# Workshop korpus českého znakového jazyka (3)

**Anička: DGS NON-tokens**

* Tokeny: konkrétní produkce × lemma (reprezentativní jednotka konkrétních realizací)
  + viděl psa -> tokeny
  + vidět pes -> (lemma/slovníkový tvar)
* Máme materiál skládající se z tokenů, článek se zabývá případy, kdy jsou tokeny zvláštní/netypické (případy, kdy je člověk normálně nepoužívá)
  + např. nejazykové projevy: já to ukazuji takto, tento to ukazuje takhle
  + autoři si myslí, že by se to v korpusu mělo vyznačit, protože to není typický projev
* video materiál byl získání ze dvou zdrojů
  + korpus německého znakového jazyka
  + specifické materiály: neslyšící měli možnost reagovat na znaky vyexcerpované z korpusů, které se připravily pro vznik slovníku -> vyjádření např. pomocí videokomentářů (tato videa byla také použita)
* Proč se zabývat non-tokeny?
  + Nejedná se o typickou produkci mluvčího -> mohlo by to zkreslovat představu o té skutečné
  + z korpusu se nevylučují, ale měla by jim být přidělena nějaká specifická značka
  + pozn.: jiná situace nastává, když jsou elicitovány jednotlivé znaky (respondent ukáže znak, ale už nevíme, jestli tento znak respondent také běžně používá, nebo ne)
* volná produkce z korpusu:
  + - metajazyková produkce:
      * např.: „já to znakuji takto, moji rodiče jinak“ -> i když respondent při promluvě takový znak použije, není to pro něho přirozené a může tak být ovlivněna celá produkce
    - převzetí znaku komunikačního partnera
    - jak se co ukazuje v cizích znakových jazycích
    - specifická skupina: negativní mimika (respondent se při produkci znaku tvářil, že znak nepoužívá)
    - další skupina: neslyšící komentují nějaká gesta (např. jaká gesta dělají slyšící)
    - pokažené znaky, přeřeknutí, falešný začátek, nedokončený znak
      * v cizojazyčné literatuře označováno jako „slip(s) of the hand“
    - znaky, které byly uzpůsobeny produkci (natáčení na webkameru) -> zúžení znakovacího prostoru
* Jak hodnotitel pozná, že se jedná o přeřeknutí?
  + označovali tak situace, kdy se respondent sám opravil, nebo když bylo z kontextu jasné, že se jedná o přeřeknutí
  + × ale samozřejmě jsou příklady, kdy to jasné nebylo (značkování zajišťují rodilí neslyšící)
* Je dobré začlenit do korpusu i přirozené odchylky (přeřeknutí, trochu jiné použití znaku)
  + když už to je do korpusu začleněné, mělo by to být také označené
  + s odchylkami se tak může následně pracovat a do korpusu by měly být zařazeny
* 2 přístupu ke korpusu:
  + zachycení přirozená produkce × elicitovaná produkce
  + pozn.: to jakým způsobem se jazykový materiál vytváří, ovlivňuje podobu korpusu
* Jak správně odlišit znaky a gesta?
  + zkušenost Honzy Š. z konference:
    - existují na to různé názory
      * gesta a znaky se překrývají × někdo zase znak a gesto striktně odděluje
    - názor neslyšícího přednášejícího: setkával se slyšícími a viděl, že gesta používají při běžné komunikaci (některé shodné nebo velmi podobné se znaky) -> zastával názor, že se gesta a znaky překrývají
    - v článku zmíněno, že má být při anotaci označeno vše a pak se s tím může nějak pracovat dál
      * např. škrábání -> kdy je to příznak toho, že mě něco skutečně svědí a kdy prostředek pro překonání nejistoty
    - pozn.: názor některých odborníků, že znaky vznikly z gest (ustálením)
  + pozn.: rozdílnost vnímání např. pojetí znaku/gesta v jiných kulturách
    - pohlazení břicha („chutnalo mi“) –> je znak, nebo gesto?
    - V některých kulturách se to může chápat jako znak, ale je to možné také vnímat jako gesto (funguje to např. jako gesto u slyšících Italů)
  + zkoumání gest a znakových jazyků
    - v současné době převažuje tendence přestat vytvářet mezi gesty a znaky hranici
    - pozn.: nejsou tím myšlena konvenční gesta tzv. emblémy (nesoucí jeden konkrétní význam, „lexikální znak“)
    - jsou tím myšlena gesta, která doprovázejí mluvenou řeč
      * dříve považována za něco, co nemá charakter jazyka (pohyby gest nelze rozdělit na jednotlivé komponenty jako znaky – gesta nemají morfologii)
    - objevuje se ale čím dál více strukturních analogií se znakovými jazyky -> např. značení lexikální sémantiky slovesných významů
      * např. gesta u sloves vyjadřující ukončení děje × v některých zn. j. chápáno jako gramatický jev

**Honza C.: DGS Where does the sign start and end**

Segmentace:

* segmentace: brána ve smyslu tokenizace (je obvykle prvním krokem při transkripci zn. j.)
  + předchází lemmatizaci
* existují dva základní přístupy k segmentaci (to je jak znakovanou promluvu rozdělit na jednotlivé znaky):
  + a) znaky následují bezprostředně za sebou (tam, kde jeden končí, začíná druhý), neexistují „volná místa“
  + b) nezapočítává se pohyb a změna tvaru ruky (přesun mezi jednotlivými znaky není považován za součást znaku)
    - přechody mezi znaky jsou brány pouze jako technický přesun (ne jazykový)
* V korpusu DGS pracují s druhou variantou
  + díky tomu je menší variantnost tokenů, než kdyby se započítávaly také přesuny mezi znaky
* Při segmentování je problém s určením začátku a konce znaku. Z toho důvodu jsou v korpusu DGS definována některá doporučení pro určení počátku a konce znaku při specifické situaci.
* Byl proveden průzkum: zkoumala se shoda určení začátku a konce znaku + jaké rozdíly přinese různě kvalitní záznam
  + průzkum videa z různých kamer (čím více snímku za sekundu, tím viditelnější jsou detaily) -> rozdíl byl ve snímkové frekvenci
  + jedna kamera 500 snímků za sekundu, všechno bylo super detailní
  + druhá 50 snímku za sekundu
  + na záznamech následně pracovali tři anotátoři (dva slyšící a jeden neslyšící rodilý uživatel zn.j.)
  + segmentace se zaměřila pouze na dominantní ruku
    - -> záznamy z výkonnější kamery se pro potřeby segmentace musely zpomalovat
    - jednotlivé snímky byly sice kvalitní, ale práci anotátorům to nijak neušetřilo (problém také s tím, že znaky nelze přehrát v přirozené rychlosti a tak je vyžadovaný větší čas na přehrání)
    - ne vždy se anotátoři shodli na tom, kde znak začíná a kde končí
    - přesto, že u videí s velkou snímkovou frekvencí, byly jednotlivé snímky ostřejší a lépe rozkouskované, nevedlo to k větší shodě mezi anotátory
    - Anotátoři se častěji shodli na určení konce znaku (někdy dokonce všichni tři určili stejné místo)
* poznámka ke shodě mezi anotátory:
  + existuje několik způsobů jak ji měřit
    - změření procentuální míry shody -> závisí na velikosti vzorku
    - -> u deseti videí bude 80% shoda znamenat, že se anotátoři neshodli ve 2 případech, když vzorek bude obsahovat 10 000 záznamů, tak neshoda nastala ve 2000 případech
  + koeficient shody: Cohenovo kappa
    - koeficient shody mezi anotátory
    - výsledkem není procentuální shoda mezi anotátory, ale číslo mezi 0 a jedničkou (1 představuje 100% shodu, 0 – 0% shoda)
    - když je shoda větší jak 0,5 tak je označována jako „přijatelná míra shody“
    - když větší jak 0,8 tak jako „velmi dobrá shoda“
* Pozn.: program ve kterém budeme anotovat (ELAN), dovoluje hloubkovou (několika úrovňovou) anotaci -> umožňuje anotovat např. simultánnost znaku
  + simultaneita je opravdu specifická pro znakové jazyky
  + např. při výčtu, nebo předčasné nastavení pasivní ruky
  + opravdu simultánní projevy -> (ŘÍDIT + VOLAT)

**Honza Š.: DGS Annotation of Mouth Activities**

Mluvní komponenty

* jak zapisovat to, co se dělá ústy (jak zapsat/anotovat)
* mouthing: pohyby vycházející z mluveného jazyka, spíše mluvní komponenty
* mouth gestures: vnímáno spíše jako orální komponent
* v korpusu používají tři možnosti zápisu:
  + zapsání slova
  + + 2 systémy vycházející z fonetické transkripce
* pracovali s avatarem, který umí rozklíčovat transkripční systém
* anotace:
  + dva transkripční systémy vypadají podobně
  + je otázkou, jak s tím umí anotátoři pracovat a jestli systémy skutečně používají všichni stejným způsobem
* pozn.: mělo by se anotovat do různých částí (např. když jeden mluvní komponent je stejný pro dva následující znaky)
  + u slovníku zn.j. se ke znaku připojí mluvní/orální komponent relativně snadno
  + × při promluvě se nemusí přesně překrývat znak a mluvní/orální komponent -> musí se tedy minimálně oddělit manuální a nemanuální vrstva
* zmíněna je snaha o zajištění automatizovaného zpracování (ušetření času – první zpracování počítačem, člověk by pouze následné kontroloval toto zpracování)
* Které komponenty bude pravděpodobně jednodušší anotovat? Orální, nebo mluvní?
  + Snadněji půjdou předdefinovat spíše orální komponenty
  + pozn.: pohyb úst má často distinktivní platnost (rozlišuje význam znaku)

**Žofka: DGS Multi-sign lexemes**

* jak v korpusu označit, že po sobě následující znaky jsou součástí nějaké fráze? Že tvoří samostatnou jednotku.
  + dva znaky u sebe –> např. přidat speciální značku
  + je nutné alespoň na jedné rovině anotace označit, že se k sobě pojí
  + dvě situace:
    - znaky se k sobě pojí (kolokace)
    - znaky se k sobě nutně pojí (typ frazému, kompozita)
      * TVRDOHLAVÝ, KANCELÁŘ (PSANÍ+MÍSTNOST -> to dohromady ukazuje jednu skutečnost)
      * v korpusu to musí být začleněno (nestačí pouze označení samotných částí -> HLAVA, TVRDÝ (netvoří to význam tvrdohlavý)
      * pozn.: příklad z čj: „Plakat strachy“
        + „strachy“ - dnes se stává typem adverbia
        + kdyby substantivum –> „strachy“ je plurál (u abstrakt plurálový tvar nebývá typický)
* -> to, že se vysegmentují jednotlivé znaky, ještě neznamená, že je hotovo, některé znaky se musí zpětně propojit

**Anotace**

* funguje v korpusech normálně to, že stejný záznam anotuje několik lidí?
* Ano, ve chvíli, kdy se jedná o veřejný korpus, nebo nějakou studii, měla by být zveřejněna také úroveň anotátorské shody
* Měli by anotovat jenom rodilí mluvčí?
  + když by součástí byly např. také překlady –> tak je vhodné uplatnit např. tlumočníky
  + to, že by anotovali pouze rodilí mluvčí, může mít i nevýhodu např. tendence vkládání vlastní interpretace („ten znak musel skončit už tady“)
* Korpus Ortofon:
  + Studenti sehnali nahrávky, odevzdávají potvrzení o tom, že všichni účastníci záznamu souhlasí s jeho zpracováním (je otázka, jestli skutečně takový souhlas mají)
    - pozn.: v případě žákovských korpusů – mnohem důslednější shánění souhlasů

**ELAN**

* Stáhnutí ELANu na stránce: tla.mpi.nl/tools/tla-tools/elan
* Existuje více anotačních softwarů:
  + iLEX – dobrý pro určité věci, nám se moc nehodí
  + ANVIL (**AN**notation of **VI**deo and **L**anguaga data)
    - jedním z prvních programů, vychází z něho ELAN
    - dnes už se s ním moc nepracuje, je komerčního charakteru
  + EXMARaLDA  (**EX**tensible **MAR**kup **L**anguage for **D**iscourse **A**nnotation)
    - dnes už se s ním také moc nepracuje
* ELAN (**E**UDICO **L**inguistic **AN**notator)
  + vznikl v rámci jazykového archivu v Holandsku (Max Planck Institute) za účelem dokumentace jazyků × ne znakových jazyků
  + vytvořili ho lingvisté, co se zabývají např. zaznamenáním ohrožených jazyků
  + výhody: je zdarma, je opensourcový (každý se může podílet na jeho vývoji, skutečně se vyvíjí a vydávají se nové verze)
    - -> když někomu něco nevyhovuje, může si program snadno upravit
  + představuje současnost a minimálně i blízkou budoucnost multimodálních korpusů
  + umožňuje práci se zvukem i obrazem
  + umožňuje synchronizovanou anotaci
  + umožňuje anotovat na časové ose
    - přesnost je tak velká, že data, která vznikají lze použít např. v rámci exaktních psycholingvistických výzkumů
  + program funguje v prostředí programovacího jazyka JAVA (Java musí být na počítači nainstalovaná)
  + výsledkem jsou data ve formátu xml
    - zdrojový kód z ELANu je jednoduchý a je velmi snadno zpracovatelný

**Práce s ELANem**

* Prvním krokem je **založení nového souboru** s anotací:
  + **„FILE“** -> „**NEW“** (tam vybrat video)
    - pozn.: novější verze umožňují přetáhnutí ikony videa přímo do rozhraní programu
  + ikonka šipka **spustit (play)**
  + zkušební anotace (levým tlačítkem myši označit část videa) –> kliknutí pravým tlačítkem myši do vybrané vrstvy (růžová) přidat novou anotaci – „**ADD NEW ANNOTATION**“
    - když založím nový anotační soubor – ELAN automaticky vygeneruje první anotační vrstvu (ta růžová)
* obvykle potřebujeme vrstev více
  + jedna je např. věnována transkripci, nebo překladu, další obsahují např. glosy, nebo lingvistickou anotaci
  + typicky první vrstva překladová, transkripční
  + druhá už morfologická
  + další vrstvy mohou být věnované dalším částem promluvy (mluvní komponenty, pohyby hlavy, rozdělení na pravou a levou ruku)
* **vytvoření nové vrstvy**:
  + horní menu: „**TIER“** -> „**ADD TIER“**
  + přidá se nová vrstva (může se konkrétně pojmenovat, obyčejně se využívá nějaká značka)
* Cílem hodiny vyzkoušet si segmentování krátké promluvy na jednotlivé znaky
  + proto jsme nevytvářeli více vrstev, ani nedefinovali jejich vzájemné vztahy
  + ELAN umožňuje pracovat s několika vrstvami pod sebou
    - lze nastavit, že to, co máme v jedné vrstvě, se zkopíruje do nějaké další vrstvy (při práci po té není potřeba některé kroky při anotaci zbytečně opakovat)
  + některé vrstvy se mohou prolínat (koartikulace, simultánní projevy)
    - × anotace může být i nezávislá na ostatních vrstvách
  + úkol: vysegmentovat tři znaky a do anotace dopsat význam.
* ELAN umožňuje přednastavit si hodnoty nějaké kategorie a z té poté následně pouze vybírat jednotlivé možnosti (nemusí se vše vepisovat ručně)
  + např. přednastavení souboru orálních komponentů
  + pozn.: vrstva transkripce, pravděpodobně nebude mít přidělenou množinu anotačních hodnot
* pro nejpočátečnější fázi segmentace je lepší, když necháme běžet video a necháme dělat **hrubou segmentaci**:
  + záložka „**OPTIONS**“ (třetí zprava)
    - několik možností základního zobrazení –> „**SEGMENTATION MODE**“
    - **„CONTROLS“** (vpravo – může se nastavit pomalejší rychlost přehrávání)
  + lze nastavit kontrolku pro **zpoždění mezi reakcí a videem**
    - např o 200 milisekund (**SEGMENTATION** –> **DELAYED MODE** –> **200**)
* jak vytvořit **vrstvu s přednastavenými hodnotami**
  + soubory s anotačními hodnotami se nazývají „**controlled vocabulary**“
  + nejdříve se musí vytvořit a až po té se mohou v jednotlivých vrstvách použít
    - záložka „**EDIT“** (4. možnost)
    - **název souboru** (např. slovní druhy -> slovní\_druh)
    - **ADD** –> vytvoříme soubor
    - jednotlivé hodnoty: pole „**ENTRY VALUE**“ a tam se můžou vkládat jednotlivé hodnoty (SUBST, ADJ, PRON, NUM, VERB)
  + **LINGUISTIC TYPE**
    - dalším krokem je vytvoření „linguistic type“ (vlastnost anotační vrstvy)
    - přisouzení vytvořených přednastavených hodnot k nějaké vrstvě (její vlastností ale není pouze to, že obsahuje vytvořený vokabulář, ale i definice oblastí, kterých se to týká)
    - záložka „**TYPE**“ -> „**ADD NEW TYPE“**
    - pojmenovat
    - „**USED CONTROLLER VOCABULARY“** –> vybrat a dát „**ADD“**
  + poslední krok
    - záložka „**TIER**“ – „**ADD NEW TIER“**
      * možnost přidat jméno participanta
      * stejně jako jméno anotátora
      * **TIER TYPE** (vybrat ten, co jsme vytvořili) a **ADD**
      * může se vybírat z předvolených hodnot

Na příští hodinu

* K otázce slovních druhů
  + rozhodnout se jaké slovní druhy pro korpus znakového jazyka využít (ty co známe z čj? Není vhodné na úrovni slovních druhů přidat nějakou další kategorii – např. klasifikátor?
* Moodle: ke stažení budou nahraná videa i seznam, kdo které bude zpracovávat + zadání cvičení, které bude procvičovat to, co jsme probrali
* Bylo by dobré, kdybychom na příští hodině mohli pracovat s vlastními notebooky (dva jsou k dispozici k zapůjčení)