

1. Uved'te definici korpusu v moderním terminologickém slova smyslu (čím se liší korpus od „sbírky textů“) 6 b.

Korpus je soubor počítačově uložených textů (v případě mluveného jazyka - přepisů záznamu mluvy), který slouží k jazykovému výzkumu. Od sbírky textů se liší tím, že je složen ze vzorků (samples), které jsou vybrány podle předem stanovených kritérií tak, aby reprezentovaly jazyk. Korpus je rozsahově i obsahově vymezen a omezen, je uložen v elektronické podobě a obsahuje standardní reference. Tato kritéria sbírka textů splňovat nemusí.

2. Co znamenají zkratky 4 b.

NLP **Natural Language Processing**

MRF **Machine Readable Dictionary**

OCR **Optical Character Recognition**

SGML (XML) **Standard Generalized Mark-up Language (eXtensible Mark-up Language)**

3. Jaký je rozdíl mezi kognitivně plausibilním a kognitivně neplausibilním systémem?

Kognitivně plausibilní systém usiluje o vytvoření poznávacího (kognitivního) modelu, pro nějž je relevantní, jakým způsobem člověk řeší nějaký úkol pomocí inteligence, a používá jej jako bázi pro stroj, který má tento problém (úkol) inteligentně řešit. Takový systém často používá úplnou sadu pravidel, které explicitně formulují znalosti, které člověk implicitně používá, ve formě tzv. báze znalostí (knowledge base). Naopak systémy, které rezignují na kognitivní plausibilitu a jsou tedy kognitivně neplausibilní, se prostě snaží vytvořit model inteligentního chování, aniž by se přihlíželo k tomu, zda systém pracuje stejným způsobem jako člověk (inteligentně). Tyto systémy často používají surová kvantitativní data k vytvoření statistického modelu napodobujícího lidské chování.

4. Uved'te v bodech obecné zásady vytyčené G. Leechem, podle nichž mají být vytvářeny značky (nejméně 3, má jich být 7) 7 b.

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

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

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

Mělo by být jasné JAK a KDO anotaci provedl (JAK – ročně x automaticky x poloautomaticky, s postkorekcí x bez korekce) (KDO – počítačový program, anotátor - člověk)

Uživatel korpusu by si měl být vědom toho, že anotace nejsou nějakou nedotknutelnou neomylnou instancí. Anotace je pouze více či méně užitečným nástrojem. INTERPRETACE. Anotační schéma by mělo být založeno na široce schvalovaných a teoreticky nezátížených principech. Není na škodu i zjednodušující přístup.

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

5. Co je to korpusový manažer? 3 b.

KORPUSOVÝ MANAŽER je program, umožňující efektivně pracovat s počítačovým korpusem, tj. vyhledávat podle zadatelných kritérií (slovní tvar, značka, lemma) ve formě KWIC, vyhledané informace třídit a statisticky zpracovávat, vytvářet subkorpora, ukládat získané informace, využívat standardních statistických metod pro vyhledávání kolokací atd.

6. Co je to značkovací jazyk? 4 b.

Značkovací jazyk je jakýkoli jazyk, který vkládá do textu značky vysvětlující význam nebo vzhled jednotlivých jeho částí. Vzhledové značky se původně používaly jen pro formátování textu v nakladatelstvích - dodnes se používá formátovací jazyk TeX (formátování knih do tisku). Dalšími jazyky jsou troff, PDF, ...

Pro potřeby KL se používal jazyk SGML, dnes XML.

7. Co znamená zkratka TEI ? 4 b.

Text encoding initiative. Jedná se o aktivitu sponzorovanou hlavními vědecky orientovanými asociacemi zabývajícími se využitím komputerů v humanitních vědách. ACL (Association for Computational Linguistics), ALLC (the Association for Literary and Linguistic Computing), ACH (the Association for Computers and Humanities). Cílem TEI je vytvoření standardní implementace pro operace s počítačově čitelnými texty. TEI za tímto účelem používá již existující formu SGML (Standard Generalised Markup Language). Byl přijat proto, že je jednoduchý, jasný, formálně přísný a již mezinárodně uznávaný. Vlastním příspěvkem TEI je detailní návod k použití přísl. standardu.

8. Jaké jsou výhody anotovaných korpusů? 4 b.

Anotovaný korpus přináší lingvistické interpretace na různých úrovních. Využitelnost těchto interpretací je dvojitá. Jednak pro výzkum jazyka (vytěžování korpusů lingvisty). Jednak pro aplikace NLP – budování modelů jazyka na základě dat získaných z korpusů.

9. Jaký je rozdíl mezi stochastickou disambiguací morfologicky označovaného korpusu a disambiguací řízenou pravidly? 6 b.

Stochastická metoda spočívá v tvorbě tzv. trénovacích dat (část korpusu se ručně disambiguuje tak, aby všem automaticky označovaným tvarům byla přiřazena jedna jediná správná značka), poté se statistický program „naučí“ na „trénovacím korpusu, tj. „učí si jakousi představu“ o pravděpodobnostech přechodu mezi jednotlivými značkami a o jejich četnostech, kterou si uloží do vnitřních tabulek, a pak „aplikuje“ tyto znalosti při další disambiguaci již nedisambiguovaného korpusu.

Podstatou disambiguace řízené pravidly je intuitivní formulace řady dílčích pravidel opřených o znalosti syntaktických konfigurací v PJ. Přístup je kognitivně plausibilní.

10. K čemu může sloužit kvantitativní analýza dat získaných z velkých jazykových korpusů? 4 b.

NLP: stochastická disambiguace, automatická analýza založená na statistických metodách. Frekvenční seznamy, výzkum kolokací - lexikografie.

CELKEM BODŮ:

Maximálně : 48

PROSPĚL: 48-32

NEPROSPĚL: 32 a méně