# Workshop korpus českého znakového jazyka (6)

Co určitě zařadíme do kategorie slovních druhů:

* Substantiva (SUBST)
* Adjektiva (ADJ)
* Číslovky (NUM)
* Slovesa (VERB)
* Příslovce (ADV)
* Předložky (PREP)
* Spojky (KONJ)
* Částice (PART)
* Specifické znaky
* Nepřímá pojmenování
* Pomocná slovesa (AUXILIÁRY)
* Konverzivní kategorie:
  + SUBST-VERB
  + ADV-ADJ

Co dále?

* dobré dodržovat kategorii konverzivních kategorií (SUBST-VERB, a ADV-ADJ)
  + při anotování se někteří z nás setkali s tím, že by danou kategorii použili
* STOJÍCÍ AUTA: patřilo by to do klasifikátoru, nebo do kategorie sloves? Jak by se to mělo vlastně přeložit
* rozhodovat se jestli je něco spíše sloveso, nebo spíše klasifikátor, by mohlo zkreslit následné vyhledávání
* U některých návrhů z předchozí hodiny (např. prstová abeceda) jsme se příliš neshodli v tom, jestli to do vrstvy slovních druhů začleňovat, jako řešení se nabízí vytvořit nové vrstvy, které by se vyplňovaly pouze v případech, kdyby to byla potřeba.
* Návrh tří dalších vrstev:
  + **Inkorporace** (vlastní vrstva)
  + pozn.: vrstva inkorporace → syntaktická kategorie (proto úplně nepatří do slovních druhů), další podtypy:
    - objektu
    - prostředku
      * vehiklu (dopravního prostředku)
    - číslovky
    - adverbia
  + **Další vyjadřovací prostředky**:
    - Prstová abeceda (v podstatě je to jenom způsob vyjádření)
    - Inicializace
    - ~~Kalk~~: je otázka jestli kalk začleňovat, nebo ne
      * asi moc široká kategorie, a začleňovat ho nebudeme
  + **Klasifikátor/specifikátor**
    - Označení jestli je někde klasifikátor, nebo specifikátor (začleněny v jedné vrstvě, protože se vzájemně vylučují)
    - pouze poznámka: objevuje se (ano × ne)
  + -> tyto vrstvy jsou na úrovni „slovního druhu“ (jako samostatné vrstvy)
    - pozn.: pokaždé nemusí být všechny 4 vrstvy vyplněny
  + jestli tyto vrstvy vytvoříme zároveň pro pravou a levou ruku, tak najednou máme 6 nových vrstev
    - je otázkou, jestli se držet spíše malého počtu vrstev (v korpusech ve světě, třeba 40 až 50 vrstev)
* u spec. znaků a nepřímých pojmenování je ohraničený počet existujících znaků
  + vytvářel by se heslář?
  + napsat pojmenování do glosy?
* konverzivní kategorie
  + budeme je určovat na základě konkrétní instance (pakliže v dané promluvě není možné určit jednu kategorii)
  + když bude z věty jasné, že se jedná o sloveso, zařadí se to do kategorie sloves
  + považujeme za výhodnější s konverzivními kategoriemi pracovat
    - zrušit nebo nechat prázdnou kolonku je asi o něco jednodušší proces, než zpětné vytváření nové položky v hesláři
* Jak pracovat s těmi vrstvami: KNIHOVNA: vyznačeno v rovině slovních druhů jako celek (KNIHA+KLAS.)
  + klasifikátor oddělit v samostatné vrstvě → stereotyp: symbolic subdivision
* Tudíž kdybychom měli „AUTA STOJÍ“
  + v rovině slovních druhů bychom použili označení pro sloveso
* Jedna osoba chodí sem a tam:
  + jak přeložit/zapsat glosu (zpětně to nikdo nevyhledá...)
  + problém v počátečních fázích anotace
  + asi by se to řešilo zkratkou (přiložený seznam pro tyto konkrétní případy)
  + na úrovni slovních druhů by bylo „chodit“
  + klasifikátor: klasifikátor pro jednu osobu
  + Inkorporace:
    - buď číslo (jedna osoba)
      * pozn.: není už číslo jedna obsaženo v klasifikátorovém tvaru ruky?
    - adverbia
    - asi by bylo vhodné využít možnosti poznámek, s nárůstem poznámek podobného charakteru by se po té pro problém mohlo vytvářet jednotné řešení
* s navrženými vrstvami jsme spokojeni, řízený heslář pro slovní druhy lze považovat hotový
  + pro inkorporace, klasifikátory + specifikátory a další vyjadřovací prostředky by se musely řízené hesláře vytvořit)
* věci, které se spolu doplňují: „AUTO JEDE“ klasifikátor, nebo inkorporace prostředku?
  + jak poznám, když jede (inkorporace vehiklu × nebo inkorporace adverbia -> poznám jak jede/pomalu, rychle)
  + můžeme nadefinovat vrstvy, které budou skryty a odkryjí se jenom v situacích, kdy je budeme potřebovat
  + nic nám nebrání v tom mít hodně vrstev, ale v aktivní a pravidelné formě používat pouze jejich omezené množství
  + pozn.: v psaných korpusech např. máme vrstvu *pád*, ale u sloves ji prostě nevyužijeme
* 2 přístupy k anotaci:
  + omezíme si počet vrstev × jsme nuceni mít rozsáhlé řízené slovníky
  + počet vrstev neomezíme × hesláře mohou být menší
    - pozn.: když třeba nechceme tři vrstvy, můžeme mít heslář složený ze všech možných kombinací (mnohonásobně roste počet položek v řízeném slovníku, moc možností vede k horšímu hledání)
    - také řízené slovníky používat nemusíme vůbec, ale po té vše trvá déle vypsat
  + více vrstev asi lepší řešení

Naše zkušenosti s určováním slovních druhů

* JAK? : v příkladu z kontextu, je to spíše řečnická otázka
  + označilo by se to jako příslovce a další informace uvedeny do poznámky
* JAK, KDO, KDE?
  + Stejně tak například spojení: „Co s tím?“
    - analogie z mluveného jazyka
    - kdybychom označili jako PRON(co)+ PREP(s) + PRON(tím), tak z fráze nedostanu plný význam
    - to jak se to používá, funguje až na rovině interpretace
    - a může se to zapsat to poznámky, nebo vytvořit vrstvu pro elipsu
  + K čemu nám je anotace slovních druhů? Pro korpus není stěžejní, dá nám to informace o frekvenci slovních druhů, na úrovni syntaxe už vypovídá méně.
  + Vždycky budou nějaké výjimky, také záleží na tom, jak se rozhodneme s těmito spornými problémy pracovat (uvádění do poznámek, nová vrstva?)
  + v ČZJ frekventované „řečnické otázky“ – asi by bylo vhodné vytvořit samostatnou vrstvu
* problém ČÍST/NOVINY/Klasifikátor
  + jak označit že někdo drží noviny, čte, a sleduje okolí
  + hranici pro levou ruku rozšířit na celou dobu, kdy levá ruka zastupuje noviny
* TO, TAM
  + co je to? → deixe
  + je to pro vlastní vrstvu, nebo pro kategorii v hesláři slovních druhů?
  + kdybychom udělali vlastní vrstvu: JÁ, TADY, TEĎ, teoreticky také např. MINULÝ TÝDEN
  + jsou to normální zájmena a s některými adverbii mají vlastní funkci, mohou být deiktické → vrstva deixe?
  + deixi bychom chápali jako explicitní odkaz, fungovalo by to jako vrstva u některých případů
  + určovalo by pouze přítomnost/nepřítomnost u zájmen a některých adverbií (tady, teď)
  + vlastní vrstva: asi rozumná věc (vyskytuje se to v různých slovních druzích, minimálně u zájmen a příslovcí)
    - pozn.: deixe by se také dala určovat u sloves pohybu
* prohození vrstev při anotování
  + exportovat data do Excelu, tam upravit a po té znovu nahrát
  + asi jde i přímo v ELANu
  + konzultovat s manuálem
  + v případě přetrvávajících problémů napsat e-mail

Návštěva Lucie Stanovské

* V rámci třetího předmětu v odborném modulu vytvářeli malý korpus
* Jak pracovat s ELANem je naučil David Lukeš (Ústav českého národního korpusu FF UK)
* cvičení nastavování kamer a natáčení
* cvičení příprav elicitačních materiálů
* natáčeli se dva respondenti, Radka Nováková fungovala jako moderátor (zasahovala i do promluv)
* inspirováno polským korpusem (hlavně anotační materiály), něco upravovali podle svých potřeb
* dotazník: respondenti ho vyplňovali s pomocí Radky Novákové
* různé elicitační úkoly (např.:
  + kalendář: každý dostal svůj různě zaplněný diář, úkol domluvit se, kdy se sejdou
  + převyprávění komiksu
    - žába: chlapeček ji má za mazlíčka, žába se ztratí, chlapeček ji jde hledat a to se mu nakonec podaří
    - zjistilo se, že část komiksu chybí, ale respondenti si s tím poradili
    - tento úkol byl anotován
  + diskuze na inkluzi vzdělávání: tuto část úkolu vypustili, jednak z časových důvodů, jednak z toho, že oba respondenti se v oblasti nepohybují a asi by pro ně bylo těžké o daném tématu hovořit
* Při natáčení (místnost 422B), stoleček na kterém byly dva notebooky (respondenti se občas k notebookům předkláněli)
* Některé elicitační materiály byly jen v prezentaci, některé na papíře
* Problémy:
  + elicitace barev: nepoužívali jednoduché kartičky, ale složitý obrázek (složený z různých geometrických obrazců) - chtěli, aby také ukazovali různé tvary →
  + moc složité → výsledek neuspokojivý
  + video komiks vyšel jako lepší materiál, než komiks statický (respondentům se snadněji převypravoval)
  + Vánoce/Velikonoce: Radka Nováková se nechala moc vtáhnout do diskuze
  + Mapa: asi by bylo vhodnější použít nějakou vymyšlenou/zjednodušenou podobu mapy
* Jednotlivé vrstvy
  + inspirovány polským korpusem
  + rozšířeno o vrstvu pro pohyby těla
  + tři zástupné značky pro gesta:
    - dlaně vzhůru (%)
    - mávnutí rukou (&)
    - třepotání prstů (@)
      * „nemohu si vybavit“
    - kdyby to bylo jiné gesto: označení jako „G“
* Vrstva pro slovní druhy
  + brána jako „ne úplně důležitá“ zaznamenány pouze podstatná jména, slovesa, přídavná jména
  + pozn.: dát si ruce pod hlavu (je to klasifikátor?) × je to ikonické vyjádření?
    - je to sloveso?
    - ve chvíli, kdy ikonicky přenáším vizuální podnět (elicitace obrázkem komiksu) tak ten znak nemusí nutně být slovesem

Jak dál?

* Potřebujeme začít korpus?
  + moc práce × čím více lidí, tím lépe
  + kompetence přítomného rodilého mluvčího znakového jazyka jsou neocenitelné
  + pokračovat v této naší pilotáži, sbírat problematická místa a do budoucna s nabytými zkušenostmi pracovat?
  + problémy s žádostí o granty: je potřeba publikovat (granty dostávají projekty, které již nějakým způsobem fungují, ne ty v počáteční fázi)
  + tento seminář může fungovat jako předpříprava článku zabývající se reliabilitou slovních druhů v ČZJ.
* představa další spolupráce:
  + jednalo se o poslední společnou hodinu, dobrovolně pokračovat v anotování
  + na začátku letního semestru se sejít a říct si, kdo, co může
  + účastníci semináře a další zájemci – vytvoření nezávazné pracovní skupiny?
  + po zkouškovém dostaneme pozvánku, kdy se sejdeme a rozhodneme se, jestli bychom s tím chtěli pokračovat
* zápočet: na jednu stránku sepsat reflexi (buď tohoto kurzu, nebo k práci při anotování)
  + stačí v bodech,
  + možnost zpracování formou videa v ČZJ