

Výberový slovník termínov z počítačovej a korpusovej lingvistiky

Mária Šimková

analýzátor a generátor tvarov (word form analyzer and generator)

softvérový nástroj (program) na analýzu a generovanie tvarov slov, obsahujúci slovník slov zatriedených do vzorov s úplnými paradigmami

anotácia (annotation)

1. pri počítačovom spracovaní textu priradovanie (bibliografických a lingvistických) informácií (metadát) jednotlivým častiam textu vo forme značky al. iného ustáleného zápisu; tagovanie, značkovanie

externá anotácia – priradovanie bibliografických informácií a informácií o štýlovo-žánrovej charakteristike textu (dokumentu)

interná anotácia – morfológická anotácia – priradovanie informácií o slovnodruhovej príslušnosti a morfológických kategóriách (tvaru) slova, textovej jednotky v kontexte

– syntaktická anotácia – priradovanie informácií o syntaktických (vetnočlenských, štruktúrnych a výpovedných) funkciách slova, textovej jednotky vo vete

2. výsledok tejto činnosti

anotátor (annotator)

1. osoba ručne vykonávajúca anotáciu (bibliografickú, štýlovo-žánrovú, morfológickú, syntaktickú a i.)

2. softvérový nástroj (anotačný program) na automatizovanú anotáciu (najčastejšie morfológickú)

dezambiguácia (disambiguation)

1. v počítačovom spracovaní automatizované odstraňovanie viacznačnosti, homonymie tvaru a priradovanie jednoznačných kategórií

– pravdepodobnostná dezambiguácia (probabilistic d.) – založená na štatistických metódach a teórii pravdepodobnosti

– pravidlová dezambiguácia (rule-based d.) – založená na (lingvistických)
pravidlách

2. výsledok tejto činnosti

excerpcia

1. excerpovanie, ručné vypisovanie skúmaných jazykových prostriedkov a javov zvyčajne
s (krátkym) kontextom a stručným bibliografickým záznamom na excerpčné lístky
usporadúvané a ukladané do kartotéky

2. výsledok tejto činnosti

externá anotácia # anotácia 1.

formálna lingvistika # lingvistika

frekvencia (frequency)

častosť, početnosť výskytov, použítí (skúmaných) prostriedkov v texte

hľadané slovo # kľúčové slovo

homonymia (homonymy)

v širokom zmysle rozdielnosť významov pri totožnosti tvaru

homonymum, homonymá (homonym)

v širokom zmysle slovo, jeho tvar al. iný jazykový prostriedok s dvoma al. viacerými
rôznymi významami

interná anotácia # anotácia 1.

inventár (inventory)

súpis, súbor jazykových jednotiek rovnakého typu

jazyková jednotka (language/linguistic unit)

najmenšia jednotka jazykového systému na konkrétnej rovine

kartotéka

súbor excerpčných lístkov zoradených podľa stanovených kritérií (najčastejšie podľa abecedy)

klúčové slovo (keyword)

hľadané slovo al. reťazec znakov zapisovaný do vyhľadávacieho okienka a po vyhľadani zobrazený zvyčajne uprostred kontextu – konkordančného výpisu

kolokácia (collocation)

spoločný výskyt (súvýskyt) lexikálnych jednotiek ako zvyčajný a do určitej miery predpovedateľný jav; spoločný výskyt lexikálnych jednotiek nad rámec početnosti výskytu očakávaného pri náhodnom rozdelení

konkordancia (concordance)

výpis kontextov klúčového, hľadaného slova al. reťazca znakov s jeho zvyčajným zobrazením spolu s určitým kontextom

kontext (context)

relevantné (textové) okolie skúmanej jazykovej jednotky

konverzia (conversion)

1. prevádzanie dát z jedného kódu do iného kódu
2. výsledok tejto činnosti

korektor

softvérový nástroj (program) na automatickú kontrolu pravopisu (spelling-checker) a/al. gramatiky (grammar-checker) alebo iných charakteristík (štylistický korektor)

korpus

1. elektronická databáza textov, zvyčajne veľkého rozsahu, spracovaná v jednotnom formáte a vybavená efektívnymi vyhľadávacími nástrojmi na získanie jazykových dát ako materiálového zdroja výskumu
- korpus hovorených textov – elektronická databáza hovorenej formy jazyka
 - korpus písaných textov – elektronická databáza písanej formy jazyka

– národný korpus – jednojazyčný korpus textov konkrétneho národného (jazykového) spoločenstva

– synchronný korpus – korpus jazyka v jeho súčasnej vývinovej fáze

– všeobecný korpus – nešpecifický, základný korpus zahŕňajúci široké spektrum jazykových štýlov a žánrov, vecných oblastí (domén), autorských generácií, vydateľských úzov, regiónov a pod.

2. súbor textov, na ktorých sa vykonáva výskum alebo spracovanie

korpusová gramatika (corpus grammar)

gramatika vypracovaná na základe štatisticky vyhodnoteného a lingvisticky opísaného materiálu z dostatočne veľkého a reprezentatívneho al. vyváženého korpusu

korpusová lingvistika

lingvistika

korpusový manažér (corpus manager)

softvérový nástroj (informačný systém) na rýchle vyhľadávanie a usporadúvanie údajov z korpusu a prehľadné zobrazovanie vyhľadaných kontextov a atribútov z jednotlivých úrovní anotácie a na získavanie a spracovanie rôznych štatistických informácií

kvantitatívna lingvistika

lingvistika

KWIC (key word in context)

zobrazenie výsledkov hľadania, pri ktorých je hľadané (kľúčové) slovo al. reťazec znakov zobrazené zvýraznene uprostred kontextu

lema (lemma, pl. lemmata/lemmas)

základný, slovníkový tvar, pri ohybných slovných druhoch reprezentujúci celú paradigm (pre slovesá neurčitok, pre menné slovné druhy nominatív singuláru, príp. mužského rodu)

lematizácia (lemmatisation/lemmatization)

1. (automatizované) priradovanie lem slovným tvarom v texte
2. výsledok tejto činnosti

lematizátor (lemmatiser/lemmatizer)

softvérový nástroj (program) na lematizáciu

lexéma (lexeme)

formálna a významová (aj viacslovná) jednotka slovnej zásoby reprezentovaná základným, slovníkovým tvarom (pre slovesá neurčitok, pre menné slovné druhy nominatív singuláru, príp. mužského rodu)

lingvistika (linguistics)

jazykoveda, veda o jazyku, jeho systéme a používaní, vývine, vzťahoch s mimojazykovými skutočnosťami a pod.

- formálna lingvistika – skúma formálne vlastnosti jazykových jednotiek a javov, zameriava sa na matematické modely gramatickej štruktúry jazyka
- korpusová lingvistika – odbor počítačovej lingvistiky, skúma a opisuje jazyk, resp. jeho použitie v reálnych kontextoch z rozsiahleho korpusového materiálu, vyhľadanych a spracovaných pomocou počítačových nástrojov a štatistických metód
- kvantitatívna (štatistická) lingvistika – odbor matematickej lingvistiky, skúma jazyk pomocou štatistických metód so zameraním na frekvenciu jazykových jednotiek a javov
- matematická lingvistika – skúma a opisuje jazyk matematickými metódami
- počítačová lingvistika – skúma jazyk pomocou počítačových nástrojov, aplikácie sa využíva najmä v oblasti strojového prekladu

matematická lingvistika # lingvistika

morfologický analyzátor a generátor tvarov # analyzátor a generátor tvarov

morfologická anotácia # anotácia 1.

morfologický anotátor # anotátor 1.

národný korpus # korpus

NLP (natural language processing) # počítačové spracovanie prirodzeného jazyka

počítačové spracovanie prirodzeného jazyka

1. systematická štruktúrovaná a uchovávaná analýza foriem a významov (funkcií) jazykových prostriedkov a javov pomocou výpočtovej techniky s potenciálnym vyústením do predikovateľnosti novej, neznámej jazykovej jednotky a generovania jej (možných) tvarov a vlastností
- ? 2. výsledok tejto činnosti

regulárny výraz (regular expression, skratka regexp, regex)

textový reťazec pozostávajúci z alfanumerických znakov a špeciálnych symbolov utvorený na základe určitých konkrétnych syntaktických pravidiel, ktorý opisuje istú množinu reťazcov vyhovujúcich danému regulárnemu výrazu, napríklad umožňuje vyhľadávať slovo vo všetkých jeho tvaroch v prípade nelematizovaného korpusu a/al. rôzne kombinácie slov, tvarov a iných reťazcov znakov.

reprezentatívnosť (representativeness)

reprezentovanie, zastupovanie celku (jazyka) vzorkou (korpusom), ktorá má rovnaké (úmerné, proporčné) zloženie a vlastnosti ako reprezentovaný celok

reťazec znakov (string)

1. v programovaní a formálnych jazykoch usporiadaná postupnosť symbolov z danej abecedy
2. dátový typ v mnohých počítačových jazykoch, zväčša umožňujúci ukladanie postupností alfanumerických znakov a iných symbolov

segment

uzuálne stanovený prvok ako (zvyčajne najmenšia) jednotka analyzovanej roviny

segmentácia (segmentation)

1. (automatizované) členenie celku na segmenty
2. výsledok tejto činnosti

synchronný korpus

korpus

syntaktická anotácia # anotácia 1.

štatistika

metóda opisu jazyka pomocou kvantitatívnych charakteristík a frekvencií skúmaných javov, založená na metódach matematickej štatistiky a teórie pravdepodobnosti

štatistická lingvistika # lingvistika

štýlovo-žánrová anotácia # anotácia 1.

tag (tag)

ustálená (dohodnutá) značka reprezentujúca konkrétne vlastnosti slova, textovej jednotky, textu

tager (tagger)

softvérový nástroj (program) na tagovanie, anotáciu

tagovanie (tagging) # anotácia 1.

tagset

súbor značiek a pravidiel na tagovanie, anotáciu

textová jednotka # token

token (token)

základná textová jednotka použitá v analýze alebo spracovaní textu, znak al. súbor znakov formálne oddelených od iného znaku al. súboru znakov napr. medzerou (tvar slova v texte, interpunkčné znamienko, číslca a pod.)

tokenizácia (tokenisation/tokenization)

1. automatizované členenie, segmentácia textu na tokeny
2. výsledok tejto činnosti

unikátny výskyt # výskyt

všeobecný korpus # korpus

vyváženosť (ballance)

vyrovnané, rovnovážne zastupovanie celku (jazyka) vzorkou (korpusom), ktorá obsahuje rôzne zložky (najčastejšie štýly) v prijateľnom pomere

výskyt (occurance)

zaznamenané vyskytnutie sa, použitie jazykovej jednotky al. javu

– unikátny výskyt – jednotlivý výskyt jazykovej jednotky al. javu, zastupujúci množinu všetkých výskytov

značkovanie # anotácia 1.