Matematická lingvistika: formální jazyky a automaty

1. Vysvětlete pojem neprojektivity v závislostních stromech.
2. Uveďte příklady dvou neprojektivních vět.

Nejpopulárnějším nástrojem pro zpracování morfologie přirozených jazyků byla v 80. letech minulého století tzv. Two-level morphology profesorů Kartunnena a Koskenniemiho.

1. Pojmenujte dvě úrovně reprezentace zmíněné v názvu teorie.
2. Na jakém základním formálním nástroji bylo zpracování morfologie založeno?
3. Uveďte alespoň dvě ze tří základních myšlenek tohoto mechanismu.

Matematická lingvistika: základní formalismy pro popis přirozených jazyků

1. Popište základní vlastnosti teorie Funkčního generativního popisu.
2. Vysvětlete pojem valence.
3. Popište základní fakta o tvorbě jazykových korpusů.
4. Rozdělte jazykové korpusy podle typu značkování, pro každý typ uveďte alespoň dva příklady.
5. Vyjmenujte a popište tři základní komponenty transformační gramatiky.
6. Vysvětlete pojem transformace v Chomského transformační gramatice. Kterými dvěma složkami je transformace definována?
7. Proč je nutné zavádět typy sestav rysů v unifikačních gramatikách?

Uveďte základní charakteristiky následujících formalismů pro popis syntaxe přirozených jazyků:

8. Lexical Functional Grammar
9. Tree-adjoining Grammar
10. Kategoriální gramatika

Základní datové typy:

11. Popište dva základní datové typy používané pro zachycení syntaxe přirozených jazyků, závislostní a složkové stromy.
12. Uveďte výhody a nevýhody obou z nich.
13. Popište problém neprojektivních konstrukcí a vysvětlete, jak se tyto konstrukce dají, příp. nedají popsat těmito dvěma typy stromů.

14. Unifikační gramatiky využívají speciální datový typ, tzv. Sestavu rysů (feature structure). Uveďte jeho základní vlastnosti.
15. Vysvětlete, jak funguje mechanismus unifikace dvou sestav rysů. Zdůvodněte, proč je při tvorbě unifikační gramatiky přirozených jazyků nutné používat typované sestavy rysů.
16. Nejvíce rozšířenou unifikační gramatikou byla Head Driven Phrase Structure Grammar (HPSG). Uveďte základní vlastnosti tohoto typu gramatiky.

17. Definujte sestavu rysů (feature structure) jako základní datovou strukturu unifikačních gramatik.
18. Vysvětlete operaci unifikace sestav rysů a zdůvodněte, proč je nutné používat typované sestavy rysů.
19. Stručně uveďte základní vlastnosti formalismu HPSG (Head Driven Phrase Structure Grammar).

Jedním ze základních konceptů teorie Funkčního generativního popisu je valence.

20. Jaký je základní rozdíl mezi aktantem a volným doplněním?
21. Uveďte názvy alespoň 3 z 5 typů aktantů používaných v teorii Funkčního generativního popisu. Sestavte českou větu, ve které budou tyto tři typy aktantů zastoupeny.
22. Definujte valenční rámec.

Matematická lingvistika: morfologická, syntaktická a sémantická analýza přirozeného jazyka

Lexikální sémantika:

1. Popište, jakými metodami můžeme popsat lexikální sémantiku jednotlivých slov přirozeného jazyka.
2. Podrobněji popište nejrozšířenější sémantické sítě a jejich kladné i záporné vlastnosti.
3. Uveďte aplikační oblasti, ve kterých sémantické sítě můžeme úspěšně použít.

4. Uveďte alespoň 4 základní požadavky, které by měl splňovat dobrý program na kontrolu překlepů, a zdůvodněte je.
5. Uveďte dvě nejčastěji používané metody (pro kontrolu překlepů) založené na využití slovníku daného jazyka a vysvětlete, pro jaké typy jazyků se hodí a proč.
6. Nedílnou součástí kontroly překlepů je také nabízení vhodných oprav. Uveďte alespoň dvě prakticky použitelné metody, pomocí kterých se uživatelům budou nabízet vhodné opravy.

7. Vysvětlete rozdíl mezi stemmingem a lematizací slov přirozeného jazyka a uveďte jazykový jev, kvůli kterému není používání stemmingu vhodné pro češtinu.
8. Co znamená pojem Universal Dependencies a proč jsou důležité?
9. Vysvětlete hlavní myšlenku algoritmu zásoby sdílených znalostí (Stock of Shared Knowledge).

Sémantika se v přirozených jazycích uplatňuje na mnoha úrovních, od významu jednotlivých slov až po význam delších textových úseků. Vysvětlete některé základní sémantické pojmy:

10. Vysvětlete pojem „ontologie“ ve zpracování sémantiky přirozeného jazyka.
11. Popište sémantickou síť Wordnet, její strukturu a historii, a uveďte alespoň dva příklady aplikace této sítě v různých oblastech zpracování přirozeného jazyka.
12. Vysvětlete pojem „anafora“ a uveďte základní kategorie anafory v textu.

Při systematickém zpracování morfologie přirozených jazyků lze postupovat různými způsoby. Nejčastější tři způsoby jsou založeny na různých základních jednotkách, jmenovitě na morfémech, lexémech nebo

slovech. Vysvětlete stručně, jak každý z těchto tří způsobů funguje a pro každý z nich uveďte příklad typu jazyků, pro který je vhodný.

Jazykové modelování

Vysvětlete následující pojmy

1. Metoda zašuměného kanálu
2. Jazykový model ve statistickém strojovém překladu
3. Vyhlazování

Matematická lingvistika: základy teorie informace

Pracujeme s textem v přirozeném jazyce a chceme automaticky klasifikovat slovní druhy slov bezprostředně následovaných slovesem. Na rozsáhlém vzorku dat bylo zjištěno, že

- pravděpodobnost, že nahodile vybrané slovo v textu je sloveso, je 1/8,
- pravděpodobnost, že slovesu bezprostředně předchází příslovce, je 1/4,
- pravděpodobnost, že slovesu bezprostředně předchází podstatné jméno, je 1/2.

1. Vypočítejte pravděpodobnost výskytu příslovce bezprostředně následovaného slovesem.
2. Náhodná veličina reprezentující slovní druh slova bezprostředně následovaného slovesem nabývá tří různých hodnot $Y \in \{N,D,X\}$, kde N je podstatné jméno, D je příslovce a X je libovolný jiný slovní druh. Vypočítejte entropii uvedené náhodné veličiny $\mathrm{H}(Y)$.

Házíme třemi pravými mincemi. Pro každou z nich platí, že pravděpodobnost, že padne panna, je 1/2. Výsledek hodu reprezentujeme jako hodnotu náhodné veličiny $\langle x_1,x_2,x_3\rangle$, kde $x_i \in \{P,O\}$.

1. Vypočítejte entropii rozdělení této náhodné veličiny.
2. Jak se entropie změní, jestliže právě jedna ze tří mincí bude falešná a padne na ní vždy panna?

Mějme dvě diskrétní náhodné veličiny X a Y. Obě nabývají čtyř různých hodnot z množiny $\{a,b,c,d\}$. Sdružené pravděpodobnostní rozdělení je následující:

	$X = a$	$X = b$	$X = c$	$X = d$
$Y = a$	1/8	1/16	1/16	1/4
$Y = b$	1/16	1/8	1/16	0
$Y = c$	1/32	1/32	1/16	0
$Y = d$	1/32	1/32	1/16	0

1. Rozhodněte, která z veličin X a Y má větší entropii. Odpověď zdůvodněte.
2. Vypočítejte podmíněnou entropii $H(Y|X = c)$.
3. Jaký vzorec použijete pro výpočet vzájemné informace $I(X;Y)$ na základě pravděpodobností uvedených v tabulce, tj.
   bez znalosti entropie?

Házíme hrací kostkou a hozené číslo z množiny $\{1,2,3,4,5,6\}$ interpretujeme jako hodnotu náhodné proměnné X. Předpokládejme, že X má rovnoměrné rozdělení. Dále uvažujme náhodnou proměnnou Y s hodnotami

sudé/liché a náhodnou proměnnou Z s hodnotami *true* (pokud padne číslo větší než 4) nebo *false* (pokud nepadne číslo větší než 4). Obory hodnot náhodných proměnných jsou shrnuty v tabulce

náhodná proměnná	hodnoty
X	$\{1,2,3,4,5,6\}$
Y	{*sudé*, *liché*}
Z	$\{true,false\}$

1. Jsou proměnné X a Y statisticky nezávislé? Zdůvodněte.
2. Která z proměnných X, Y , Z má největší entropii? Odpověď přesně zdůvodněte.
3. Určete vzájemnou informaci $\mathrm{I}(X;Y)$. Výsledek zdůvodněte.

1. Která z proměnných $X\,Y\,Z$ má největší entropii? Která má nejmenší entropii? Odpověď přesně zdůvodněte.
2. Vypočtěte podmíněnou entropii $\mathrm{H}(X|Z)$. Uveďte postup výpočtu.

Házíme dvěma hracími kostkami, které označujeme X a Y. Kostka X je dokonalá, tj. rozdělení hodnot z množiny $\{1,2,3,4,5,6\}$ je rovnoměrné.

Kostka Y je falešná tak, že sudá čísla mají dvakrát větší pravděpodobnost než lichá:

$\Pr\{1\} = \Pr\{3\} = \Pr\{5\}$ a $\Pr\{2\} = \Pr\{4\} = \Pr\{6\}$, přičemž $\Pr\{2\} = 2 \cdot \Pr\{1\}$.

Čísla, která padají, interpretujeme jako hodnoty náhodných proměnných X a Y . Podle čísel na kostce Y uvažujeme také binární náhodnou proměnnou Y_b s hodnotami *sudé* a *liché*. Rozdělení náhodných proměnných je shrnuto v tabulce:

náhodná proměnná	hodnoty	rozdělení
X	$\{1,2,3,4,5,6\}$	rovnoměrné
Y	$\{1,2,3,4,5,6\}$	nerovnoměrné
Y_b	{*sudé*, *liché*}	nerovnoměrné

1. Co je jednotkou entropie? Co je jednotkou *podmíněné* entropie?
2. Jsou proměnné X a Y statisticky nezávislé? Svoji odpověď zdůvodněte.
3. Co můžeme říci o entropii $\mathrm{H}(Y_b)$? Vyberte jednu z následujících možností a svoji odpověď zdůvodněte.

a) $\mathrm{H}(Y_b) > 2$ b) $\mathrm{H}(Y_b) < 2$ c) $\mathrm{H}(Y_b) = 2$