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

- vzorky (sampling) a reprezentativnost
- konečná velikost (omezený a vymezený rozsah)
- strojově čitelná forma (MRF)
- standardní reference

Korpus je soubor počítačově uložených textů (v případě mluveného jazyka - přepisů záznamu mluvy), který slouží k jazykovému výzkumu.

2. Co znamenají zkratky 4 b.

NLP **N**atural **L**anguage **P**rocessing (strojové zpracování přirozeného jazyka)

MRF **M**achine **R**eadable **D**ictionary (elektronicky čitelná forma)

OCR **O**ptical **C**haracter **R**ecognition (metoda převodu textu ve formě obrázku do formy znakové)

SGML (XML) **S**tandard **G**eneralized **M**ark-up **L**anguage (**eX**tensible **M**ark-up **L**anguage) – značkovací jazyk

3. Uveďte 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 nezatížených 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.**

4. Stanovte si kriterium a vyjmenujte podle něj typy korpusů 6 b.

časové hledisko: synchronní – diachronní

jazyk: psaný – mluvený

texty: obecný – autorský

5. Co je to korpusový manažer? 3 b.

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

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 formátovcí 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 Computional 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. Jaký je rozdíl mezi surovým a anotovaným korpusem? 4 b.

Surový korpus neobsahuje přídatné informace týkající se lingvistické interpretace na různých úrovních lingvistické analýzy textu. Anotovaný korpus takové interpretace obsahuje. Nejčastějším typem anotací jsou morfologické anotace (značky, tagy), informace POS, menší korpusy jsou anotovány syntakticky (tree-banks).

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

Stochastická metoda spočívá v tvorbě tzv. trénovacích dat (část korpusu se ručně disambiguuje tak, aby všem automaticky označkovaným tvarům byla přiřazena jedna jediná správná značka), poté se statistický program „naučí“ na „trénovacím korpusu, tj. „učiní 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.
Lingvistika: lexikologie, kolokace, lexikografie : frekvenční slovníky .

CELKEM BODŮ: 48

48-46 b.....A
45-43 b.....B
42-40 b.....C
39-37 b.....D
36-32 b.....E
31 b. a méně .....F