-terminál speciální neterminál gramatiky, který se přímo přepisuje na terminálový řetězec znaků, tj. pravidla tvaru X → w, označuje všechny slova, která odpovídají určitému lexikálnímu symbolu (všechna podstatná jména, přídavná jména, . . . )

N → pes | člověk | dům …

V → nese | chodit | psal …

ADJ → …

PREP → …

* **frázová kategorie (phrasal category)** – neterminální symbol gramatiky, který nevyjadřuje lexikální kategorii
* **větný člen (constituent)** – lexikální nebo frázová kategorie
* **větná struktura (sentence structure)** – strukturovaný popis větných členů
  + povrchová struktura (surface structure) - derivační/složkový strom jako výsledek bezkontextové (CF) analýzy
  + závislostní struktura (dependency structure) – zobrazuje závislosti mezi větnými členy
  + hloubková struktura (deep structure) – sémantická interpretace fráze. Popisuje role větných členů (agens, patiens, donor, cause, . . . )

#### Složkový a závislostní přístup

- dva základní způsoby zadávání gramatik

##### Složkový přístup

* skupiny slov tvoří větné jednotky, které jsou označovány jako fráze, a jako větné členy (složky, constituents) formují větu
* např. podstatné jméno – součást jmenné fráze (noun phrase – NP), jmenná fráze spolu s předložkou – tvoří předložkovou frázi (prepositional phrase – PP)
* syntaktická struktura věty je zachycována jako složkový strom

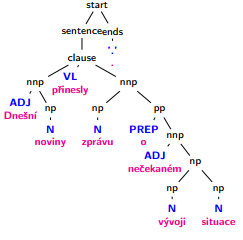
##### Závislostní přístup

* jeden člen vazby je označován jako řídící, druhý jako závislý
* např. přídavné jméno závisí na řídícím podstatném jménu
* syntaktická struktura věty je zachycována pomocí závislostního stromu:
  + uzly odpovídají elementárním jednotkám vstupu (často slovům)
  + hrany označují vztahy závislosti mezi elementárními jednotkami
* závislost není relací mezi jednotlivými slovy, ale obecně relací mezi jedním slovem a frází řízenou druhým slovem. např. vazba mezi konkrétním slovesem a podmětem nebo vazba mezi slovesem a předmětem věty
* technicky vzato, závislostní relace je vztahem mezi uzly a podstromy (uzlem a všemi uzly, které na tomto uzlu závisí)

Jen zřídka se používá čistě složkový či striktně závislostní přístup. Ve složkovém jsou závislosti zpravidla vyjádřeny přidáním označení, která složka je řídící pro danou frázi. Závislostní strom bývá doplněn o informaci určující lineární precedenci. Je možné pak mezi těmito přístupy výsledek převádět.

#### Uzly syntaktického stromu

Označení uzlu (název neterminálu) podle zvoleného přístupu reprezentuje:

* **gramatická role (gramatická funkce)**
  + charakterizují vztahy mezi větnými složkami na povrchové úrovni
  + určujeme, zda daný větný člen je NP v roli podmětu, NP v roli předmětu, ADVP určující lokaci atd.
  + v češtině (a jazycích se systémem gramatických pádů) pomáhá k určení gramatické role právě informace o pádu
  + ovšem přiřazení gramatických rolí ke gramatickým pádům a naopak není zdaleka jednoznačné.
* **tematická role** (též hloubkový/sémantický pád)
  + na rozdíl od gramatické role se jedná o sémantickou kategorii
  + určujeme např.:
    - Agens – kdo je životným původcem nějaké cílevědomé činnosti
    - Patiens – co hraje roli entity, na kterou se působí
    - Donor – osoba, která dává
    - Cause – entita, která způsobuje, že je něco děláno

#### Příznaky a příznakové struktury

Informace v uzlu syntaktického stromu:

* příznaky/rysy (features) – zaznamenávají syntaktické nebo sémantické informace o slovu nebo frázi.
* např. test na shodu:

Malý Petr přišel domů.

podmět (Petr) je ve shodě s přísudkem (přišel) v čísle a rodě, přídavné jméno (malý) a podstaté jméno (Petr) se shodují v pádě, čísle a rodě

S(n, g) → NP(\_, n, g) VP(n, g)

NP(c, n, g) → ADJ(c, n, g) N(c, n, g)

* gramatické znaky (slovní druh, gramatický pád, rod, číslo, osoba, . . . ) je výhodné začlenit do gramatiky ve formě dvojic atribut–hodnota
* potom je možné **zobecňovat**, např. vyjádřit shodu v pádě, čísle a rodě výhradně pomocí atributů
* aplikace – v mnoha gramatických formalismech
* jazykové objekty jsou zde modelovány jako **příznakové struktury** (feature structures), tedy právě matice dvojic atribut–hodnota.
* u složitějších struktur – nestačí pak běžné porovnání instanciace jde oběma směry → použije se **uniﬁkace**

#### Možnosti zadávání gramatik

* nejčastější formát speciﬁkace gramatik – **produkční pravidla** gramatika se skládá z pravidel generujících správně utvořené řetězce
* **cíl analyzátoru** – najít odvození vstupního řetězce ze zadaného neterminálu (označovaného obyčejně velkým písmenem *S*
* z anglického sentence – věta) na základě daných pravidel pokud je tohoto cíle dosaženo, vstup je akceptován a je mu přiřazena odpovídající struktura
* v minulosti rovněž populární – **přechodové sítě** (transition networks) přechody sítě = lingvistické jednotky, uzly sítě = stavy analyzátoru v procesu analýzy vstupu. Přechody jsou označeny symboly deﬁnujícími, za jakých podmínek se analyzátor může přesunout z jednoho stavu do stavu druhého.
* **rozšířené přechodové sítě** (ATN – Augmented TN) jsou doplněny o podmínky a procedury – ekvivalentní deklarativním gramatikám

#### Chomského teorie syntaxe

* 50. léta 20. stol. – Noam Chomsky vytvořil formální teorii syntaxe
* jedna ze základních tezí – autonomie syntaxe ⇐ k ověření syntaktické správnosti věty nepotřebujeme znát její význam

Colorless green ideas sleep furiously.

vs.

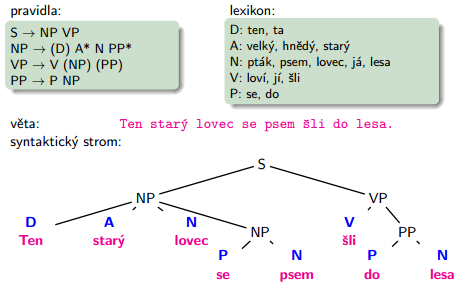
Furiously sleep ideas green colorless.

* syntaktické principy mají univerzální platnost pro různé přirozené jazyky
* znalost jazyka = gramatika
* Noam Chomsky, Aspects of the Theory of Syntax, 1965 – standardní teorie syntaxe – transformační generativní gramatika (TGG)
* snaží se řešit i zachycení sémantických vztahů v hloubkové struktuře
* postupně se vyvinula:
  + v rozšířenou standardní teorii (1968)
  + později tzv. Government & Binding Theory (teorie nadřazení a vázání, 1981), která zakládá na pojmu univerzální gramatiky
  + 90. léta – teorie minimalismu (snaha po úspornosti popisného aparátu)

Chomského předpoklady o rozumu:

* rozum má vrozenou strukturu
* rozum je modulární
* rozum obsahuje speciální modul pro jazyk, porozumění jazyku je oddělitelné od jiných aktivit
* syntaxe je formální, nezávislá na významu a komunikačních funkcích
* znalost jazyka je modulární, obsahuje moduly pro jednotlivé fáze analýzy jazyka

##### Základní části standardní teorie:

* bázová komponenta
  + bezkontextová pravidla a schémata pravidel generují základní strukturu větných členů
  + lexikon popisuje lexikální kategorie a syntaktické rysy lexikálních položek
* transformační pravidla – vložení, smazání, přesun, změna-rysu, kopie-rysu; transformace převádí hloubkové struktury na struktury povrchové
* transformace:
  + obligatorní – např. přesun slovesné koncovky za sloveso
  + fakultativní – např. pasivizace, tvorba otázek, negace (změna významu)
* pravidla bázové komponenty – popisují strom hloubkové struktury v obvyklém pořadí
* transformace umožňují jeho změny na různé povrchové varianty (trpný rod, otázka, . . . )
* stopa (trace) – ukazuje, kde byl prvek před přemístěním

#### Východiska syntaktické analýzy

Návrh podkladů a datových struktur

* syntaktický (odvozovací, derivační) frázový strom – kompletní hierarchický popis struktury věty
* úkol syntaktické analýzy = pro danou gramatiku a daný vstup (větu) dát všechny odvozovací stromy
* existují techniky pro kompaktní uložení lesa takových stromů (chart parsing)
* jelikož se zabýváme výhradně syntaktickou strukturou a nevylučujeme a priori derivační stromy s absurdní interpretací, má většina vět mnoho různých syntaktických stromů

Automatická analýza syntaxe musí vždy projít třemi fázemi:

1. musí být zvolena notace pro zápis gramatiky – gramatický formalismus
2. musí být ve zvoleném formalismu napsána gramatika pro každý jazyk, který bude zpracováván
3. musí být vybrán nebo navržen algoritmus, který určí, zda daný vstup odpovídá gramatice, a pokud ano, jaký popis mu odpovídá

### Algoritmy syntaktické analýzy

Základní postupy pro syntaktickou analýzu obecných bezkontextových gramatik

* obecná CFG – rozsáhlá, (silně) víceznačná, s ∈-pravidly
* všechny uvedené algoritmy pracují s polynomiální časovou a prostorovou složitostí
* algoritmus CKY – Cocke, Kasami, Younger;
* tabulková (chart) analýza (neplést s LR tabulkou):
  + shora dolů (top-down);
  + zdola nahoru (bottom-up);
  + analýza řízená hlavou pravidla (head-driven);
* Tomitův zobecněný algoritmus LR

Vstupem syntaktické analýzy je řetězec lexikálních kategorií (pretermínálních symbolů, např. ADJ CONJ ADJ N V PREP N .) a bezkontextová gramatika. Výstupem je efektivní reprezentace derivačních stromů.

#### Algoritmus CKY

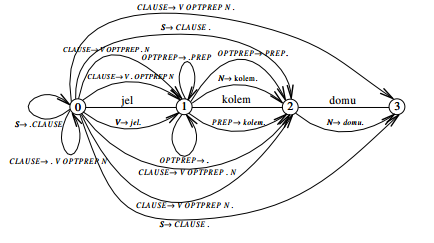
* Gramatika musí být v Chomského normální formě.

CNF (každá CFG jde do ní převést): A → BC

D → ‘d’

* Pro daný vstup délky n derivujeme podřetězce symbolů délky q na pozici p, značíme wp,q, 1 ≤ p, q ≤ n.
* Derivace řetězců délky 1, A ⇒ wp,1, je prováděno prohledáváním terminálních pravidel.
* Derivace delších řetězců A ⇒∗ wp,q, q ≥ 2 vyžaduje aby platilo A ⇒ BC ⇒∗ wp,q. Tedy z B derivujeme řetězec délky k, 1 ≤ k ≤ q, a z C derivujeme zbytek, řetězec délky q − k. Tzn. B ⇒∗ wp,k a C ⇒∗ wp+k,q−k. Kratší řetězce máme tedy vždy “předpočítané.”
* složitost CKY je O(*n3*)
* příklad: vstupní řetězec vstupní gramatika

#### Tabulkové (chart) analyzátory

* Rozlišujeme tři základní typy tabulkových analyzátorů:
  + shora dolů;
  + zdola nahoru;
  + analýza řízená hlavou pravidla.
* Neklade se žádné omezení na gramatiku.
* Analyzátory typu “chart” v sobě většinou obsahují dvě datové struktury **chart** a **agendu**. Chart a agenda obsahují tzv. hrany
* Hrana je trojice [A→ α•β, i, j], kde:
  + i, j jsou celá čísla, 0 ≤ i ≤ j ≤ n pro n slov ve vstupní větě
  + a A → αβ je pravidlem vstupní gramatiky.
* chart po analýze shora dolů

#### Tomitův zobecněný analyzátor LR

* generalized LR parser (GLR)
* Masaru Tomita: Eﬃcient parsing for natural language, 1986
* standardní LR tabulka, která může obsahovat konﬂikty;
* zásobník je reprezentován acyklickým orientovaným grafem (DAG);
* derivační stromy jsou uloženy ve sbaleném “lese” stromů.
* v podstatě stejný, jako algoritmus LR;
* udržujeme si seznam aktivních uzlů zásobníku (grafu);
* akce redukce provádíme vždy před akcemi čtení;
* akci čtení provádíme pro všechny aktivní uzly najednou;
* kde je to možné, tam uzly slučujeme.

**6. Budování a indexace korpusů, webové korpusy, statistické nástroje pro korpusy, jazykové modelování**

**Korpus** je rozsáhlý vnitřně strukturovaný a ucelený soubor textů daného jazyka elektronicky uložený a zpracovávaný. Jsou vytvářeny se zřetelem k zvolenému cíli a vychází z těchto předpokladů:

* jazyková data jsou v korpusu uložena ve své přirozené textové podobě, lze je všestranně a opakovaně zkoumat a vyvozovat z nich příslušné teoretické generalizace
* velký rozsah dat minimalizuje nebezpečí, že by mohlo dojít k převaze jevů okrajových nad základními
* velký rozsah dat je podmínkou reprezentativnosti

**Token** – výskyt slovního tvaru v korpusu, slovní tvar jako takový.

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

#### Budování korpusů

Zdrojem dat je psaný i mluvený jazyk. Z psaných textů se data získávají třemi způsoby:

* konverzi ze sázecích disket a pásek, které lze získat od většiny nakladatelství
* užitím OCR – závisí na kvalitě scanneru a programového vybavení a typografické složitosti textu
* klasickým opisováním do počítače

#### Typy korpusů a standardizace

**Elektronické archivy** – volné kolekce různorodých textů, různé formáty a jazyky.

**Vlastní korpusy** – tvoří úplné celky. Většina Evropských jazyků už má dnes svůj vlastní korpus. Existují také korpusy paralelní – dvojjazyčné, obecné a specifické, velké korpusy obsahující subkorpusy jazyka psaného, mluveného, nářečí, synchronní – diachronní aj.

S rostoucím počtem korpusů vzniká potřeba jejich standradizace → vznik **Text Encoding Initiative (TEI)** – vydala doporučení pro společný výměnný formát, zásady kódování, znakové sady, navrhla společný kódovací – značkovací jazyk **Standard Generalized Markup Language (SGML)**.

#### Obsah korpusu

* text
* metainformace
* struktura dokumentu – odstavce, nadpisy, verše, věty
* značkování – informace o slovech/pozicích, morfologie, základní tvary, syntaktické vazby, …

#### Tokenizace

Rozdělení textu do pozic

* může silně ovlivnit výsledky dotazování, četnosti i značkování
* token (pozice) = základní prvek korpusu
* většinou slovo, číslo, interpunkce
  + bude-li, don’t – 4 možnosti:

1. |don’t|

2. |don| |’t|

3. |don| |’| |t|

4. |do| |n’t| – v BNC

* + zkratky (s tečkama?)
  + data
  + desetinná čísla, ...

**7. Počítačová lexikografie – systémy pro editaci slovníků, značkování slovníkového hesla**

**8. Paralelní korpusy, zarovnání**

**9. Strojové učení – princip, techniky, vyhodnocování, křížová validace**

**10. Reprezentace znalostí, jazyková znalost, znalost o světě, encyklopedická znalost, common-sense**

## Reprezentace znalostí (ontologie, wordnet, rámce)

* hledá způsob vyjádření znalostí počítačově zpracovatelnou formou (za účelem odvozování)
* vyvozování znalostí (reasoning) – zpracovává znalosti uložené v bázi znalosti (knwledge base, KB) a provádí odvození (inference) nových závěrů
  + odpovědi na dotazy
  + zjištění faktů, které vyplývají z faktů a pravidel v KB
  + odvození akcí, které vyplývají z dodaných znalostí
* možnosti: ontologie, wordnet, rámce

### Ontologie

* explicitní a formalizovaný popis určité problematiky
* je slovníkem, který slouží k uchovávání a předávání znalosti týkající se určité problematiky
* obsahuje **glosář** (definici pojmů) a **tezaurus** (definici vztahů mezi jednotlivými pojmy)
* datový model reprezentující určitou znalost nebo její část

#### Typy prvků

* **jedinci** – živý i neživý objekt, abstraktní pojem atp.
* **třídy** – množina jedinců určitého typu nebo další podtřídy
* **atributy** – popisují určitou vlastnost, charakteristiku či parametr jedince, obsahují název a hodnotu a jsou určeny pro uložení určité informace vztahující se k danému jedinci
* **vazby** – jednosměrné nebo obousměrné propojení dvou jedinců

### 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.

#### 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 tématické hierarchi, 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 konzistenciú, 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)

### Rámce

* varianta sémantických sítí
* všechny informace se ukládají do univerzálních struktur – rámců
* stejně jako sémantické sítě podporují dědičnost
* dobře vyjadřují statické znalosti, tedy nějakou hierarchii pojmů
* rámec obsahuje objekty, sloty a hodnoty

|  |  |  |
| --- | --- | --- |
| **Objekt** | **Slot** | **Hodnota** |
| **Savec** | podtřída | zvíře |
|  | část | hlava |
|  | \*má kožich | ano |
| **Slon** | podtřída | savec |
|  | \*barva | šedá |
|  | \*velikost | velký |

\* vzorové hodnoty, které mohou měnit hodnoty u podtříd a instancí

#### Sémantické sítě vs. rámce

Sémantické sítě Rámce

- uzly - objekty

- spoje - sloty

- uzel na druhém konci spoje - hodnota slotu

**11. Dialogové systémy – dialogové strategie, typy dialogů, značkovací jazyk pro popis dialogu, využití dialogových systémů, komunikační agenty (chatbots)**

## Komunikační agenty (chatbots)

* počítačové programy simulující psanou konverzaci s člověkem v přirozeném jazyce
* snaha o vytvoření dojmu, že agent tématu konverzace skutečně rozumí
* snaha o splnění Turingova testu
* pro zahnání volné chvíle, ale i jako virtuální asistent, průvodce internetovým nakupováním atp.

### Eliza

* 1964 - 1966 vyvíjena profesorem Josephem Weizenbaumem působícím na Massachusettském institutu technologie
* parodie psychoterapeuta rogeriánské školy - přeměna pacientových výpovědí do podoby otázek, jež jsou pacientům pokládány
* významné procento testerů považovalo program za člověka
* http://epanel.cz/eliza/eliza.php

#### Fungování Elizy

* vstupní text je analyzován a vyhledají se v něm klíčová slova
* když je takové slovo nalezeno, je věta převedena podle transformačních pravidel přiřazených k onomu klíčovému slovu a je vygenerován výstupní text
* výstupní text bývá často kombinován s uživatelovou předchozí odpovědí

#### Problémy Elizy

* nalezení nejdůležitějšího klíčového slova
* výběr odpovídajícího transformačního pravidla a transformace samotná
* ustanovení mechanismu, který umožní Elize odpovědět srozumitelně, když se v textu nevyskytuje klíčové slovo

### Parry

* 1972 vytvořen Kennethem Colbym
* simuluje paranoidního schizofrenika
* vyhledá klíčové slovo ve vstupu a přiřadí mu nějaký vnitřní stav, dle vývoje konverzace např. hněv, strach, nedůvěra
* pokročilejší než Eliza

**12. Automatické rozpoznání emocí – techniky, jazykové zdroje**