4. Korpusová lingvistika II.
Základní pojmy

Četba: Marco Baroni, *Distributions in text*, celý text ke stažení na stránkách M. Baroniho: http://clic.cimec.unitn.it/marco/publications/hsk_39_dist_rev2.pdf

1) ZPRACOVÁNÍ TEXTŮ

- **značkování (tagging)**: texty (po konverzi do požadovaných formátů) jsou *anotovány*:
- vnější značkování:
  - správní anotace = údaje o textu, tj. autor, název, rok atd.
- vnitřní značkování:
  - strukturní = text je rozdělen do úseků - kapitola, odstavec, věta, slovo atd. strukturní značka vymezuje daný typ úseku, např. ...
  - lingvistické = každé grafické slovo je vybaveno lingvistickou anotací (např. slovní druh - tzv. POS tagging), morfologické značkování se řídí sekvencí pozic, např. 1. pozice slovní druh, 2. pozice detailní určení slovního druhu atd., např. „Ve“ = sloveso + tvar přechodníku přítomného)
  - Popis morfologických značek ke stažení na http://www.korpus.cz/doc/popis_znacek.pdf

- lemmatizace

- každá pozice je vybavena informací o „slovníkové podobě“ slova, tj. každý tvar je přiřazen lexému, k němuž patří.
- Jakmile je korpus lemmatizován, máme zásadní distinkci *lemma / výskyt* (angl. TYPE / TOKEN)

2) POJMY

- **Type frequency (V) = počet lemmat**
- **Token frequency (N) = počet výskytů** (absolutní / relativní - počet vyskytů na 100, na 1000, na 1000000 = i.p.m., apod.; absolutní frekvence / velikost korpusu * 100, nebo 1000, apod.). Relativní frekvence - umožňuje srovnání mezi různě velikými korpusy. (Sofistikovanější srovnání frekvence napříč korpusy bere v úvahu i různé složení korpusů, k tomu Michal Křen na http://ucnk.ff.cuni.cz/srovnani10.php)
- **Velikost korpusu (F)** = počet všech výskytů (tokens) /

lemma = INTERNET	absolutní frekvence	relativní frekvence (‰)
SYN2000 (F = ~ 121 mil.)	9506	**0,0786**
SYN2005 (F = ~ 122 mil.)	10475	**0,0855**
SYN2006PUB (F = ~ 361 mil.)	34419	**0,0952**

- **Rozložení** = graf rozložení všech výskytů v rámci celého korpusu

SYN2000 (15% beletrie, 25% odborná literatura, 60% publicistika)

SYN2005 (40% beletrie, 27% odborná literatura, 33% publicistika)

SYN2006PUB (pouze publicistika; od r. 1989 do r. 2004, první čárečka koresponduje s rokem 1994, max. odpovídá r. 2001)

3) VYHLEDÁVÁNÍ (KonText)

- konkordanční řádek = výsledek „dotazu“, který se zadává do dotazového řádku
- KWIC = key word in context
- Vyhledávání podle a) implicitního atributu (word/lemma atd.), b) pomocí morfologických značek + regulární výrazy
- word (slovní tvar) - „učil“
- lemma (lexém) - „učit“
- tag (podle morfologické značky) - [tag=“Ve“] (tj. všechna slovesa, která jsou přítomnými přechodníky)
- tag + regulární výraz - [(lemma=".*ář") & (tag="N.M.*")] (tj. všechna slova, která končí na -*ář* a jsou životnými substantivy)

(Vynikající příručka: *Jak využívat Český národní korpus*, Lidové noviny 2005)

4) VYHLEDÁVÁNÍ KOLOKACÍ A ZÁKLADNÍ STATISTICKÉ FUNKCE

- KOLOKACE = spojení dvou či více slov, které je nějak *pevnější* než obvyklé syntagma (pokusů o definici spousta; jeden z nich - M. Baroni, *Classificazione preliminare delle collocazioni*, kurs a.a. 2005/2006, vychází z E. Ježek, *Lessico*, Bologna, il Mulino 2005, s. 173-190; u nás srov. *Kolokace*, F. Čermák - M. Šulc (eds.), Praha, Lidové noviny 2006)
  a) VOLNÉ KOMBINACE (combinazioni libere): *mýt auto, modrý sešit* atd.
  b) TĚSNÉ KOMBINACE (combinazioni ristrette): *parkovat auto, blonďaté vlasy*
  c) VLASTNÍ KOLOKACE (collocazioni propriamente dette):
    i. kolokace typu *pioggia battente, bujná fantazie*
    ii. frazémy, idiomatické konstrukce typu *lune de miel*
    iii. konstrukce s kategoriálním slovesem (fr. *verbe support*): *fare una telefonata* atd.

- Ad a) volnost kombinace, nahraditelnost, kompozicionální sémantika, restrikce konceptuální (hledat X, mýt Y atd.) (Ježek, 2005, s. 175)
- Ad b) adjektivum / sloveso specifikuje vlastnost *inherentní* danému předmětu; specifikace se týká velmi omezeného počtu možností; adjektivum / sloveso jsou převážně monosémní, sémanticky specializované jednotky: *parkovat X, oblékat si Y, pasterizovat Z* atd., restrikce konceptuální + „implicazione sintagmatica di contenuto" (Ježek, op. cit., s. 176-177)
- Ad c) kolokace = „kombinace dvou či více slov, která je omezena takovou lexikální restrikcí, kvůli níž je volba jednoho slova (kolokátu) podmíněna volbou druhého slova (báze)." (Ježek, op. cit., s. 178)
- Rozdíl mezi TĚSNOU KOMBINACÍ a KOLOKACÍ? (podle Ježkové, op. cit. 179-180):
    - V sémantické implikaci (*parkovat* samo o sobě implikuje vůz, *oblékat si* implikuje oděv, atd.; versus *battente* neimplikuje *pioggia*, *bujná* neimplikuje *fantazii* atd.), a proto těsnou kombinaci nejspíše **uhádneme**
    - Kolokace je typem těsné kombinace jen díky tomu, co Ježková nazývá „solidaritou, která je stabilizována jen územ", a proto kolokaci nejspíš **neuhádneme** (jako rodilí mluvčí většinou ano, ale protože ji známe; M. Baroni: a co second language learners?)
- A co korpus? Hlavní myšlenka je, že rozdíl mezi těmito spojeními může být zachycen rozdílem ve *frekvenci*.
- Výsledky vyhledávání kolokací obsahují hodnoty tzv. MI-score a T-score.

- MI-score = mutual information; je hodnotou, která udává pravděpodobnost toho, že dvě slova /či více/ se vyskytnou současně vedle sebe;
- T-score je testem náhodného („random") rozložení výskytů (předpoklad toho, že korpus je tzv. *random sample*; jde o složitou otázku, nedávno k tomu Stefan Evert, *The Statistics of Word Cooccurences*, 2005):

    - Čím vyšší je hodnota MI-score, tím pravděpodobnější je, že jde o pevnou, ustálenou kombinaci.
    - Vysoká hodnota T-score naznačuje, že rozložení frekvencí není náhodné, tzn. že jsou zde tzv. *clustering effects*, např. to, že jde o pevnější spojení (slova jsou více u sebe, než by byla, kdyby bylo jejich rozložení nezávislé jedno na druhém, tj. zcela *náhodné*.

- Hypoteticky tedy:
    1) VOLNÉ KOMBINACE → nízké MI-score i T-score
    2) TĚSNÉ KOMBINACE → spíš vyšší MI-score (problematické, protože často je komplement implicitní - viz definice) a vyšší T-score
    3) KOLOKACE → vysoké MI-score (obvykle nad 10.00; problematické pro nízkofrekvenční jednotky)
- Příklady (ze SYN2000): 1) *mýt auto*, 2) *parkovat auto*, 3) *mhouřit oči* (rozdíly v hodnotách v závislosti na velikosti kontextu; zadáno shodně 0-3, tj. do 3 pozic napravo)
    1) f(mýt) = 667, f(auto) = 21154, f(mýt, auto) = 11, MI-score = 6.559, T-score = 3.281
    2) f(parkovat) = 439, f(auto) = 21154, f(parkovat, auto) = 36, MI-score= 8.873, T-score = 5.987
    3) f(mhouřit) = 93, f(oko) = 40597, f(mhouřit, oko) = 79, MI-score = 11.305, T-score = 8.885
- Co je pro cizince nejdůležitější? Nejspíš naučit se právě kolokace ...
- Příště: „case study" - morfologická produktivita a korpusová lingvistika