**Počítačové zpracování přirozeného jazyka**

*1. Techniky word sense disambiguation a algoritmy.*

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

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

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

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

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

*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.*

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

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

# 1. Techniky word sense disambiguation a algoritmy

* ***WSD*** – lexikální disambiguace, nalezení významu slova v daném kontextu
* pro počítače je to AI complete problém (aby bylo možné tuto úlohu vyřešit, je potřeba vyřešit všechny problémy AI, jde o problém neřešitelný v polynomiálním čase)
* **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
* jde o ***klasifikační úlohu***
	+ jednotlivé významy tvoří třídy
	+ podle kontextu se rozhodujeme, do kterých tříd slovo na vstupu patří
* ***předpoklady***
	+ významy jsou diskrétní
	+ významů je konečný počet
	+ máme nějaký inventář významů
* ***aplikace WSD***
	+ strojový překlad
	+ inteligentní vyhledávání – information retrieval
	+ inteligentní korektor překlepů a další
* ***přístupy WSD –*** zpravidla ignorují teoretické, psychologické, logické aj. aspekty významu
	+ *hloubkové* – zahrnují znalosti o jazyce i o světě
		- náročnější, nejsou funkční, je třeba zpracovat a formalizovat příliš mnoho dat
		- zdlouhavé a drahé
		- jsou funkční v omezených doménách, protože jde o „malý svět“
	+ *povrchové* – bez dalších znalostí, počítají s okolím
		- v současnosti jsou používanější – jsou relativně levné
	+ *historický (expertní) přístup* – v jakých kontextech může slovo nabývat jakých významů? (např. kohoutek – v botanice rostlina, v chovatelství lopatková kost, v technice ruční uzávěr atd.)
		- velmi náročný
		- nelze jednoduše vyjmenovat všechny možné významy a kontexty
		- významy i kontexty neustále přibývají
* ***algoritmy založené na znalostech***
	+ *Leskův algoritmus* ***–*** slovo w, jehož okolí sdílí nejvíc slov s definicí (nebo s příklady užití) určitého významu, má význam si
		- porovnává vstupní text se definicemi pro daný význam, jako správný určí význam s největším počtem shodných slov
		- vylepšené verze přidávají váhy (TF-IDF, vzdálenost od daného slova)
			* čím vyšší TF-IDF má nějaké slovo v dokumentu, tím vyšší se předpokládá jeho relevance
			* TF-IDF je součinem tf (četnost znaku *t* v určitém dokumentu) a idft (inverse document frekvency – počet dokumentů, ve kterých se vyskytuje *t*, čím běžněji se slovo vyskytuje, čím je nižší)
* ***algoritmy strojového učení***
	+ *strojové učení (ML)* – algoritmy a techniky, které způsobí 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 ML jsou klasifikační úlohy
	+ *pravidlové* – typické a intuitivní pro lidi, příkladem je hra „myslím si zvíře“
		- uplatňuje se zde redukce – pomocí otázek redukuji množinu správných odpovědí
		- strategie kladení otázek – neklást otázky, které dělí množinu odpovědí na stejné části, je třeba ji výrazně zredukovat
		- r o z h o d o v a c í s e z n a m y – jsou jednodušší na implementaci
		- r o z h o d o v a c í s t r o m y – přehlednější i při vyšší složitosti, pracuje se s nimi častěji
	+ *matematické*
		- pravděpodobnostní: naivní Bayesovský – předpokládá nezávislost znaků, je rychlý, počítá s podmíněnými pravděpodobnostmi
		- maximální entropie
		- podobnostní
	+ *promluvové*
		- one sense per discourse
		- one sense per collocation
		- redundance atributů
	+ *Yarkowského algoritmus strojového učení*
		- vezmi všechny výskyty slova *w* z korpusu včetně jejich kontextů
		- pro každý možný význam vytvoř sadu příkladů
		- vytvoř rozhodovací strom s pravděpodobnostmi pro další slova, která se vyskytují v kontextech
		- aplikuj tento seznam na celý korpus
		- nově zařazená slova obsahují další slova v kontextech
		- algoritmus můžeme upravit pomocí zařazení předpokladu one-sense-per-discourse
		- opakuj kroky 3–6
		- jakmile množiny přestanou narůstat, zastav
		- systém je nyní natrénovaný i na jiný korpus
		- závisí na první volbě kolokací, způsobu určení pravděpodobnosti, prahu pro pravděpodobnost a správnosti předpokladu one-sense-per-discourse
* všechny algoritmy závisí na inventáři a popisu významů (nevýhoda)
* všechny algoritmy WSD pracují s kolokacemi a určitým oknem, ve kterém kolokace sledují
* ***soutěž SENSEVAL*** – vyhodnocení systémů pro WSD, zatím chybí čeština

# 2. Regulární jazyky a regulární výrazy, CQL

* ***regulární jazyky***
	+ nejjednodušší formální jazyky
	+ je akceptovaný deterministickým i nedeterministickým konečným automatem
	+ může být popsán regulárním výrazem
	+ může být vygenerován regulární gramatikou
	+ všechny regulární jazyky jsou bezkontextové
* ***regulární výraz*** – řetězec popisující celou množinu řetězců (regulární jazyk)
	+ využití – vyhledávání textu, manipulace s textem
	+ skládá se z literálů textu, které se mají shodovat, a speciálních znaků, které nejsou součástí hledaného textu, ale slouží pro popis alternativ, množin, počtů výskytů a přepínačů
	+ *metaznaky*
		- **.** (tečka) – libovolný znak
		- [abc] – povolujeme jeden ze skupiny znaků *a, b, c*
		- [^ ] = odpovídá jednomu znaku, neuvedenému v závorkách ( [^abc] odpovídá libovolnému znaku krom *a, b, c*)
		- skupinu znaků můžeme zapsat také jako interval, např. [1–5], [A–Z]
		- opakování určité sekvence – výraz uzavřeme do závorek (a) a za pravou závorku doplníme kvantifikátor, např *ko(ko)?* odpovídá *kos* a *kokos*
		- **|** - slouží jako oddělovač variant, mezi nimiž si můžeme vybrat, např. Petr|Pavel
		- kvantifikátory

|  |  |
| --- | --- |
| * ?
 | * minimálně 0krát, maximálně 1krát
 |
| * \*
 | * minimálně 0krát (maximálně neomezeno)
 |
| * +
 | * minimálně 1krát (maximálně neomezeno)
 |
| * {n}
 | * právě nkrát
 |
| * {m,n}
 | * minimálně mkrát, maximálně nkrát
 |
| * {m,}
 | * minimálně mkrát (maximálně neomezeno)
 |

* + - předdefinované skupiny znaků

|  |  |
| --- | --- |
| * \d
 | * číslice 0-9
 |
| * \D
 | * jakýkoliv znak kromě číslic 0-9
 |
| * \w
 | * znaky „slova” (ekvivalentní zápisu [a-zA-Z0-9\_])
 |
| * \W
 | * jakýkoliv znak kromě znaků „slova” (ekvivalentní zápisu [^a-zA-Z0-9\_])
 |
| * \s
 | * „bílé” znaky (mezera, tabulátor, znaky pro zalomení řádků)
 |
| * \S
 | * jakýkoliv znak kromě „bílých” znaků
 |

* + - příklady

|  |  |
| --- | --- |
| * a+
 | * sekvence písmen a (1 a více znaků)
 |
| * a\*
 | * sekvence písmen a (0 a více znaků)
 |
| * o?kov
 | * okov či kov
 |
| * tel(efon)?
 | * tel či telefon
 |
| * telef(on|ax)
 | * telefon či telefax
 |
| * [0-9]|[1-9][0-9]
 | * čísla 0 až 99
 |
| * \d{2}
 | * sekvence dvou číslic desítkové soustavy (00, 01, …,98, 99)
 |
| * [0-9a-fA-F]|[1-9a-fA-F][0-9a-fA-F]+
 | * hexadecimální čísla
 |
| * (19|20)\d{2}
 | * letopočty 1900-2099
 |
| * \d{2,6}
 | * sekvence dvou až šesti číslic
 |
| * [^ ,.]+
 | * neprázdná sekvence znaků mezi nimiž nesmí být mezera (), čárka (,) či tečka (.)
 |
| * ^P.\*
 | * řetězec, který začíná písmenem P za nímž následuje libovolný (i nulový) počet libovolných znaků
 |
| * \d+0$
 | * řetězec, který končí znakem 0 (nula), kterému předchází minimálně jedna číslice
 |
| * a+b
 | * ab, aab, aaab atd.
 |
| * a\+b
 | * a+b
 |

* ***CQL*** – dotazovací jazyk pro práci s korpusy, dotazy jsou zadávány ve formě [atribut="hodnota"]
	+ *atribut* – word, lemma, tag apod.
	+ *hodnota* – samotný výraz nebo vzor specifikovaný regulárním výrazem
	+ dotaz můžeme omezit na strukturní atributy (věta, doc apod.)
	+ dotaz na více pozic současně vznikne zřetězením dotazů, např. [lemma="mít"][][lemma="srdce"]
	+ další viz <https://www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying>

Corpora query language (CQL) je dotazovací jazyk, pomocí něhož jsou tvořeny jednotlivé dotazy na korpus. S jeho pomocí se v korpusu vyhledávají slova dle jejich atributů (většinou word, lemma, tag). Jazyk dovoluje i kombinování podmínek pro vyhledávání slova. Při vytváření dotazu se používá regulárních výrazů, které umožňují poměrně velkou flexibilitu v parametrech hledání slova.

Každá pozice (slovo) je ohraničena hranatými závorkami, v nichž se nachází specifikace jednoho nebí více atributů.

*[atribut=“value“]*  kde value je regulérní výraz

**Vyhledávání podle tvaru slova nebo slovního spojení:**

*[word=“holubí“]*

Korpus najde všechny výskyty slova holubí

K vyhledávání více než jednoho slova použít pajpu

[word = „admin|amidst“]

[word = „struggle|battle|fight“]

**Vyhledávání podle lemmatu (základního slovníkového tvaru):**

*[lemma=“holubí“]*

Korpus najde: holubího, holubí, holubích

**Vyhledávání podle morfologické značky**

Každá značka obsahuje 15 pozic, každá pozice odpovídá jedné morfologické kategorii

[tag=“….4.\*] hledáme-li všechny akuzativy

[tag=“ACYS.\*] hledáme-li všechna adjektiva ve jmenném tvaru mužského rodu životného i neživotného v singuláru

**Vyhledávání spojení:**

*[lemma="mořsk."][word="ryba|živočich"]*

Korpus najde: mořská ryba, mořský živočich

**Kombinace elementů**

Výraz vybere všechny tvary slovesa chodit následované dvěma až čtyřmi slovy (započítává se i interpunkce) a zakončené podstatným jménem ve třetím nebo sedmém pádě. Pro zkrácení zápisu se dále využívá značky (?) ekvivalentní zápisu {0,1} a [pos=”A”] ekvivalentní [tag=”A.\*”].

[lemma=”chodit”][]{2,4}[tag=”N…3.\*”|tag=“N…7.\*“]

Např. pila je buď sloveso nebo podstatné jméno. Dotaz, chceme-li najít pila jako sloveso:

[lemma = „pila“ & tag = „V..“]

**Řetězce elementů:**

Která předložka následuje za pila? [lemma = „pila“] [tag „PRP“]

Která předložka následuje za podstatným jménem pila? [lemma = „pila“ & tag = „N..“] [tag „PRP“]

**Vložení slova mezi řetězce:**

[lemma=““] [] [lemma = „approach“]

Číslo v hranatých závorkách {x} označuje, kolik nechat volných slov (počet period) mezi slovy učení a nuda: [lemma=“učení“] []{3}[lemma = „nuda“]

Závorky {1,3} určují rozsah – od jedné do tří. Tento dotaz hledá jedno, dvě nebo tři slova mezi slovy patra a koupit: [lemma=“parta“] []{1,3}[lemma = „koupit“]

**Vyloučení mezi řetězci:**

[lemma = „koupit“] [word != „alkohol“]

**Vyhledávání podle interpunkce:**

[word=“\,“][word = „which“]

**Najít pozici v korpusu (číslo tokenu):**

Najdi pozici 100 a 210: [#100|#210]

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

* návrh, implementace a aplikace automatických systémů pro překlad textů
* ***klasifikace strojového překladu***
	+ *podle přístupu*
		- p r a v i d l o v ý (rule-based, knowledge-based)
			* transferový
			* iterlingua
		- s t a t i s t i c k ý (SMT)
		- h y b r i d n í (HMT, HyTran)
	+ *podle interakce s uživatelem*
		- r u č n í s p o m o c í p o č í t a č e (MAHT)
		- a u t o m a t i c k ý s i n t e r a g u j í c í m p ř e k l a d a t e l e m (HAMT)
		- p l n ě a u t o m a t i c k ý
		- HAMT + MAHT = CAT
	+ *podle četnosti překladu*
		- d v o j j a z y č n é
		- v í c e j a z y č n é
	+ *podle směru překladu*
		- j e d n o s m ě r n é
		- o b o u s m ě r n é
* ***Georgetown experiment*** – první funkční prototyp strojového překladu, RJ-AJ
* ***strojový překlad v současnosti***
	+ sběr paralelních dat
	+ práce s miliardami slov
	+ Google Translate jako gold standard
	+ dobré výsledky na velkých párech (EN-SP, EN-FR)
	+ morfologicky bohaté jazyky jsou opomíjeny
	+ *EuroMatrix 2006–2009* – překlad všech párů jazyků EU, open source
* patří mez UI-kompletní problémy, k dispozici je velká výpočetní síla, zdají se být vhodnější statistické přístupy
* ***výzvy pro strojový překlad***
	+ *lexikální výběr* – výběr správného překladového ekvivalentu
	+ *volný slovosled*
* ***rule-based machine translation (pravidlový strojový překlad)***
	+ lingvistické znalosti formou pravidel
		- pro analýzu
		- převod struktur mezi jazyky
		- syntézu
* ***KBMT – knowledge-based MT***
	+ je důležité správně analyzovat kompletní význam zdrojového textu
	+ dělí se na
		- *přímý překlad* – nejstarší, např. Georgetown experiment, METEO
		- *systémy s interlinguou* – zdrojový jazyk je přeložen do abstraktní jazykově nezávislé reprezentace (interlingua) a z ní je generován cílový jazyk, např. Rosetta, KMBT-89
		- *transferové systémy* – přidává třetí transferový krok
* ***systém přímého překladu***
	+ hledá korespondence mezi zdrojovými a cílovými jazykovými jednotkami
	+ omezen na konkrétní jazykový pár
	+ skládá se z velkého překladového slovníku a monolitického programu pro syntézu a analýzu
	+ jednosměrný
* ***přístup pomocí interlinguy***
	+ zdrojový jazyk konvertuje do nezávislé jazykové reprezentace
	+ interlingua musí být jednoznačná
	+ z interlignuy se generuje cílový jazyk
	+ cílový a zdrojový jazyk nepřichází do styku
* ***transferové systémy***
	+ analýza se provádí pro určitou úroveň
	+ transferová pravidla převádí zdrojové jednotky na cílové
	+ převod na syntaktické úrovni umožňuje zavádět kontextové omezení
* ***proces analýzy pravidlových systémů*** – tokenizace → morfologická analýza → morfologická desambiguace → syntaktická analýza → sémantická analýza = lexikální desambiguace
* ***syntéza pravidlových systémů*** – vyčlenění obsahu → pořadí propozic → lexikální výběr → syntaktický výběr → uspořádání konstituent → koreference → generování povrchových struktur
* ***statistický strojový překlad***
	+ užívá teorie informace a statistiky
	+ *gisting* – není třeba přesný význam, stačí, má-li překlad nějaký užitek
	+ využití bilinguálních textů a překladových modelů
	+ *data pro SMT* – paralelní korpusy
		- LDC (Linguistic Data consorcium) – paralelní korpusy pro jazykové páry
		- Europarl – texty Evropského parlamentu
		- OPUS – různý původ
		- Acquis Communautaire – právní dokumenty EU
		- vícejazyčné stránky (Wikipedie)
		- Kapradí – Shakespearova dramata (FI)
		- InterCorp – beletrie, ručně zarovnané texty (ČNK, FF UK)
	+ *základy SMT*
		- s l o v a – základní jednotka, uplatnění Zipfova zákona (10 nejčastějších slov tvoří 30 % textu)
		- v ě t y – v různých jazycích se syntaktická struktura liší, je třeba přerovnávání
		- p a r a l e l n í k o r p u s y – základní datový zdroj
		- z a r o v n á v á n í vět – věty si neodpovídají 1 : 1
	+ *základy pravděpodobnosti pro SMT*
		- pravděpodobnostní rozložení – graf hodnot pro elementární jevy náhodné veličiny
		- podmíněná pravděpodobnost
		- Shannonova hra – pravděpodobnostní rozložení pro následující znak v textu se liší v závislosti na předchozích znacích
		- Bayesovo pravidlo
	+ *princip noisy channel* – původně korekce signálu na zašuměných kanálech, využití znalosti typu vznikajících chyb a informací o původní zprávě, komponenty:
		- jazykový model
		- překladový model
		- dekódovací algoritmus
	+ *jazykové modely* – pro plynulost výstupu (správný slovosled), pomáhají s WSD v obecných případech a pomocí kontextu
		- n – g r a m o v é m o d e l y – využití statistického pozorování dat, některá slova se často vyskytují ve dvojicích, jde o pravděpodobnostní modely pro předpovídání následujících prvků
		- m a r k o v ů v ř e t ě z e c a p ř e d p o k l a d – posloupnost slov modelujeme postupně slovo po slovu užitím pravidla řetězu, historii slov omezíme použitím markovova předpokladu, nejčastěji se používají trigramové modely
		- t r i g r a m o v ý m o d e l – používá se pro určení pravděpodobnosti slova dvě slova předcházející s pomocí odhadu maximální věrohodnosti (maximum likelihood)
		- dobrý model přisuzuje dobrému textu vyšší pravděpodobnost než špatnému
		- c r o s s – e n t r o p y (křížová entropie) – průměrná hodnota záporných logaritmů pravděpodobností slov v testovacím textu, odpovídá míře nejistoty pravděpodobnostního rozložení; čím menší, tím lepší
		- p e r p l e x i t a – jednoduchá transformace křížové entropie; čím nižší, tím lépe
	+ *vyhlazování jazykových modelů* – je třeba rozlišovat pravděpodobnosti i pro neviděná data, reálný počet n-gramů upravujeme na očekávaný počet
		- a d d – o n e – nadhodnocuje neviděné n-gramy, přidává koeficient 1
		- ad d – a l p h a – nepřidává koeficient 1 ale α
		- d e l e t e d e s t i m a t i o n – umělé vytvoření neviděných n-gramů použitím dalšího korpusu nebo trénovacího korpusu
		- G o o d – T u r i n g – upravení počtu výskytů v korpusu tak, aby odpovídal obecnému výskytu v text, používá se frekvence frekvencí, tzn. počet různých n‑gramů, které se vyskytují n-krát
		- i n t e r p o l a c e – využití n-gramů nižších řádů pro stanovení pravděpodobnosti n‑gramů vyšších řádů
	+ *výpočet překladové pravděpodobnosti* – využívá se paralelních korpusů, EM (expectation maximization) algoritmus
* ***překladové modely***
	+ potřebujeme lexikální překladové pravděpodobnostní rozložení, tzn. jak často se dané slovo překládá na jednotlivé své ekvivalenty
	+ zavádí se zarovnávací funkce – překlady si často neodpovídají v počtu slov ani slovosledu
	+ *IBM modely* – rozkládají překlad do menších kroků
		- E M a l g o r i t m u s – používá se k zarovnávání slov
			* inicializuj model
			* aplikuj model na data
			* uprav model podle dat (úprava zarovnání)
			* opakuj kroky, dokud je co zlepšovat
		- IBM-1 – lexikální překlad, neuvažuje kontext, neumí přidávat a vypouštět slova, všechna různá zarovnání považuje za stejně pravděpodobná
		- IBM-2 – přidává model absolutního zarovnání, překlad rozdělí na dva kroky, nejprve se přeloží lexikální jednotky, poté se podle modelu zarovnání přeskupí přeložená slova
		- IBM-3 – přidává model fertility – řeší problém, kdy se jedno slovo přeloží na více slov, nebo se nepřeloží vůbec, zavádí vkládání tokenu NULL
		- IBM-4 – přidává model relativního zarovnání, kdy změny pozic slov závisí na přecházejících slovech
		- IBM-5 – ošetřuje nedostatečnosti předchozích modelů
	+ *frázový překladový model* – překlad sekvencí slov, statisticky motivované fráze
		- slovo je nevhodná jednotka, často si odpovídá n : m slov
		- pomáhá řešit překladové víceznačnosti
		- je jednodušší
* ***dekódování*** – z exponenciálního množství překladů musíme vyhledat ten, kterému modely přiřazují nejvyšší pravděpodobnost, používá se heuristické vyhledávání
* ***chyby v překladu*** jsou způsobeny chybou v prohledávání (není nalezen nejlepší překlad v celém prohledávacím prostoru) nebo chybou v modelech (nejlepší překlad podle pravděpodobnostních funkcí není správný
* ***hodnocení kvality překladu***
	+ záleží na plynulosti, adekvátnosti a srozumitelnosti, ruční hodnocení je drahé
	+ *gold standard* – ručně připravené referenční překlady, kandidáty srovnáváme s referenčními překlady
	+ *přesnost* – correct/output-length
	+ *pokrytí* – correct/reference-length
	+ *f-score* – precision x recall / (precision + recall)/2
	+ *BLEU* – nejznámější a nejpoužívanější standard, n-gramová shoda mezi referencí a kandidáty, extra postih za krátkost
	+ *další metriky* – NIST (vážení shod n-gramu podle informační hodnoty), NEVA (úprava BLEU skóre pro kratší věty), WAFT (editační vzdálenost), TER (nejmenší počet kroků), METEOR (synonyma a morf. varianty)
* ***hybridní systémy strojového překladu*** – kombinace pravidlových a statistických, pravidlový překlad s post-editací statistickým systémem

Strojový překlad v ČR (od 50. let 20. st.)

• ČSSR – od r. 1957, Oddělení teorie strojového překladu FF UK (P. Sgall, P. Novák, B. Palek)

• r. 1960 překlad AJ – ČJ na čs. počítači SAPO (několik desítek slov, odborné texty, ustálená slovní zásoba),

• lingvisté z UK Petr Sgall, Eva Hajičová, Jarmila Panevová, Petr Piťha, Zdeněk Kirschner,

Výzkumný ústav matematických strojů – projekt APAČ, 70. léta, ČJ – AJ – projekt RUSLAN, 80. léta, ČJ – RJ

# 4. Morfologická analýza přirozeného jazyka a desambiguace, přístupy

**Morfologická analýza** jazyka – interní struktura slov, skládání slov z menších jednotek. Morfém – nejmenší jednotka, která může nést význam. Např. slovo *pří-lež-it-ost-n-ými* (*pří* je prefix označující blízkost, *lež* – lexikální kořen slova *ležet*, *it* – adjektivní derivační sufix, *ost* – substantivní derivační sufix, *n* – adjektivní derivační sufix, *ými* – gramatický afix instrumentálu plurálu).

Morfologická analýza klasifikuje (značkuje, *tag*) slovní tvary jednotlivých kategorií (**Part of Speech/PoS tags**). **Kategorie pro účely** analýzy můžeme dělit na dvě skupiny:

* ***lexikální kategorie*** – pojmenovávají věci, akce, myšlenky… → substantiva, verba, adjektiva, adverbia, …
* ***gramatické kategorie*** – vyjadřují vztahy mezi ostatními větnými členy – předložky, spojky, částice, anglické členy

Jazyky můžeme rozdělit na:

* **jazyky s jednoduchou morfologií** – např. angličtina, několik desítek kategorií
* **jazyky s bohatou morfologií** – hierarchický systém, kde vedle základních slovních druhů určujeme nejrůznější subklasifikace (pád, číslo, rod, osoba, druhy příslovcí, …) - celkově tisíce značek

Morfologická analýza provádí

* **rozpoznávání slovních tvarů** - nástroj se nazývá **morfologický analyzátor** (Part-of-Speech/PoS tagger)
* **lemmatizaci** – přiřazuje k rozpoznaným slovním tvarům základní tvar (lemma)
* charakterizuje morfo-syntaktické vlastnosti nalezených slovních tvarů
* kvalita morfologické analýzy ovlivňuje všechny následující analytické roviny

Úkol morfologické analýzy zahrnuje 3 podúkoly:

* vypsat všechny možné analýzy – klasický **morfologický analyzátor**
* vybrat jednu nejpravděpodobnější analýzu – značkovač (**tagger**)
* analýzy pro neznámé slovo podle koncovky – **guesser**

Brillův značkovač

* učí se podle trénovacích dat:

1. přiřaď nejčastější značku

2. zkontroluj, kde jsou chyby (podle trénovacích dat)

3. ohodnoť pravidla pro opravu chyb → vyber nejlepší → oprav zpětně chybné značky

4. opakuj, dokud se daří odvozovat dobrá pravidla

* používá **učení založené na transformacích** (transformation-based learning)
* analogie – malování obrazu: nejprve pozadí a pak přes něj stále drobnější detaily
* značkuje 36 různých PoS značek, úspěšnost přes 90 %

Algoritmicky popis české formální morfologie

V češtině nestačí pravidla podle obecných morfémů – je potřebné mít lexikon, který ke každému kmenu obsahuje jeho přiřazení ke vzoru.

**Morfologické (tvaroslovné) paradigma** – soubor tvarů ohebného slova vyjadřující systém jeho mluvnických kategorií. **Vzor** – reprezentace tvaroslovného paradigmatu paradigmatem určitého konkrétního slova.

Algoritmický popis:

* definice koncovkových množin
* definice vzorů prostřednictvím vzorových slov rozdělených na
	+ neměnnou část vzorového slova – **kmenový základ**
	+ proměnlivé části vzorového slova – **intersegmenty**
	+ **koncovkové množiny** obsahující utříděné seznamy všech přípustných koncovek vzorového slova spolu s jejich gramatickými významy

**Popis vzoru** – formální pravidlo, které specifikuje přípustné kombinace těchto komponent (segmentů) ohebného slova.

Segmentace slova pro potřeby algoritmického popisu

* segmentace od začátku slova
	+ a) segmenty se snadno formalizovatelným výskytem vázaným gramaticky:
		- negativní preﬁx *ne-*
		- superlativní preﬁx *nej*-
		- futurální slovesný preﬁx *po*-
	+ b) segmenty s nesnadno formalizovatelným výskytem vázáným sémanticky:
		- preﬁxy
		- první členy kompozit
		- preﬁxy *ni-, ně-* zájmen neurčitých a záporných
* segmentace od konce slova
	+ c) rozdělení slovního tvaru na **kmen** a **koncovku**
	+ d) další segmentace kmene na **kmenový základ** a **intersegment**

Efektivní implementace morfologického lexikonu – trie

****Struktura trie:

* uspořádaný strom nad danou abecedou A
* v každém uzlu je různé písmeno z abecedy A
* klíč je v trie uložen jako cesta od kořene
* výhody:
	+ sdílení společných preﬁxů
	+ v každém případě nalezení nejdelšího shodného preﬁxu

Jiná efektivní implementace ML – konečný automat

* použití mírně pozměněných volně dostupných knihoven pro práci s KA od Jana Daciuka – FSA library
* vstupní data se generují ze slovníku ajky převedeného do tvaru “slovo<TAB>lemma<TAB>značka” (cca 33 mil. řádků)
* data se dále upravují pro KA – slovo+zkr.lemma+značky
* v lemmatu – 1. písmeno je počet znaků, které se odtrhnou jako předpona, 2. písmeno je počet znaků, které se trhají od konce a ostatní znaky se přidají
* tím se sníží počet řádků na 6.7 mil. řádků, ze kterých se přímo generuje (a minimalizuje) konečný automat
* výsledný slovník má 4.3MB
* rychlost je cca o 1/4 lepší než u trie, velikost řádově srovnatelná



**Algoritmický popis deklinace a konjugace, lemmatizace**

* algoritmický popis českého tvarosloví zahrnuje deklinaci, konjugaci a stupňování a některé pravidelné derivační (tvarotvorné) procesy
* v češtině nestačí pravidla podle obecných morfémů – je potřebné mít lexikon, který obsahuje pro každý kmen přiřazení ke vzoru
	+ **morfologické (tvaroslovné paradigma)**: soubor tvarů ohebného slova vyjadřující systém jeho mluvnických kategorií
	+ **vzor:** reprezentace tvaroslovného paradigmatu paradigmatem určitého konkrétního slova
		- **popis vzoru:** *formální pravidlo*, které specifikuje přípustné kombinace těchto komponent (segmentů) ohebného slova
* v češtině je kombinatorika morfémů do značné míry pravidelná, a proto i systematicky popsatelná souborem formálních pravidel, které z gramatik známe jako vzory. Popisujeme pomocí deklinačních vzorů (ohýbání substantiv apod.) a konjugačních (ohýbání sloves)
* formální prostředky, které umožňují vhodně reprezentovat morfémové struktury českých slov: některé typy *konečných automatů* a *trie struktury*
	+ **výchozími datovými strukturami jsou:**
		- vzory (asi 830)
		- kmeny (cca 164 tisíc)
		- intersegmenty (cca 460)
		- koncovkové množiny (počet koncovek cca 127) a prefixy (cca 140)
			* **koncovkové množiny** obsahující utříděné seznamy všech přípustných koncovek vzorového slova spolu s jejich gramatickými významy
* **ohýbací vzory:** potřeba základní soubor vzorů (jak známe ze školy) rozšířit a zjemnit jejich klasifikaci
	+ zavedení dostatečného počtu podvzorů, které zachycují příslušné hláskové změny (vlk – vlci apod.)
* počítačový slovník českých kmenů, jádro automatického morfologického analyzátoru, v současnosti obsahuje cca 36 tisíc slovesných kmenů
* vlastní segmentace:
	+ nejprve se hledají prefixy *nej-* a *ne-*, ev. I další (*vy-, roz-, po-, na-*)
	+ v případě úspěchu se poté činí pokus najít kmen, pokud je ovšem neúspěšný, odtrhávají se koncovky od konce slova, dokud se kmen nenajde ve slovníku

Jednotlivých **modifikací morfologického programu** LEMMA se užívá v následujících softwarových produktech jako samostatného modulu:

* v textových procesorech
	+ korekce překlepů
	+ nabídka možných tvarů
	+ nabídka synonym a antonym (thesaurus)
	+ dělení slov
* v sázecích systémech:
	+ text apod. (hlavně dělení slov)
* ve fulltextových aplikacích využívajících **lemmatizaci**, tj. přiřazení základního tvaru k libovolnému vstupnímu dotazu
* OCR systémy
* Překladové programy a překladové elektronické slovníky (např. oboustranný anglicko-český a německo-český slovník Lingea Lexicon 2)

**Tagsety pro morfologickou analýzu (značky)**

* tagset: soubor pravidel a hodnot, které se můžou vyskytovat v tagu
* **poziční systém (ČNK):**
	+ Hajič, Hlaváčová
	+ značka kóduje hodnoty kategorií
	+ kategorie je určena pozicí
	+ pražský systém (15 pozic)
* **atributový systém:**
	+ dvojice atribut-hodnota bez ohledu na pořadí, brněnský systém
	+ přehlednější, úspornější, snadno rozšiřitelný, čitelné RE
* **heterogenní:**
	+ bratislavský
	+ vynechává prázdné pozice
	+ nejkratší značky, ale malá rozšiřitelnost a složitější programové zpracování
* **BNC tagset**
	+ pevná množina hotových značek
	+ např. AJ0 Adjective (general or positive); AJC Comparative adjective

**Problémy MA**

* neshody v systému
	+ poziční systém (ČNK)
	+ atributový systém
	+ heterogenní
	+ BNC tagset
* různé možnosti lemmatizace (*ženu* vs. *ženu*; podstatné jméno *žena* nebo *ženu* od slovesa *hnát*)
	+ různé SD, dublety, diachronie, variety
* různé možnosti volby gramatických kategorií a jejich hodnot

České 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 Šmerka, na principu konečných automatů, s novým mechanizmem vzorů

**Majka**

* nový morfologický analyzátor, ajka byla příliš složitá a nerozšiřitelná
* Pavel Šmerk
* **princip konečných automatů**
* analýza:
	+ vyhledávání tvaru v seznamu WLT
	+ seznam lze chápat jako konečný jazyk, musí být minimalizovaný
	+ minimalizovaný deterministický automat
	+ FSA
	+ rychlejší, používá je *Seznam.cz*, *IS*
* výhody:
	+ jednoduchost: průchod automatem je nezávislý na konkrétních datech; funkcionalita rozšiřována změnou datových souborů, nikoli kódu
	+ zrychlení
	+ oddělení popisu dat a analyzátoru
	+ důkaz, že pro čj není třeba specializovaný datových struktur nebo algoritmů

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)

**Analyzátor MORČE (MORfologie ČEštiny)**

* Jan Raab ÚFAL MFF UK Praha
* statistický tagger
* včetně desambiguace

**Morfologická analýza**

* nejnižší rovina zpracování jazyka v textové podobě
* pro každý slovní tvar by měla vrátit **lemma** a možné gramatické významy v podobě značky, může přiřazovat i více údajů, je-li tvar slovnědruhové nebo morfologicky homonymní
* největší problém – stanovení pravidel pro určení slovního tvaru v konkrétním větném kontextu
* ***značky*** – reprezentují gramatickou informaci
	+ *poziční systém –* značka kóduje jen hodnoty kategorií (pražský systém)
	+ *atributový systém –* dvojice atribut-hodnota bez ohledu na pořadí (brněnský systém), je přehlednější, úspornější a snadno rozšiřitelný, čitelné RE
	+ *heterogenní systém* **–** bratislavský, vynechává prázdné pozice, má závazné pořadí, první znak určuje slovní druh, další kódují kategorii i hodnotu, nejkratší značky, ale malá rozšiřitelnost a složitější programové zpracování
	+ *BNC tagset* – pevná množina hotových značek
* ***ajka***
	+ slovník, definice vzorů, koncovkové množiny
	+ slovní tvary lemmatu se rozdělí na společný základ a koncovky, lemmata mající shodné množiny koncovek patří k témuž vzoru
	+ mezi základem a koncovkou je ještě intersegment
	+ intersegment i koncovka můžou být nulové
	+ po odtržení intersegmentů a koncovek vzniká slovní základ, k němuž se připojují koncovky daného vzoru
	+ počítá s prefixy nej-, ne- i postfixy (-s)
	+ problémy
		- redundance popisu
		- nekonzistence při doplňování
		- formální, strukturní nekonzistence – možnost popsat tutéž věc různými způsoby
* ***nový formát dat***
	+ zůstává slovník a soubor vzorů, koncovky jsou uspořádány do vzorů
	+ spojuje se část značky společná pro celý vzor s částí specifickou pro danou koncovku
	+ po spojení základu s koncovkou se slovní tvar získá aplikací předdefinovaných pravidel
	+ možnost omezit použitelnost koncovek podmínkou na zakončení základu
	+ *od slovníku vzorů, ke slovníku rysů* – nepoužíváme vzory, ale jiné sémantické, strukturní a hláskové informace, skloňování podle slovotvorných přípon → implicitní pravidla pro typická zakončení základů nebo jejich rysů vyznačených značkou ve slovníku
	+ *přínosy nového formátu* – redukce redundance, vyšší lingvistická přijatelnost (řazení slov k tradičním vzorům, zachycení alternací, možnost popisu slovotvorných vztahů, rozlišení periferních a okrajových jevů)
* ***majka***
	+ ajka příliš složitá a proto nerozšiřitelná
	+ využití práce Jana Daciuka – minimalistické deterministické automaty
	+ rychlejší
	+ jednoduchost – průchod automatem je nezávislý na konkrétních datech, kód není třeba při úpravách měnit
	+ odděluje popis dat a analyzátor
	+ dokazuje, že pro češtinu není třeba speciálních datových struktur nebo algoritmů

# 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, …)

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

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

**Složkový strom**

* 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**
* modré věci neterminály/konstituenty – každý modrý uzel popisuje nějakou frázi, která může k něčemu být
* každému slovu je přiřazena nějaká složka (neterminál) = lingvistická intepretace, kterou zachycuji svou znalost o struktuře věty
* Chomsky psal o tom, jestli je to univerzální, jestli ve všech jazycích máme jmenné skupiny, slovesa… (oficiální čínština nemá sloveso, tuto funkci slovo nabývá až v kontextu
* popisuje vnořené fráze, dává jim označení
* složkový popis využívá synt
* fráze jsou nositeli lepšího významu než samostatná slova. Syntaktickou informaci můžeme použít jako vstup nějaké další analýzy. Aplikace je třeba identifikace frází → kvalitnější vyhledávání, odpovídání na otázky… (něco, co není fráze, na to se uživatel neptá)
* je třeba morfologická analýza, syntaktická analýza na ní závisí
* pozn.: deskriptivní adekvátnost – jak přesně popisujeme strukturu věty x explikativní adekvátnost

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)

**Závislostní formalismus**

* 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)
* informaci kóduje jako „čáry“ mezi vstupními slovy
* v základním stromě nejsou neterminály, jen slova a šipky mezi nimi
* závislost je vyjádřena čárami mezi slovy
* vyjadřuje to relaci závislosti a podřízenosti
* každý podstrom je jedna platná fráze
* u závislostního formalismu problém s doplňky (nelze závislost označit dvěma čarami – Pivo je dobré studené.)
* ze závislostního formalismu nejsme dobře schopni určit hranice vět – nevíme, ve které větě je to sloveso. Rozpoznávání hranic vět je dost složitý problém

Závislostní je starší

V současnosti se preferuje závislostní, hlavně v Americe

Složková analýza Chomsky

Závislostní – Šmilauer, Bloomfield

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.

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 podstatné 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 zezákladních 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 (preterminálních 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 vstupních gramatik

****

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

**Slovesné valenční rámce a jejich využití v syntaktické analýze**

* sloveso je klíčové ve formální analýze věty (může mít až 5 elementů)
* valenční rámce sloves
* **Valenční slovník českých sloves**
	+ zdroje:
		- počítačový slovník českých kmenů, jádro automatického morfologického analyzátoru, v současnosti obsahuje cca 36 tisíc slovesných kmenů
		- Slovník českých synonym (Pala, Všianský), SSČ
	+ výchozí soubor cca 10 000 českých sloves, později 12.000, přidány slovesa z DESAM => 15.000 položek
	+ VSČS orientován na algoritmický popis české syntaxe a její počítačové zpracování – je proto zachycen pomocí formální notace
* kolokabilita: obecná schopnost slova (a dalších jednotek) spojovat se v textu s jinými
* u sloves se mluví o valenci zejména proto, že ji lze vyjadřovat morfologickými prostředky, tj. pády
* otázka, zda je valence jev syntaktický nebo sémantický? Valence – syntaktická záležitost, protože je vyjadřována morfologickými prostředky – pády apod.
* ale schopnost slova kombinovat se v textu s jinými slovy je primárně dána sémanticky a chápeme ji jako významem determinovanou schopnost slova kombinovat se s jinými slovy

**Šest typů valenční vázanosti**

* **integrální, lexikálně determinován, nevypustitelná doplnění:** nepřipouštějí substituce argumentů blízkými synonymy a modifikace adjektivy, nevstupují do syntagmatických substitučních paradigmat; frazeologická spojení apod.
	+ symbol #
	+ *držet* # *hubu* apod.
* **obligatorní, nutná doplnění:** forma přímých a předložkových pádů a vedlejších vět uvozených
	+ předložkové pády: symbol & (*rozkázat komu & co*)
	+ větná doplnění: symbol $ (*rozkázat $ aby, co*)
* **fakultativní, nepovinná:** realizovaná přímými a předložkovými pády; vypustitelné
	+ symbol *?* (*dopisovat si = s kým ? o čem*)
* **střední doplnění:** široce determinovaná sémantickou třídou; nejčastěji přímý instrumentál s širokým nástrojovým významem, popř. výrazy časové, místní, popř. způsobové
	+ symbol: []; (*dopovat = koho [čím]*)
* **volná doplnění:** volná doplnění časová, místní a způsobová; nijak sémanticky nevyplývají z významu slovesa
	+ *platit = komu & co [čím] [za co] <kdy=dnes> <kde=v obchodě, jak=hotově>*
* **periferní doplnění:** částice různého typu; částice mající pragmatickou povahu
	+ *poslat ? co | <asi, možná, patrně> |*

**Využití**

* seznam 12 tisíc českých sloves může posloužit jako východisko k vytvoření řady slovesných tříd získaných na základě pádových příznaků (a jejich kombinací), s nimiž se slovesa pojí
* **znalost vazeb – výchozím předpokladem pro počítačovou analýzu vět** → potřeba mít datové zdroje s příslušnými informacemi
* třídění sloves s jejich valenčními vzorci je velice komplikovaný úkol, je potřeba napsat program; příliš komplexní věc
* může sloužit například k vytřízení sloves pohybu pomocí valencí
* pomocí valenčních vzorců bude možno získat širší sémantickou klasifikaci českých sloves, která bude velmi užitečná pro softwarové aplikace

**VerbNet**

* největší on-line slovník anglických sloves
* rysy: hierarchický, doménově nezávislý (vyvážený), široké pokrytí
* propojení do PropBank (90% pokrytí), WordNetu, FrameNetu

**FrameNet**

* on-line zdroj anglických slov - slovesa, podstatná jména, přídavná jména, předložky
* cílem je dokumentovat sémantické a syntaktické možnosti kombinace možných významů jednotlivých slov
* sémantické rámce, sémantické role

**PDT-VALLEX**

* valenční rámce / významy pro slovesa, podstatná jména a přídavná jména
* významy, které se vyskytly v textech v PDT (bottom-up přístup - důraz na pokrytí textu)

**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
* jedná
* přístupný prostřednictvím webového prohlížeče, v pdf formátu a jako xml soubor

Anglický valenční slovník **VerbNet** (8537 total verbs)

**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, obouvidová)
* sémantické třídy sloves
* způsob užití slovesa (základní, přenesené, idiomatické)
* reflexivita (refl. tantum, reciprocita, syntakt. refl., absol. synon.)
* doloženo konkrétními příklady

**Syntaktická analýza**

* odhalení povrchové struktury věty, je základem pro analýzu jazyka na vyšších úrovních
* úkolem je pro danou gramatiku a daný vstup dát všechny odvozovací stromy
* využití – hledání vztahů mezi slovy, identifikace frází v textu
* ***základní termíny***
	+ *lexikální symbol, lexikální kategorie* – speciální neterminál gramatiky, který se přímo přepisuje na terminálový řetězec znaků (*N → pes*)
	+ frázová kategorie – neterminální symbol gramatiky, tkerý nevyjadřuje lexikální kategorii (*NP* *→ ADJP N*)
	+ větný člen – lexikální nebo frázová kategorie
	+ větná struktura – strukturovaný popis větných členů
		- povrchová – derivační/složkový strom jako výsledek bezkontextové analýzy
		- závislostní – zobrazuje závislosti mezi větnými členy
		- hloubková – sémantická interpretace fráze – role větných členů
* ***notace syntaxe přirozeného jazyka***
	+ *závislostní formalismus* – strukturní vztahy jsou kódovány závislostmi mezi slovy na vstupu, závislostní stromy
		- jeden člen je označen ajko řídící a druhý jako závislý
		- uzly závislostního stromu odpovídají slovům
		- hrany stromu označují vztahy závislosti mezi slovy
		- např. PDT – Pražský závislostní korpus
			* 2 mil. slovních jednotek
			* rovina anotace – slovní, morfologická, syntaktická (analytická), sémantická (tektogramatická), aktuální větné členění, koreferenční vztahy
	+ *složkový formalismus* – strukturní vztahy jsou kódovány stromem odvození z gramatiky
		- skupiny slov tvoří větné jednotky, které jsou označovány jako fráze, a jako větné členy formují větu
		- N →NP, NP P → PP, V PP → VP
		- např. synt
	+ *parciální syntaktická analýza* – nezajímá nás kompletní strom, jen některé vztahy (VaDis, Word Sketches)
* ***automatická syntaktická analýza***
	+ hl. problém – víceznačnost
	+ možnosti
		- manuálně vytvořené gramatiky (bezkontextové, závislostní)
		- statisticky naučené gramatiky
		- statistické odhadování podoby stromu
	+ *gramatické formalismy*
		- bezkontextová gramatika (CKY, chart parser)
			* algoritmus CKY – používá gramatiku v Chomského normální formě, pro daný vstup délky *n* derivujeme podřetězce symbolů délky *q* na pozici *p*, řešení pomocí matice
			* tabulkové analyzátory
				+ obsahují dvě datové struktury – chart a agendu, které obsahují hrany (trojice, kterou tvoří začátek a konec hrany a pravidlo vstupní gramatiky)
				+ shora dolů
				+ zdola nahoru – začíná od terminálů (slov ve větě)
				+ analýza řízená hlavou pravidla – hlava pravidla je libovolný (určený) symbol z pravé strany pravidla, od ní začíná analýza
			* Tomitův zobecněný analyzátor LR
				+ zdola nahoru, zleva doprava
				+ zásobník je reprezentován acyklickým orientovaným grafem
				+ derivační stromy jsou uloženy ve sbaleném lese stromů
				+ udržuje si seznam aktivních uzlů zásobníku
		- závislostní syntax – aplikace algoritmu hledání nejmenší kostry
		- **definite clause grammar (Prolog) – rozšiřují CFG**
			* neterminálům můžeme přidat argumenty a tvořit tak složité termy
			* na pravé straně pravidla můžeme volat procedury (logické podprogramy)
			* generování i rozpoznávání, vyšší síla
* ***NLP Lab***
	+ *synt* – složkový analyzátor, bezkontextová gramatika + chart parser
	+ *SET* – hybridní, detekce konečných vzorků z textu, využívá parciální analýzy
		- proces analýzy je rozdělen do několika vrstev
		- p a t t e r n m a t c h i n g – postupně vyhledáváme vzorky v textu a vybíráme nejpravděpodobnější z nich
		- hybridní stromy
	+ *Dis/VaDis* – založen na DCG, použití při morfologické desambiguaci
	+ *IOBBER* – chunker, statistický rozpoznávač frází
* ***CKY***
	+ algoritmus syntaktické analýzy pro bezkontextové gramatiky
	+ analýza zdola nahoru
	+ gramatika musí být v Chomského normální formě (všechna pravidla jsou ve tvaru A → BC, A → α, S → ε a jsou nezkracující)
* ***Combinatory categorial grammar (CCG)***
	+ gramatický formalismus, generuje frázovou strukturu (složkový přístup)
* ***Head driven phrase structure grammar (HPSG)***
	+ lexikalizovaná nederivační generativní gramatika
	+ složkový (frázový) přístup
	+ obsahuje lexikon a gramatická pravidla
		- lexikon je strukturovaný
		- položky lexikonu jsou značeny typem
	+ slova jsou definována jednak fonetickou a jednak syntakticko-sémantickou formou

**Stromové banky (treebanks)**

* textové soubory tvořené větami, u nichž je vyznačena syntaktická struktura, např. ve tvaru syntaktického (složkového) stromu

**Notace syntaxe přirozeného jazyka**

**Závislostní formalismus**

* strukturní vztahy kódovány závislostmi mezi slovy na
* vstupu
* pražský korpus závislostních stromů PDT

**Složkový formalismus**

* strukturní vztahy kódovány stromem odvození
* z gramatiky
* brněnský analyzátor **synt**

**Parciální syntaktická analýza**

* nezajímá nás kompletní strom, jen některé vztahy
* např. systém VaDis, Word Sketches

**NLP**

**Synt**

* složkový analyzátor
* základ = bezkontextová gramatika + chart parser
* koncept metagramatiky, kontextové akce

**SET**

* hybridní analyzátor
* založen na detekci konečných vzorků v textu

**Dis/VaDis**

* založen na DCG
* použití při morfologické desambiguaci

**IOBBER**

* statistický rozpoznávač frází (chunker)

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

Uložení korpusu v počítači

Formáty korpusů

* archiv/kolekce
	+ různé formáty, podle zdroje/typu
* textové banky
	+ jednotný formát a základní struktura
	+ dokumenty/texty, základní metainformace
* vertikální text
* binární data v aplikaci
	+ pomocná data pro rychlejší zpracování
		- indexy
		- statistiky

Kódování znaků

* 8 bitů ≈ 256 znaků
	+ ASCII – základ 7 bitů
	+ kódování pro češtinu
		- ISO-Latin-2, Windows-1250, 852
* Unicode
	+ 32bitů na znak
	+ UTF-8
		- 1 až 4 byty na znak
	+ UTF-16
		- 2 až 4 byty na znak

Kódování metainformací

* escape-sekvence – speciální znak mění význam následujících znaků\n, \t, &amp;, <tag>
* SGML – • Standard Generalised Markup Language, ISO 8879:1986(E)
* XML – Extensible Markup Language, W3C, 1998
	+ struktura popsána v DTD
	+ elementy
		- počáteční, koncová značka
		- <doc>, <head>, </head>, <g/>
	+ atributy elementů/značek
		- <doc title="Jak pejsek …" author="čapek">
		- <head type="main">
	+ entity – &gt;, &lt;, &amp;, &eacute;

Standardy pro ukládání textů

* SGML/XML
* TEI
	+ Text Encoding Initiative (1994)
	+ TEI Guidelines for Electronic Text Encoding and Interchange
* CES, XCES
	+ Corpus Encoding Standard

Vertikální text

* jednoduchý formát i jeho zpracování
	+ každý token na samostatném řádku
	+ struktury formou XML značek
	+ značkování odděleno tabulátorem (různé atributy k dané pozici)
* podrobnosti na: <http://nlp.fi.muni.cz/> → Informace pro současné a potenciální spolupracovníky → Textové korpusy → Popis vertikálů

Indexování a prohledávání korpusu

*Nenašla jsem nic konkrétního v dostupných PLIN materiálech, tak alespoň popíšu, co je INDEX, jak se pomocí něj vyhledává apod. Analogicky by to mohlo fungovat i na korpusech… snad :)*

**Index** – někdy označován jako KEY (klíč), jde o databázovou konstrukci, která usnadňuje vyhledávání a dotazování v databázích; definován výběrem tabulky a konkrétního sloupce, nad kterými se má vyhledávání urychlit. Vytvoření indexu je náročnější na operační paměť a prostor na disku (při tvoření a zápisu indexu je ukládání do databáze mnohem pomalejší, vyhledávání je však mnohonásobně rychlejší).

Index se tvoří mj. **pomocí příkazu jazyku SQL** a datový server do něj poté uloží informace o rozmístění hodnot indexovaných sloupců. Při vyhledávání dotazu pak není tabulka prohledávána podle toho, jak je sestavena, ale pomocí informací, které jsou uloženy v paměťovém prostoru indexu, se rovnou přejde na relevantní řádky. Zjednodušeně **jde o rejstřík** (jako v knize), který sice neurčí přesné místo výskytu dotazu, ale **zmenší prohledávanou oblast** (odkáže na paměťové místo s hledanými daty).

Na první pohled se může zdát, že čím více indexů, tím lepší a rychlejší vyhledávání. Opak je však pravdou, a to kvůli zmiňované potřebě paměťového místa pro indexy. Při vytvoření dostatečného množství indexů by pak mohlo dojít k situaci, že paměť zabraná pro jejich chod je shodná s pamětí zabranou samotnými daty. Druhým úskalím je fakt, že každý **index zpomaluje práci databáze**, protože při jakýchkoliv změnách v databázi musí být provedena i **změna dat odpovídajících indexů** (např. přidání řádků do tabulky apod.).

Druhy indexů:

* PRIMARY (primární index) – tvoří ho sloupec (kombinace více sloupců) obsahující primární klíč, v každé tabulce se smí vyskytnout pouze jednou, dle konvence nazýván ID; jedná se o zvláštní případ druhu klíče UNIQUE
* UNIQUE (unikátní index) – podobně jako předchozí typ, rozdíl: indexů může být více (příklad: kromě ID záznamu o osobě můžeme požadovat i unikátnost sloupce s loginovým jménem osoby; kombinace rodového a druhového jména zvířat apod.)
* INDEX (též SECONDARY) – sloupce obsahující sekundární/druhotný klíč (též vedlejší index), umožňuje vyhledávání pomocí dalších sloupců – nejen pomocí primárních nebo unikátních indexů; na rozdíl od předchozích lze vkládat do tabulky záznamy, které nejsou v sekundárním indexu unikátní
* FULLTEXT – optimalizace full.textového vyhledávání, způsob provedení záleží na samotném databázovém serveru

**Korpusy**

* ***korpus*** – rozsáhlý vnitřně strukturovaný a ucelený soubor textu jazyka, elektronicky uložený a zpracovávaný, textová podoba, velký rozsah dat, který zaručuje dostatečnou reprezentativnost
* ***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
	+ s rostoucím počtem korpusů vzniká potřeba jejich standardizace
	+ *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
* ***přípravné kroky před zpřístupněním korpusu***
	+ *akvizice* – vlastní získání textu (elektronická podoba, skenování, pořízení nahrávek), součástí je také získání souhlasu s použitím textu nebo nahrávky
		- v případě w e b o v ý c h korpusů jde o automatické stažení (crawl) webových stránek s texty
	+ *konverze* – převod do potřebného formátu spojený s čištěním
		- cílem je standardizace – sjednocení různých formátů do jediného (zpravidla XML)
		- čištění – odstranění částí netextového charakteru (obrázky, grafy…)
		- detekce a odstraňování duplicit, částí textů v jiném jazyce, seznamů, tabulek
		- opravy chyb způsobené sazbou
		- přepis nahrávky
	+ *vnější anotace* – přidání informací o textech a autorech ve formě metadat
	+ *vnitřní anotace* – přidání lingvistické interpretace ke slovním tvarům v textu (lemmatizace, tagování, parsing, u paralelních korpusů zarovnání)
	+ *selekce –* výběr textů do korpusu, zajištění reprezentativnosti
	+ *indexace* – technický převod textů do zaindexované podoby, formát vytvářený pomocí korpusových manažerů nezbytný pro rychlé vyhledávání ve velkých souborech dat
* ***vnitřní struktura korpusu***
	+ *poziční atributy*
	+ *strukturní atributy –* hranice vět, odstavců
* ***formy uložení korpusu***
	+ *archiv (kolekce) textů* – nejvolnější forma
	+ *textové banky* – texty jsou uloženy v jednotném formátu, základní značkování
	+ *použití korpusového manažeru* – texty zakóduje do určité databáze a uživateli umožňuje prohlížení korpusu z různých úhlů
	+ *vertikální text* – textový soubor, kde je rozdělení pozic vyznačeno tak, že se každá pozice nachází na samostatném řádku
		- atributy jsou uvedeny na jednom řádku a odděleny od sebe tabulátorem
		- strukturní značky jsou každá na samostatném řádku
		- výhodou je jednoduchost – je čitelný ve všech textových editorech a programovacích jazycích
* ***značkování korpusů***
	+ *gramatické značkování (tagging)* – přiřazení značek slovních druhů každému výskytu slova v korpusu
	+ *morfologické značkování* – pomocí morfologických analyzátorů (ajka, majka)
	+ *syntaktické značkování*
		- s t r o m o v é b a n k y (treebanks) – soubory textu, kde je u vět vyznačena syntaktická struktura, např. pomocí syntaktického stromu, způsob analýzy je dán předem danou gramatikou
* ***korpusové nástroje***
	+ *konkordanční programy* – třídí a počítají objekty nalezené v korpusu, využívají značkování
	+ *značkovací programy (taggers)* – každému sovu v korpusu přiřadí gramatickou značku
	+ *korpusové manažery*
* ***korpusový manažer***
	+ speciální počítačový program, který umožňuje vyhledávání v korpusu
	+ pracuje buď lokálně (pak na počítači musí být uložen celý korpus) nebo na principu klient-server (korpus je uložen na vzdáleném serveru)
	+ *klientský program* – určen pro práci s korpus, nainstalován buď lokálně na počítači, nebo jde o webovou aplikaci
	+ *logická struktura korpusu*
		- korpus jako posloupnost pozic
		- pro každý korpus definujeme množinu pozičních atributů s nějakou hodnotou
			* atribut uchovávající vlastní slova textu
			* systémy gramatických značek
		- korpus může obsahovat značkování ve formě struktur, jsou nad pozicemi
			* plochá – nevnořuje se a nepřekrývá, rozděluje korpus na části, např. doc
			* stromová – lze je vnořovat do sebe
			* prázdná – pouhé oddělovače, např. pre, g
	+ *konkordance* – seznam všech výskytů jistého jevu ve zvoleném korpusu, formát KWIC
	+ umožňuje vytvářet *statistiky*
	+ je s ním možné pracovat pomocí *dotazovacího jazyka*
	+ *korpusové manažery v ČNK*
		- M a n a t e e – serverová část, autor Pavel Rychlý, slouží k vyhodnocování dotazů zadaných pomocí klientské části
		- B o n i t o – klientská část, kterou si uživatel lokálně nainstaluje, autor Pavel Rychlý, může běžet na různých operačních systémech, prostřednictvím dotazovacího jazyka umožňuje vyhledávat slova nebo slovní spojení, které zobrazuje formou konkordančních řádku (KWIC), umožňuje také použití statistických funkcí
		- S k e t c h E n g i n e – klientská část v podobě webové aplikace, umožňuje zobrazovat slovní profily (wordsketchees), ty jsou plně integrovány s kolokacemi, dále vytváření vlastních webových korpusů a práci s nimi
			* word sketches - seznamy kolokací pro jednotlivé gramatické vztahy
		- N o S k e t c h E n g i n e – klientská část v podobě webové aplikace, podobně funkce jako Bonito, nekomerční verze SketchEngine, používá ji ČNK
* ***konkordanční programy***
	+ speciální programy sloužící k vyhledávání slov a jejich kolokací
	+ např. WordSmithTools, MonoconcPro, ConcApp, AntConc, Multiconcor: the Lingua Parallel Concordancer, OCP, WordCruncher, TactWeb, WebAsCorpus
* ***indexace***
	+ pro přímý přístup k pozicím s danými slovy se využívá dodatečných datových struktur
	+ urychluje vyhledávání
	+ tvorba tzv. lexikonu – jednotlivé řetězce se očíslují a operace se provádí nad čísly místo nad řetězci
	+ *reverzní index* – pro každé slov (type) z korpusu obsahuje samostatně seznam pozic, na kterých se dané slovo vyskytuje
* ***tradiční textové korpusy***
	+ zdroje – nakladatelství, skenování knih, přepisy rozhovorů
	+ vyvážená reprezentace žánrů a stylů
	+ informace o datech (autor, rok vydání…)
	+ možnost opravy chyb
	+ obtížné získávání dat, vysoké náklady, problémy s autorskými právy
	+ Brown Corpus (1964), BNC (1991–4), Corpus of Contemporary American English (COCA, 1990)
* ***webové korpusy*** – iniciativa Web as Corpus
	+ velké množství dat, dokumenty různých druhů, aktuální podoba jazyka, snadná dostupnost, nízké náklady (výhody)
	+ získávají se automaticky z textů na webu
	+ neuspořádanost, nežádoucí obsah, chyby, víme, co stahujeme (nevýhody)

**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%3Akolokace) 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

**Příprava korpusu**

Obecně však lze vyčlenit následujících 6 fází, z nichž ne všechny jsou nezbytné

(1) ***Akvizice***, vlastní získání textu.

(2) ***Konverze***, převod do potřebného formátu spojený většinou s čištěním. **Cílem konverze** je především **standardizace**, **sjednocení nejrůznějších vstupních formátů do jediného**, kterým je zpravidla XML. Součástí konverze je zmíněné čištění, tj. odstranění částí netextového charakteru (obrázků, grafů atd.), poté většinou následuje řada dalších (polo)automatických čistících procedur: detekce a odstraňování duplicit, částí textu psaných cizím jaz. n. takových, které obsahují velké množství čísel, tabulek, seznamů apod.; obecně jde o části textu, jejichž přítomnost v korpusu by byla spíše kontraproduktivní. Jejich součástí mohou být také opravy některých chyb způsobených sazbou (dělení slov) n. nevhodným technickým zpracováním, které nebyly záměrem autora. V případě mluvených korpusů lze za konverzi považovat přepis nahrávky; pro zajištění kvality přepisu je žádoucí, aby prošel několika nezávislými kontrolami.

(3) ***Vnější anotace***, přidání informací o textech a autorech (mluvčích) ve formě metadat (bibliografických, sociolingvistických aj.). Vnější anotace bývá nedostatečná u webových korpusů, uváděno je zpravidla pouze URL.

(4) ***Vnitřní anotace***, přidání lingvistické interpretace ke slovním tvarům v textu. Vnitřní anotace představuje dodatečnou informaci, která nijak nemění původní text, a přitom rozšiřuje možnosti využití korpusových dat. Jde zejména o lemmatizaci, tagování (morfosyntaktické značkování), a případně také parsing; u paralelních korpusů je nezbytné zarovnání celků (zpravidla vět), které si v jednotlivých jaz. odpovídají.

(5) ***Selekce***, výběr textů do korpusu. Dochází k ní u reprezentativních a vyvážených korpusů, kdy je potřeba dodržet celkové zastoupení jednotlivých variet. Provádí se buď na závěr jako výběr z celého shromážděného materiálu, n. průběžně již při jeho získávání; to je žádoucí, zejména je‑li korpus sestavován z obtížně dostupných, cenných dat.

(6) ***Indexace***, technický převod textů do zaindexované podoby. Jde o formát vytvářený většinou korpusových manažerů, nezbytný pro rychlé vyhledávání ve velkých souborech dat.

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

**Počítačová lexikografie**

* ***lexikální jednotka*** je reprezentována lexikální formou a asociována s určitým lexikálním významem (izolovaný význam slova bez ohledu na význam věty a gramatické kategorie)
* ***počítačová lexikografie*** – vytváření strojově čitelných slovníků za účelem jejich využití v různých úlohách NLP, ruční tvorba nebo automatická extrakce
* lexikografie korpusová
* vytváření a zpracování slovníků na základě dat z jazykového korpusu, a to jak slovníků knižních, tak elektronických
* rozsáhlé textové zdroje → slovník budovaný na spolehlivějším, bohatším, věrnějším a odstíněnějším základě bez korpusu
* autentické použití slov
* korpus ovšem musí být kvalitně vytvořen z dostatečně reprezentativních dat, aby zachytil všechny významy slova ve správných kontextech
* slovník vzniká rychleji a je aktuálnější
* informace o frekvenci a kolokacích
* **problémy při tvorbě slovníků:**
	+ neúplnost slovníku - opomenutí nebo přehlédnutí některého hesla
	+ jaký zvolit přístup k vlastním jménům - vynechat nebo zahrnout?
	+ *ghost words* -- neexistující slova, ve slovníku omylem (př. *Dord* = density ('D or d'))
	+ rozlišení významu hesel (P.Hanks: *A serious problem for computer applications is that dictionaries compiled for human users focus on giving lists of meanings for each entry, without saying much about how one meaning may be distinguished from another in text.*)
	+ jak podchytit „základní“ význam (definice) slova při automatické extrakci z korpusových dat?
* rozlišování hesel
	+ **homografie** - vlastnost dvou nebo více termínů, které mají stejnou grafickou formu, ale rozdílnou výslovnost - např. *pro-udit* vs. *proud-it*
	+ **polysémie** - jednomu lexému odpovídají dva nebo více významů (v souvislosti s jejich *výskytem v různých kontextech*), např. *vyšel z místnosti* vs. *vyšel z předpokladu*
	+ **homonymie** - *šetřit* (= *spořit*) vs. *šetřit* (= *zjišťovat*)
	+ P.Hanks: *No generally agreed criteria exist for what counts as a sense, or for how to distinguish one sense from another*
* ***slovníky*** – soubor lexikálních jednotek (výkladové, překladové, synonymické, etymologické, …)
	+ *struktura slovníkového hesla* – lexikální forma, gramatické vlastnosti, definice, kolokace, příklady, odvozené lexikální formy (hnízdování)
	+ *kolokace –* v NLP jsou to multiword expression
	+ *slovníkové definice*
		- pomocí synonym
		- klasické (genus proximum, differentia specifica)
		- hyperonymie vztah nadřazenosti
		- troponymie – modifikace míry významu
* ***slovníky a počítače***
	+ 60. léta – ruční přepis do databází
	+ 1978 – *Longman Dictionary of Contemporary English*, strojová kontrola
		- první s omezeným slovníkem definicí, kontrolováno strojově
	+ 1980 – *COBUILD* – založeno na korpusu Bank of English
		- University of Birmingham + Collins
		- první slovník založený na korpusových datech
	+ 90. léta – vývoj specializovaných systémů pro tvorbu slovníků
	+ 1987 – TEI – Text Encoding Initiative – XML formát pro sémantický popis textových dokumentů
* ***style guide*** – pravidla, doporučení pro jevy a informace, které se ve slovníku zobrazují často, zajišťují konzistentnost a snadnou navigaci
* ***XML***
	+ značkovací metajazyk
	+ *DTD –* seznam elementů a atributu a vztahů mezi nimi (Document Type Definition)
	+ *XML Schema* – popis obsahu a struktury SML dokumentu, je možné určit vlastní typy obsahu, kontrola obsahu
	+ *XSLT* (eXtensible Stylesheet Language) – převod XML na jiné formáty
* ***Dictionary Writing Systems*** – aplikace pro tvorbu slovníků
	+ komerční – IDM DPS, iLex, Tlex
	+ *DEB (Dictionary Editor and Browser)*
		- platforma pro slovníkové aplikace
		- klient-server, knihovny, moduly
		- DEBDict, DEBVisDic, Internetová jazyková příručka
* ***lexikografické podklady***
	+ *důkazy o použití jazyka* – intuice, excerpta, výpisky, korpusy
	+ *intuice*
* ***lexikální databáze*** – podrobná strukturovaná jazyková databáze, doklady z korpusu, gram. údaje, valence, vzory, styl, podklady pro slovníky a výzkum (PraLeD, DANTE)
* ***struktura hesla:*** makrostruktura a mikrostruktura
* ***makrostruktura*** – heslář (+ předmluva a přílohy), heslo (lemma, entry term, heslové slovo), heslo (heslová stať)
* ***mikrostruktura*** – struktura jednoho záznamu pomocí slovníku, lze kontrolovat pomocí softwaru
	+ lemma a přidružené informace
	+ významy – číslo, dělení, definice, vysvětlivky, poznámky k použití, příklady, odkazy
	+ etymologie
	+ ilustrace
	+ příznaky
* ***elektronické slovníky*** – multimédia, více informací, vyhledávání a navigace

**DEB**

* Dictionary editor and browser – prohlížeč a editor slovníků
* platforma pro slov. apl.
* FI MU
* rozšíření Mozilla Firefox, klient-server architektura
* DEBDict
	+ obecný prohlížeč slovníků
	+ dostupné slovníky
		- PSJČ, SSJČ, SSČ
		- Bohuslav Havránek et al., editors. *Příruční slovník jazyka českého* 1935-1957
		- Bohuslav Havránek et al. *Slovník spisovného jazyka českého* 1960-1971
		- J. Filipec et al. *Slovník spisovné češtiny* 1995
		- J. Kraus, V. Petráčková, et al. *Akademický slovník cizích slov* 1999.
		- F. Čermák et al. *Slovník české frazeologie a idiomatiky I-IV* 1983
		- K. Pala and J. Všianský. *Slovník českých synonym* 2000
	+ dále:
		- Ajka
		- Český WordNet (sémantická síť)
* DEBVisDic
	+ zobrazení WordNetů
	+ WordSketche
* Internetová jazyková příručka
* DEBGrid
	+ webové rozhraní pro Global WordNet Grid
* TeDi
	+ nástroj použitý pro vícejazyčné glosáře výtvarných termínů
* Praled
	+ lexikální databáze

**Další**

* Longman Dictionary of Contemporary English (LDOCE) – three-level embedded structure for sense distinctions (homographs,senses,optional subsenses)
* Roget’s Thesaurus (1852)
* Cambridge International Dictionary of English
* DANTE (Database of ANalysed Text of English)
* COBUILD English Language Dictionary

# 8. Paralelní korpusy, zarovnání

* ***paralelní korpus*** – elektronicky zpracovávaná a uchovávaná databáze v několika (minimálně dvou) různých jazycích, přičemž dvě odpovídající si časti (odstavce, věty, apod.) jsou nějakým způsobem zarovnány a je jednoznačně určeno, ke které části textu jiného jazyka patří
	+ nejčastěji vznikají překladem z jednoho jazyka do druhého
	+ jsou odlišné od běžných korpusů jejich použitím a pracuje se s nimi jinými metodami a prostředky
	+ každý z jazyků můžeme brát jako samostatný korpus
* ***zarovnání***
	+ první systém vznikl v roce 1987 – zarovnání vět na základě pravděpodobnostní teorie
	+ rozvinuly se tři směry – lingvistický, statistický a prostorový
	+ *lingvistický přístup* – používá gramatiky, regulární výrazy, gramatické kategorie, je ale velmi drahý a náročný
	+ *prostorový přístup* – pracuje s rozložením slov ve větě, s blízkostí určitých slov (adjektiva + substantiva apod.)
	+ *statistický přístup* – využívá frekvenci slov v paralelních větách, v současnosti je nejvyužívanější
	+ nejlepších výsledků je dosahováno při kombinaci přístupů
* ***typy zarovnání***
	+ *zarovnání souvislých částí textu* – odstavců, vět, souvislých slovních skupin
	+ *zarovnání jednotlivých slov (pozic)* – jistá forma ukazatelů, které k sobě spojují vzájemné překlady
	+ snažíme se o nalezení vztahu 1 : 1, v praxi se však spíše setkáváme s 1 : n či n : 1
* ***zarovnávání vět*** – oproti zarovnání na slova má výhodu menší chybovosti, pro další počítačové zpracování by však vyžadovalo použití složitých algoritmů, aby mohl být paralelní korpus využit pro tvorbu slovníku nebo překlad
	+ využívá se interpunkčních znamének a jiných značek
	+ *kotevní slova* – slova nebo řetězce, u nichž je pravděpodobné, že budou shodné nebo velmi podobné v paralelních textech (názvy, značky, data, čísla, adresy…)
* ***zarovnání slov***
	+ *heuristický přístup –* získává zarovnání slovo pomocí informací nezávislých na textu, ale pouze na slově (tokenu), např. metoda *Cognates* – využívá etymologické podobnosti řetězců
	+ *statistický přístup* – závistí na množině neznámých parametrů, které jsou získány pomocí učení na tréninkových datech, používají se model IBM 1–6 (viz strojový překlad)
* ***Hansards*** – francouzsko-anglický paralelní korpus, 50 mil. slov, pochází z jednání kanadského parlamentu v 80. letech, de facto se stal gold standardem
* ***Multext*** – diskuse v parlamentu Evropského společenství z let 1994–95
* ***InterCorp*** – jazykové páry zarovnané na větách, 10–30 mil. slovních, součástí ČNK, cílem projektu je vybudovat paralelní synchronní korpus pro většinu jazyků studovaných na FF
	+ jádro korpusu je ručně zarovnané
	+ obsahuje také kolekce, které jsou zpracované a zarovnané automaticky
	+ čeština je pivot, každý cizojazyčný text má českou verzi
	+ nejnovější verze 7 z r. 2014 obsahuje publicistické a zpravodajské články ze stránek Project Syndicate a Presseurop, právní texty EU z korpusu Acquis Communautaire, zápisy jednání Evropského Parlamentu z let 2007–2011 z korpusu Europarl a filmové titulky z databáze Open Subtitles
	+ zarovnaný na věty
* ***Europarl –*** 11 jazyků, vznikl z jednání Evropského parlamentu
* ***1984*** – Orwellův román
* ***Google Books Ngrams*** – vyhledávání ve skenovaných knihách, pouze n-tice slov
* ***Kačenka*** – FI MU, anglicko-český, vytvořen s úmyslem umožnit analýzu překladu na uceleném textu, cca 3,5 mil. pozic, především literární texty, zarovnání na úrovni vět

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

* počítač se učí řešit úlohu třídy T se zkušeností E a s úspěšností měřenou kritériem P, jestliže se jeho úspěšnost se vzrůstající zkušeností zvyšuje
* je třeba určit druh znalostí, které se bude počítač učit, a jejich reprezentaci a navrhnout konkrétní algoritmus učení
* typicky ***aproximace nějaké cílové funkce***, která realizuje úkol, jenž chceme počítač naučit
* ***zkušenost*** může být
	+ *přímá* – zápis pravidel
	+ *nepřímá* – dobré postupy na základě trénovacích dat
* ***učení založené na konceptu*** – na základě pozitivních a negativních trénovacích příkladů
	+ něco jako vyhledávání – prostor možností (version space), v němž se nachází optimální hypotéza, hledáme takovou, která odpovídá cílové funkci na všech trénovacích příkladech
* ***dělí se na učení***
	+ *s učitelem* – vstupní trénovací data obsahují správné řešení či výsledek
	+ *bez učitele* – vstupní trénovací data správný výsledek neobsahují
	+ obě metody lze kombinovat
	+ *zpětnovazebné* – učení posilováním
* ***základní druhy úloh***
	+ *klasifikace* (rozdělení dat do tříd)
	+ *regrese* (odhad číselné hodnoty výstupu dle vstupu)
	+ *shlukování* (zařazení objektů do skupin s podobnými vlastnostmi)
	+ *ranking* – určuje pořadí datových bodů
	+ *učení strukturovaných dat* – učení syntaktických stromů, převod řeči na textový řetězec apod.
* výsledkem učení je počítačový model obsahující extrahovanou informaci a její parametry
* účelem je extrahovanou informaci využít v dalším algoritmu jako součást pro automatické řešení konkrétního problému
* využívá se zejména ve strojovém překladu, extrakci informací nebo dialogových systémech
* ***modely*** – rozhodovací stromy, algoritmus k-nejbližších sousedů, SV, perceptron, Bayesovské sítě, neuronové sítě
* ***markovovy modely*** – používají se často pro časové řady nebo kontextově závislé posloupnosti dat
	+ *skrytý markovův model (HMM)* – modeluje systém za předpokladu, že jde o Markovův proces se skrytými stavy
		- stav není pozorovateli viditelný výstup ano
		- každý stav má pravděpodobnostní vliv na výstup systému
		- posloupnost výstupů vypovídá o posloupnosti vnitřních stavů
		- rozpoznávání řeči, písma, gest, POS tagging
* ***rozhodovací stromy*** – induktivní algoritmy i v trénovacíh příkladech mohou být chyby, cílová funkce je reprezentována rozhodovacím stromem (if – then – else), postupuje od kořene k listům, v každém uzlu testuje hodnoty jednoho atributu a určuje další směr
	+ technika data miningu, identifikace klíčových položek, vyhledávání zajímavých segmentů
	+ identifikuje objekty a zařazuje je do tříd
	+ přehledná a snadno interpretovatelná technika
	+ učení s učitelem – vytváří se na základě množiny objektů zařazených do skupin
* ***support vector machines***
	+ v úloze klasifikace hledá nadrovinu, která v prostoru příznaků optimálně rozděluje trénovací data
	+ rozdělující nadroviny je popsána dvěma body – podpůrnými vektory a je lineární funkcí
* ***bayesovská síť***
	+ pravděpodobnostní mode., který využívá grafickou reprezentaci pro zobrazení pravděpodobnostních vztahů mezi jednotlivými jevy
	+ acyklický orientovaný graf, každý uzel odpovídá jedné náhodné veličině
* ***neuronové sítě*** – propojená množina jednotek (neuronů), každý vydá z několika hodnot na vstupu pouze jednu na výstupu
	+ orientovaný acyklický graf
	+ vhodné i pro složitá a potenciálně zašuměná data
	+ trénování trvá dlouho, ale klasifikace nových instancí je velmi rychlá
	+ *perceptron* – druh neuronu
* ***cross-validation*** – metoda zjišťování, jak moc bude model statistické analýzy ovlivňovat nezávislé vzorky dat (významné pro predikaci neznámých vzorků po předchozí klasifikaci známých vzorků)
	+ vstupní množina je rozdělena na podmnožiny – jedna je testovací, zbytek trénovací
	+ chybu může způsobit malé množství dat, velké množství parametrů dat

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

* v oblasti umělé inteligence jde o proces získávání, zaznamenávání a ukládání znalostí do strojem srozumitelné podoby a jejich následné využívání v rámci expertního systému
* ***znalosti jazykové*** – znaky jazyka, segmentace na věty, segmentace na slova, určení slovních druhů, syntax jazyka, seznam slov jazyka (slovník)
* ***znalosti o světě –*** na určité úrovni tuto znalost potřebujeme alespoň na úrovni vazby mezi slovy
	+ encyklopedické (hlavní město ČR)
	+ common-sense (co si obleču dneska)
* ***common sense***
	+ sdílená znalost, ne vždy v souladu s vědeckými fakty, např. v noci nesvítí slunce
	+ 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)
	+ **projekty** – CyC (reprezentace pomocí vlastního jazyka CyCL), OpenMind, Concept Net (syntaktická analýza OpenMind, propojení s Wiktionary), NELL (Never-ending Language Learning, prochází web a odvozuje)
* ***implicitní znalost*** – znalost, která není přímo vyjádřena v určitém zdroji, primárně skryta, ale je sdělitelná, obvykle je zahrnuta v jednání, způsobu řešení úloh, souboru dat
* ***explicitní znalost*** – zaznamenaná v určitém jazyce a dostupná přímo v určitém informačním zdroji, znalosti uložené v bázi znalostí
* ***deklarativní znalost*** – formálně verifikovatelná, obecně platná (např. mapa objektu)
* ***procedurální znalost*** – implicitní, méně obecná (např. dojdi k oknu)
* ***reprezentace znalostí*** obsahuje znalostní bázi a odvozovací pravidla
* ***znalostní báze*** – knowledge base (KB) obsahuje fakta, která jsou premisami v deduktivním uvažování
	+ lidmi čitelné – how-to, FAQ, recepty, návody, diagramy
	+ strojově čitelné ontologie, sémantické sítě, dbPedia, ConceptNet
* ***ontologie*** – jeden z modelů zachycení znalostí, označení domluvené terminologie pro určitou aplikační oblast, která umožňuje sdílení znalostí z této oblasti, umožňují formalizovat doménové znalosti
	+ *slovník* – glosář, inventář pojmů
	+ *taxonomie* – tezaurus, inventář relací
		- uzly jsou třídy (organismů, entit, …)
		- třídy jsou strukturované do stromu (podtřída, nadtřída)
		- uzly na stejné úrovni se vzájemně vylučují (implicitní předpoklad)
		- princip klasické slovníkové definice – genus proximum + differentia specifica
	+ *SUMO* – Suggested Upper Merged Ontology, největší volná ontologie, specializované domény (vojenství, ekonomie, geografie apod.)
* ***sémantické sítě***
	+ popisuje realitu jako objekty, které jsou navzájem v nějakých vztazích
	+ má přirozenou grafofou reprezentaci
	+ *uzly* – entity, objekty (třídy nebo instance), jednomu konceptu odpovídá jeden uzel
	+ *hrany* – vztahy mezi uzly (binární relace)
	+ vyznačují se dědičností vlastností
	+ zahrnuje vztahy
		- is\_a – objekt patří do určité třídy objektů
		- a\_kind\_of – vyjadřuje hierarchii tříd
		- part\_of – umožňuje vyjádřit, že objekt (třída objektů) je tvořen částmi
		- hypo/hyperonymie – nadtyp/podtyp
		- instance třídy, member of
		- holo/meronymie – část/celek
		- upřesnění akce – troponymie
		- příčina/následek
	+ např. WordNet, EuroWordNet, sítě z Wikipedie, dbpedia
	+ *dbpedia* – sémantiká síť automaticky vytvořená z Wikipedie
* ***rámce***
	+ umožňují dobře vyjádřit statické znalosti s nějakou hierarchií pojmů
	+ vazba mezi rámci se dá podobně jako u sémantických sítí zobrazit grafem
	+ uzly v grafu mají vnitřní strukturu
	+ pronikly do programovacích jazyků, kde se požívají jako objekty (objektově orientované programování
	+ největší význam má *dědičnost* – lze dědit položky i jejich hodnoty, obvyklé je dědění shora dolů
	+ *zapouzdření* – součástí rámce jsou kromě datových struktur i procedury pro práci s nimi
	+ můžeme je použít pro desambiguaci slov i celých vět
	+ můžeme použít pro doplnění implicitní (nezmiňované) znalosti
* ***skripty, scénáře –*** popisují typické situace, pořadí ve scénáři je chronologické

Reprezentace znalostí

* hledá způsob vyjádření znalostí počítačově zpracovatelnou formou (za účelem odvozování)
* vyvozování znalostí (reasoning) – zpracovává znalosti uložené v bázi znalosti (knowledge 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

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).

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)

* ***dialogový systém*** – počítačový program, jehož účelem je komunikovat s člověkem pomocí mluvené řeči
	+ široké možnosti uplatnění – automatizace hlasových kontaktních center, interaktivní zábava, výuka jazyků, pomoc hendikepovaným atd.
	+ typicky se skládá z:
		- rozpoznávání mluvené řeči
		- porozumění přirozenému jazyku
		- řízení dialogu
		- generování odezvy
		- syntézy mluvené řeči
	+ dialogový systém musí pracovat se spontánní řečí, která je typická používáním nespisovného jazyka a negramatických vět, přeřeknutími apod.
* ***komponent rozpoznávání řeči* –** převod spontánní řeči do textové podoby, nejčastěji se používají statistické přístupy založené na skrytých markovových modelech a umělých neuronových sítích
	+ model rozpoznávání řeči se dělí na
		- model akustický – to, jak se hlásky slov vyslovují
		- model jazykový – jak se slova řadí do vět
	+ dialogové úloze se připravuje většinou na míru na základě dat podobných řešené úlohy, nejlépe na základě anotovaných dat komunikace uživatele a prototypu vyvíjeného dialogového systému
* ***komponent porozumění přirozenému jazyku*** – převod textového přepisu řeči do sémantické reprezentace
	+ *formalismy reprezentace sémantické informace*
		- řečové (dialogové) akty – nejčastější
			* řečový akt se skládá z typu aktu a jeho atributů a hodnot
			* typ aktu reprezentuje základní význam promluvy (pozdrav, poděkování…)
		- sémantické rámce
		- lambda kalkul
	+ rozpoznání mluvené řeči je také složité kvůli vysoké chybovosti rozpoznaného textu, používají se *sémantické interprety*
* ***komponent řízení dialogu*** – reaguje na vstup uživatele a generuje systémový akt jako odpověď,
	+ má dvě části:
		- *model dialogu* – modeluje stav dialogu
		- *strategie řízení dialogu* – určuje následující akci dialogového systému na základě odhadnutého stavu
	+ *stav dialogu* reprezentuje všechnu informaci potřebnou k úspěšnému pokračování dialogu, tj. cíl uživatele a co bylo už řečeno a potvrzeno
	+ většinou je založen na využití ručně psaných deterministických algoritmů a stavových automatů
	+ *standard VoiceXML* – využívá stavových automatů, stavy automatu odpovídají stavům dialogu a přechody mezi stavy akcím dialogového systému, vývoj je nákladný a kvalita není na úrovni, aby umožňovala plynulou konverzaci
	+ *metody založené na částečně pozorovatelném markovském rozhodovacím procesu (POMDP)* – předpokládají, že stav dialogu včetně cíle uživatele nemůže být znám, počítá pravděpodobnostní funkci přes všechny možné stavy a na ní zakládá všechna rozhodnutí, strategie řízení navrhuje optimální odpověď systému, používá se zpětnovazební učení
		- zpětnovazebné učení je typicky dvou druhů
			* učení v interakci s uživatelem - nejpoužívanější
			* učení z korpusu dat
* ***komponent generování odezvy*** – transformuje řečový akt do textové podoby včetně stylových a možných charakteristik
	+ většinou založen na *šablonách* – ručně napsané věty pro různé řečové akty
		- některé části jsou nahrazeny proměnnými
		- při realizaci řeč. aktu se vybírá nejvhodnější šablona a proměnné jsou nahrazeny atributy z řeč. aktu
		- nevýhoda – malá flexibilita, při generování alternativních realizací se každá alternativa musí ručně zakódovat jako nová šablona
	+ rozvoj metod založených na mapování řečových aktů do syntaktických stromů, z nichž se vytváří povrchové formy
	+ variabilitu výstupu lze také zvýšit tvorbou statistických modelů generování založených na frázích
* ***komponent syntézy mluvené řeči*** – převádí promluvu v textové podobě na akustický signál
	+ *předzpracování textu* – používají se ručně tvořená pravidla a slovníky
	+ *syntézy zvukového signálu* – používá se
		- metoda výběru akustických jednotek
			* syntetizuje požadovanou větu pospojováním akustických jednotek frází, slov n. fonémů
			* musí mít rozsáhlou databází jednotek – vytváří se automaticky z nahraných promluv řečníka a jejich přepisů
			* je potřeba dbát na správnou prozodii věty
		- skryté markovovy modely
			* opírá se také o systém rozpoznávání řeči optimalizovaný pro daného řečníka
			* změna prozodických vlastností se dá dosáhnout manipulací parametrů modelu
* ***multimodální dialogové systémy*** – využívají dalších výrazových prostředků ke komunikaci (psaný text, gesta dotykových obrazovek, pohybová gesta zaznamenaná kamerou…)

**Chatboty**

* počítačové programy simulující psanou konverzaci s člověkem v přirozeném jazyce
* funkce: ne porozumět textu, ale snaha o vytvoření dojmu, že agent tématu konverzace skutečně rozumí
* snaha o splnění Turingova testu (dvě komory, v jedné je PC, v druhém člověk; soudce má na základě počítačového dialogu rozhodnout, kdo je kdo)
* Loebnerova cena: každoroční soutěž programů umělé inteligence (od roku 1991), založena na Turingově 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
* využívala známé psychologické skutečnosti, že většina lidí hovoří nejraději o sobě
* významné procento testerů považovalo program za člověka -> Eliza se dá považovat za první počítačový program, který prošel Turingovým testem, i když jen částečně
* 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
	+ určuje nejmenší možný kontext, který je ke klíčovému slovu potřebný
* 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

Oba roboti spolu dokonce vedli několikrát rozhovor

Další chatboti

* **Alice:** třikrát získala Loebnerovu cenu (2000, 2001, 2004). Vytvořena v r. 1995, představuje ženu, komunikuje v angličtině
* **Elbot:** angličtina, němčina, od r. 2000; mnozí jej považují za nejúspěšnějšího chatbota vůbec; úspěch tkví v používání žertů, v nichž naznačuje, že je stroj -> mate
* **Cleverbot:** učí se z předešlých konverzací, které si ukládá do databází, v nich pak vyhledává odpověď na současný vstup uživatele (v databázi vyhledává, jak uživatelé reagovali na Cleverbotův dotaz, a jejich reakci použije jako odpověď na aktuální vstup); vlastně zopakuje to, co na současný vstup odpověděli lidé v dřívějších konverzacích – odpověď tedy nevytváří program, ale lidé.

Princip chatbotů

* **natural language processing** (zpracování přirozeného jazyka): rozkládání vstupní repliky na jednotlivé rozpoznatelné části struktury věty
* **case-based reasoning**: vyhledávání klíčových frází a slov, které spouštějí kýženou, tedy co nejvíce lidské reakci podobnou odpověď
* **jazyk AIML (Artificial Intelligence Markup Language)**
* **vyhledávání podle klíčových slov**
	+ rozpoznává je: používá keyword ranking: cesta, s jejíž pomocí chatbot vybere nejlepší klíčové slovo ve své databázi
	+ nerozpoznává je: řekne, že nerozumí; změní téma, položí další otázku, o které něco ví; metoda vytahování starých témat
	+ mohou ztroskotat na překlepech a chybách -> používají **levensteinovy (editační) vzdálenosti**; metoda, jež dokáže stanovit slovo, které je nejvíce podobné slovu chybnému – určuje podobnost mezi dvěma řetězci znaků
	+ **další zpracování uživatelova vstupu:** nezpracovává uživatelův vstup, pouze zjišťuje další informace („řekni mi víc“), používání krátkodobé paměti, citace člověka přeformulují do zjišťovacích otázek; využívání encyklopedických databází (např. na otázku, kdo je Einstein si najdou odpověď); schopnost učit se během konverzace

# 12. Automatické rozpoznání emocí – techniky, jazykové zdroje

* ***postojová analýza*** – automatická extrakce subjektivní informace z textu nebo řeči
	+ cílem je detekce postojů a emočního naladění pisatele či mluvčího
	+ využití při průzkumech veřejného mínění, získávání zpětné vazby od zákazníků, monitoringu sociálních sítí, predikce trendů v marketingu apod.
* ***detekce subjektivity*** – základní disciplína, odlišení objektivního sdělení od subjektivně laděného textu
* ***klasifikace polarity*** – nejčastější úloha, určení pozitivní nebo negativní orientace hodnocení na úrovní věty nebo delšího úseku textu, je třeba zohlednit také negaci, která obrací orientaci hodnocení
* ***hodnotící konstrukce***
	+ *zdroj hodnocení*
	+ *hodnotící výraz* – bývají sdružovány ve slovnících hodnotících výrazů
	+ *cíl hodnocení*
* ***základní přístupy –*** detekce klíčových slova a jejich využití při určování polarity, klíčová slova lze generovat na zákl. ručně připraveného seznamu, pomocí slovníku synonym nebo pomocí korpusu
* ***pro získání hodnotících výrazů se využívají***
	+ *metody neřízeného učení*
	+ *metody řízeného učení –* jako trénovací data se používají internetová data anotovaná samotnými uživateli
* ***hodnotící výrazy*** – jsou považovány za doménově závislé, je vhodné je získávat pro každou doménu zvlášť