# Workshop: Korpus českého znakového jazyka (2)

Organizační:

* Nahrávky: třetím rodilým mluvčím bude Jan Wirth (natočen bude v úterý ráno)
* na příští hodinu (30. 10.) každý dostane zadaný odborný text, z kterého v bodech vybere ty nejvýznamnější informace (poslat přípravu tlumočnici do čtvrtka 26. 10.)
	+ více info viz moodle
* Do příští hodiny nám také budou přiděleny nahrávky, které budeme zpracovávat
* Nahrávat se bude: převyprávění videa, vlastní vyprávění
* V prezentaci na moodlu je aktualizovaná verze programu hodin

Dnes – představení čtyř korpusů znakových jazyků (se zaměřením na zpracovávaný materiál)

* to znamená → jak se pracovalo s produkcí zn. j. a jak bylo řešeno nahrávání
	+ to jsou dvě části korpusů zn. j., o kterých se moc nemluví
* zmíněné korpusy jsou jen výběrem existujících korpusů, ale mohou posloužit jako referenční body, existuje o nich spousta literatury

Korpus BSL

* sestaven z produkce čtyř částí:
	+ narativy
		- vyprávění: respondent např. vypráví, kam chodil do školy, o Neslyšících apod.
		- narativy spontánní × elicitované
			* rozděleno podle toho, jak vypadají instrukce
			* elicitované: vyprávějte nám něco, o nějakém tématu, nebo vyprávění na základě stimulu (přehranné video, obrázková knížka → co se děje na obrázku)
	+ lexikální elicitace
		- elicitují se jednotlivé znaky → jednak pro tvorbu slovníku, jednak pro získání regionálních variant (např. Severní Irsko, Skotsko)
	+ rozhovory výzkumníka/lingvisty s nahrávanými osobami na předem daná témata
	+ konverzace mezi dvěma participanty
* Rozestavění kamer
	+ byly tři
	+ jedna na každého participanta zvlášť + jedna na celou scénu (nejzákladnější varianta)
	+ \* u některých nahrávek byly kamery čtyři
* autoři: Schrembri, Thomson

Korpus PJM

* polský korpus
* Složeno z:
	+ baterie elicitačních testů (obrázky, filmy, opakování rozhovoru) a to jak lexikální, tak i obecnější produkce
	+ diskuze na dvě předem zadaná témata
	+ neřízená konverzace
	+ lexikální elicitace (regionální rozdíly)
* korpus stále v přípravě, snaží se být na vrcholu technických možností → nahrávání složitější
	+ z práce Pawla Rutkowského
	+ 5 kamer
		- jedna pro celkovou scénu
		- 2 pro každého účastníka (ptačí a přímý pohled)
		- Dotaz: Proč zrovna kamery z ptačí perspektivy?
			* některé aspekty artikulace je lepší zachytit ze dvou perspektiv × nebyl by lepší boční pohled?
			* pohled z ptačí perspektivy zvolen možná kvůli prostorovému zobrazení (využití prostoru, vzdálenost rukou)
			* videa se budou zpracovávat asi strojově, a aby to bylo možné, je nutný vrchní pohled

Korpus NGT

* nizozemský
* je veřejně dostupný a v odborných textech bohatě popsán
* tvořen několika postupy
	+ nejdříve se participant představí, řekne něco o sobě
		- většinou jsou získané informace součástí metadat, tímto úvodem budou také natočeny
		- další výhodou je např. získání znaků pro osobní jména či jmenné znaky
	+ diskuze na téma „deaf issues“
	+ převyprávění znakovaného příběhu
	+ převyprávění kresleného filmu
	+ převyprávění hrané grotesky
		- pozn.: to si vyzkoušíme v rámci nahrávání
	+ převyprávění obrázkového příběhu (aby tam nebylo to zkreslení pohybem)
	+ úloha „najdi 5 rozdílů“
	+ pozn.: odpadají rozhovory
* 4 kamery namířeny na participanty (přední a horní pohled)

Korpus DGS

* německý
* tvořen z:
	+ narativů
	+ konverzace
	+ diskuze
* nejsložitější kombinace kamer (7)
	+ 3× 2 na participanty (2× 2 přední + 2 horní) + 1 celkový pohled)
	+ více viz práce Christiana Rathmanna

Korpusy:

* Nejmladší z korpusů je ten polský (pořád vzniká)
* Německý: kolik mluvčí a kolik je to materiálu?
* 330 mluvčích ze všech možných částí Německa
* asi 500 hodin materiálu
	+ pozn.: je otázka, co je moc a co je málo mluvčích
* materiál: cca 3 miliony tokenů (vyrovná se to nějakému menšímu korpusu mluveného jazyka)
* Pozn.: je nutné si uvědomit, že uživatelů znakového jazyka bude v Německu oproti ČR více

Praktická ukázka NGT

* Radboud University, v městě Nijmegen
* samotný korpus je k dispozici na samostatných stránkách:
	+ Ústavu Maxe Plancka (Max Planck Institute for psycholinguistics)
* k dispozici jsou video materiály celého ústavu (včetně celého korpusu)
	+ pozn.: k dispozici není pouze korpus NGT, ale i další (ne všechny jsou ale veřejně přístupné, u některých je potřeba se registrovat)
	+ funguje to jako úložiště dalších korpusů
		- např.: Korpus afrických zn. j., Austrálie
	+ obecně jsou problémy s úložišti dat
		- např.: internetová jazyková příručka (veškerá data příručky jsou umístěna na Masarykově univerzitě)
		- problém je s financemi + kdo to bude spravovat
		- jedna z prvních otázek ve chvíli, kdyby se dělal korpus čzj, by tedy také byla např. „Kde bude korpus uložen?“
	+ korpusy jsou v různých sekcích
	+ více info viz moodle: článek De Vosové
		- pozn.: u Kata Kolok specifické, že tam nebyli jenom rodilí mluvčí, ale také sekundární uživatelé
	+ v korpusu NGT
		- jednotlivé nahrávky nesou označení S → „session“
		- teď asi 91 mluvčích
	+ informace v Elanu
	+ videa nejsou velká, dají se stáhnout
	+ neexistuje online prostředí, ve kterém by se dala přímo přehrát (× to je něco, co bychom od korpusu čekali), nebo kdyby se to dalo přehrávat rovnou i s tou anotací (to Elan zatím neumožňuje)
	+ Pozn.: s Elanem budeme pracovat

Graf:



* Na ose y:
	+ ekonomická validita (přirozenost nahrávek)
	+ nejvyšší by mělo to, co by se např. nahrálo tajně venku na ulici
	+ nízká: to, co je natočeno ve studiu a mluvčí má navíc k tomu na sobě připevněny snímací elektrody
* Na ose x:
	+ komplexnost dat, která z materiálu získáme
	+ také možno označit jako „zpracovatelnost“, nebo „použitelnost“
* Osy vytváří napětí mezi tím, co můžeme získat, co chceme získat a s čím si poradíme v rámci zpracování
* V bodě 0 jsou velmi komplexní data
	+ komplexní jsou tak moc, že je nelze použít v empirické studii
	+ čím více vpravo na ose x budou, tím více jsou použitelná
* Chceme taková data, která by šla zpracovat strojově
* Když skrytou kamerou nahrajeme dění na ulici, získáme velké množství dat, které ale vyžaduje velké úsilí k tomu, aby se dalo nějak použít
* Naším cílem je vysoká ekonomická validita, ale také určitá zpracovatelnost
	+ abychom měli kvantitativní data (to znamená, aby se dalo měření zopakovat a aby se to dalo porovnávat s dalšími výzkumy)
* Bod A: „polní nahrávky“
* Bod B: neurologický experiment
* Bod C: náš cíl
	+ velká validita, ale také uchopitelná /jednoduchá data, s kterými si budeme vědět rady (to znamená, co nejmenší počet proměnných, které budeme muset kontrolovat)
* Bod D: není nedosažitelný cíl, ale je to hudba budoucnosti
	+ např. automatická detekce pohybu
		- → validita bude vysoká
		- → detekce bude automatická (snadnou cestou se zjistí např.: použité tvary rukou, pohyby, rychlost pohybů, vzdálenosti rukou)

Při nahrávání:

* participanti při práci na korpusu podepisují formulář, jehož součástí je např.:
* potvrzuji, že jsem přečetl poučení o výzkumu
* jsem si vědom toho, že má účast je dobrovolná,
* dávám svolení ke zpracování svých anonymizovaných odpovědí
* souhlasím s pořízením audio-video nahrávek
* souhlasím s tím, aby nahrávky byly zdarma k dispozici zájemcům z odborné veřejnosti
	+ pozn.: jednou by měl existovat něco jako etický kodex, který by byl vzorem pro všechny výzkumy na fakultě
	+ laboratoř LABELS má vzory – tam už je problém vyřešen

Ukázka nahrávky z korpusu NGT

* narativ se často převypravuje někomu, kdo není přímou součástí výzkumu → pro participanta je příjemnější, že na někoho mluví
* v ukázce: na obrazovce dvě ženy, jedna reaguje na druhou,
	+ pravděpodobně seděly naproti sobě, ale video vytvořené tak, že vidíme obě ženy najednou

Praktické natáčení:

* nutné zajistit, aby byla nabitá kamera a aby byl k dispozici dostatek volné paměti
	+ → na nahrání je pouze jeden pokus, poté už se jedná a promyšlenou/nevyhovující promluvu
* kdyby byla např. různě velká prodleva mezi participanty od zhlédnutí videa, které mají následně převyprávět, nebo by byla prodleva příliš velká, také by se jednalo o nevhodný materiál
* které nahrávky jsme točili:
	+ převyprávění hrané grotesky
	+ rozhovor: vyprávění o svém zážitku z léta, nebo co jsem dělal o víkendu
	+ zodpovězení dotazu ohledně blížících se voleb
		- očekávání, že se v promluvě objeví znaky pro názvy stran nebo jména politiků