Korpusy pro překladatele (španělština)

Jarmila Fictumová
KAA FF MU

Korpusy

- **JEDNOJAZYČNÉ**: cizojazyčné, české; všeobecné, specializované (dle žánru či odborné, ad hoc)
- **DVOJJAZYČNÉ**: paralelní (= překladové); srovnatelné (pro vyhledávání odborné terminologie)
- **STUDENTSKÉ** (neboli žákovské – tj. *learner*): jednojazyčné; překladové (výuka překladu)

ZÁKLADNÍ TERMINOLOGIE UŽÍVANÁ V KORPUSOVÉ LINGVISTICE

- **ANOTACE** (značkování; tagování): přidělení explicitní lingvistické informace k textu (větné členy a sémantická anotace)
- **CÍLOVÝ JAZYK (CJ):** jazyk, do kterého překládáme
- **CQL** (**CONTEXTUAL QUERY LANGUAGE**): vyhledávací jazyk
- **DIACHRONNÍ KORPUS:** studuje vývoj jazyka za určité delší časové období
- **KONKORDANCE:** bezprostřední kontext dané lexikální jednotky

Základní terminologie užívaná v korpusové lingvistice

- **Korpus:** obsáhlá sbírka autentických textů v elektronické podobě (psaných textů nebo transkripce mluveného jazyka), shromážděných podle určitých kritérií)
- **Korpusový manažer:** software, který nachází konkordance pro hledané termíny, nachází všechny jejich výskyty v daném korpusu
- **KWIC (Key word in context):** zobrazení slova v jeho kontextu; obvykle ve středu monitoru
- **Lemma:** slovní tvar, který je dohodou vybrán jako reprezentant všech souvisejících forem
- **Lexém (angl. type):** každé různé slovo v korpusu (např. *Žil u lesa u řeky* - 4 lexémy, ale 5 tvarů slova / tokenů)
- **Otevřený korpus:** korpus, který je pravidelně doplňován

ZÁKLADNÍ TERMINOLOGIE UŽÍVANÁ V KORPUSOVÉ LINGVISTICE

- **PARALELNÍ KORPUS:** různojazyčné texty – překlady, zarovnané většinou po větách, podobně jako překladové paměti
- **SROVNATELNÝ KORPUS:** různojazyčné texty, které nejsou překladem, ale mají některé společné rysy (typ textu, téma...)
- **SYNCHRONNÍ KORPUS:** nestuduje vývojové změny jazyka
- **TOKENY:** všechna slova obsažená v korpusu bez ohledu na tvar
- **VÝCHOZÍ / ZDROJOVÝ JAZYK (VJ):** jazyk ze kterého překládáme
- **ZAROVNÁVÁNÍ (ALIGNMENT):** nalezení a přiřazení odpovídajících úseků jinojazyčných textů v paralelních korpusech
- Více na Přehled základních pojmů korpusové lingvistiky

Korpusové nástroje

Informace z článku ***Jazykové korpusy pomáhají s psaním textů*** **(MUNI.CZ 14.1.2014)**

Zuzany Nevěřilové, *která působí v* ***Centru zpracování přirozeného jazyka*** *Fakulty informatiky MU a vyučuje v* ***Centru počítačové lingvistiky*** *Filozofické fakulty MU.*

Software s názvem **Sketch Engine** vyvíjí společnost Lexical Computing ve spolupráci s **Centrem zpracování přirozeného jazyka** Masarykovy univerzity už přes deset let. Všichni studenti a zaměstnanci mají k tomuto programu, založeném na jazykových korpusech, **volný přístup**. …

… Sketch Engine vypočítá slovní profil (takzvaný word sketch), který ukáže, s jakými slovy-partnery se kandidát vyskytuje společně, jak často a v jakých kontextech. …

… Ze slovních profilů pak dokáže Sketch Engine vypočítat na větších celcích (frázích) dobré partnery. Výsledkem takového výpočtu je tedy Tezaurus, pomocí kterého najdeme významově blízká slova. …

… Software ale obsahuje i řadu dalších pokročilých funkcí pro práci s vlastními korpusy (automatickou extrakci klíčových slov, subkorpusy podle délky dokumentu či atributů autora) nebo pro práci s vícejazyčnými (paralelními) korpusy. Sketch Engine nabízí v současnosti více než 400 korpusů pro 70 jazyků. Všechny funkce jsou **popsány v dokumentaci**.

Dokumentace ve španělštině

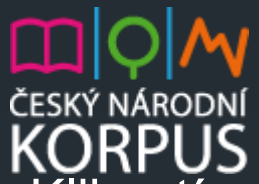
WEBOVÉ KORPUSY

(VÍCE INFORMACÍ v článku o přednášce ING VLADIMÍRA BENKA)

- **ŠPANĚLSKÉ KORPUSY**
- esTenTen11 (American, Freeling)
- esTenTen11 (European, TreeTagger)
- EUROPARL5, Spanish-English
- OPUS2 Spanish
- Spanish web corpus

- **ČESKÉ KORPUSY**
- **czTenTen12 [v. 7]**
- OPUS2 Czech
- CzechParl 2012
- Bruna Bohemica Minor (czes 14.04) 121 M

ČNK



Kliknutím na logo se zobrazí formulář registrace.

- MANUÁL PRÁCE S KORPUSOVÝM ROZHRANÍM

Lze očekávat, že bude zpřístupněn i webový korpus **Araneum Hispanicum**. Zatím jen Araneum Slovacum.

Cizojazyčné korpusy

Mark Davies: Professor, Corpus Linguistics, Brigham Young University

Corpus del Español

Google Books: Spanish

Další možnosti

IntelliText

- University of Leeds: tutoriál v angličtině

WebCorp Live

Pilotní provoz

hypal

Nástroj na vytváření paralelních korpusů a studentských korpusů jedno- i dvojjazyčných s označováním chyb

Dotazy?

Děkuji vám za pozornost!